Hindawi Publishing Corporation
Mobile Information Systems
Volume 2016, Article ID 8068209, 11 pages
http://dx.doi.org/10.1155/2016/8068209

Research Article

A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes: A Case Study of Mexico City

Félix Mata, Miguel Torres-Ruiz, Giovanni Guzmán, Rolando Quintero, Roberto Zagal-Flores, Marco Moreno-Ibarra, and Eduardo Loza

Instituto Politécnico Nacional, UPALM, Zacatenco, 07320 Mexico City, DF, Mexico

Correspondence should be addressed to Miguel Torres-Ruiz; miguel.torres.ruiz@gmail.com

Received 20 November 2015; Revised 12 February 2016; Accepted 14 February 2016

Academic Editor: Salil Kanhere

Mobile information systems agendas are increasingly becoming an essential part of human life and they play an important role in several daily activities. These have been developed for different contexts such as public facilities in smart cities, health care, traffic congestions, e-commerce, financial security, user-generated content, and crowdsourcing. In GIScience, problems related to routing systems have been deeply explored by using several techniques, but they are not focused on security or crime rates. In this paper, an approach to provide estimations defined by crime rates for generating safe routes in mobile devices is proposed. It consists of integrating crowd-sensed and official crime data with a mobile application. Thus, data are semantically processed by an ontology and classified by the Bayes algorithm. A geospatial repository was used to store tweets related to crime events of Mexico City and official reports that were geocoded for obtaining safe routes. A forecast related to crime events that can occur in a certain place with the collected information was performed. The novelty is a hybrid approach based on semantic processing to retrieve relevant data from unstructured data sources and a classifier algorithm to collect relevant crime data from official government reports with a mobile application.

1. Introduction

Nowadays, millions of citizens go through the streets of Mexico City, taking some specific routes that are planned by using either public or private transportation or even walking. Although there are well-known routes that citizens most often take for their traveling, new routes (probably unsafe) might be experimented especially for newcomers. This generation is not an easy task; generally it is made by routing systems that search the shortest or fastest paths [1]. Nevertheless, a feature like security is not taken into account when the route is generated by these systems. In particular, safety is a critical characteristic that should be considered for these mobile applications, especially in overcrowded large cities. Many routing systems are based on processing linguistic techniques and treating with name of places to define the origin and destination. Thus, these works have faced well-known linguistic problems (e.g., polysemy). In [2], a word sense disambiguation method used to name places was applied. Our work is oriented towards crime prevention by using information retrieval techniques and a clustering classification method. Other works are only focused on generating a better performance [3] and computing the shortest path [1]. Other approaches deal with crime information by using location recognition, network information, geographic information, and technologies such as augmented reality, short message service (SMS), and near field communication (NFC) [4]. However, clustering approaches are used to obtain the risk level in a specific area. So, the novelty of our proposal is the use of two heterogeneous data sources: *corpus* of *tweets* and government official reports, which were integrated and processed by the Bayes algorithm for obtaining a risk level applied to routes in a mobile system.

There are different approaches for computing the safe path and analyzing crime events, which are based on mathematical models, machine learning, geospatial analysis, and

crowdsourcing methods in mobile information systems [5]. In [6], a risk model for urban road network is proposed; it uses a mathematical model based on civic datasets of criminal activity for mobility traces in the city. But it ignores the temporal dimension for filtering data crime. In [7, 8] historical crime data are processed for identifying patterns that help in crime prediction. In our work, we identify the risk level without considering patterns directly. Commercial spatial analysis tools are presented in [9, 10]. They are very useful to study and analyze crime patterns for strategies against crime. Crowdsourcing is an interesting approach for analyzing crime data, including social information [11]. According to [12], there are important relationships between crime and people's activities in the spatiotemporal context.

This paper introduces a framework based on an approach for integrating official crime reports given by public authorities with crowdsourcing data, which were obtained from the Twitter streaming. The goal is to provide safe routes with a mobile information system in order to increase the confidence level of the citizens that use the urban infrastructure of the city. Summing up, the approach combines statistical data obtained from official information, with the perception of people derived on a daily basis and reflected by crowdsourcing data. Thus, this framework takes into account events for a particular time, and continuous updates are performed in order to provide recent events that occur at specific places, which are reported by a social network community. The crowdsourcing-based method considers a large database of tweets, which were collected by the mobile application, while official data were given by local authorities. Thus, the approach applies a social mining technique in order to extract features from the Twitter dataset and enriches the characteristics that emerged from a statistical database that qualifies the safety in Mexico City.

The remainder of the paper is organized as follows. Section 2 presents a discussion of the related work. Section 3 describes the general framework of the mobile information system. Section 4 depicts the experimental results of the mobile application, as well as the comparison with other systems. Section 5 highlights the conclusion and future work.

2. Discussion about the Related Work

The study of crime prevention in several community environments has been explored. The Department of Police Services of the California State University offers a list of safety mobile applications [5]. Another example is the Sentinel Campus, which is a free mobile application that provides crime statistics for more than 4,400 universities [13]. In the social web context, Wikicrimes is a Brazilian site that presents a crime status based on *hot spot* maps, in which users can report regional crime data [14]. The proposal retrieves information in a structured way by specialized system. However, our approach retrieves information in unstructured way by a nonspecialized system. In [15], a study of crime prevention systems with an analysis of the main web and mobile crime information systems is described. Here, the information retrieval task is collaborative; however, the authors do not use a hybrid approach which integrates social and official sources. Other applications have used spatial statistics for identifying factors that affect directly the crime occurrence, in order to generate a public map of crime prevention [10]. Nevertheless, it does not consider the case when a person needs to cross an area or point walking or in a car.

On the other hand, mobile applications for routing and planning in city environments are increasingly becoming essential for improving urban spaces. Thus, this indicates the appearance of the next generation of mobile information systems, in which recommendations are focused on decision making in order to adequately support the growth of big cities. Applications like CrowdPlanner [16] and DroidOpp-PathFinder [17] are addressed to generate crowd-sensed route recommendation systems, which request from users to evaluate candidate routes recommended by different sources and methods, for determining the best route based on the feedback of those users. The routes generated by these applications are evaluated by people; in our case the assessment is made by a clustering algorithm and by the crime occurrence.

Recently, a large range of experimental applications, which take into account the social network and crowd-sensed data, were developed. In [18], a spatio- and crowd-based routing system is proposed in order to improve the recommendation quality; however, it is only based on volunteered information and the enrichment of the crime data sources is not considered. A similar work is presented in [11], in which the sentiments are considered as a perception of users with respect to the security in certain geographic places [19]. Moreover, a system for routing of police patrols based on genetic algorithms is presented in [3]. It generates particular routes that minimize the number of crime occurrences in a given area. The route is generated by using the shortest path with data from a fixed time window.

In [20], a crime mapping and prediction based on historical data is proposed. This work makes centered analyses with spatiotemporal techniques. The analysis of crime prevention considers the network traffic and nodes. Other proposal has included the position of CCTV cameras to detect crime events [21]. Thus, [22] developed a police patrolling strategy based on the Bayesian method and ant colony algorithm in order to reduce the average time between two consecutive visits to *hot spots*. In [23], a crime ontology is described by using two scenarios of analysis: (1) the circumstances of a crime, its mechanism, information of criminals, and knowledge about the methods of crime investigation and (2) the penal procedure and other methodological and tactical recommendations of criminality, crime features, and events. Another ontology-based system in crime analysis is explained in [24], in which an ontology to represent digital incidents, associated with digital investigation and legal requirements, is described.

The cited approaches have been using spatial data of crimes, but they are not able to analyze unstructured data. In this work, we propose a hybrid approach embedded in a mobile application, which automatically combines social and official crime data, by using an ontology exploration method and the Bayes algorithm to classify crime activities, according to the Mexican penal code.

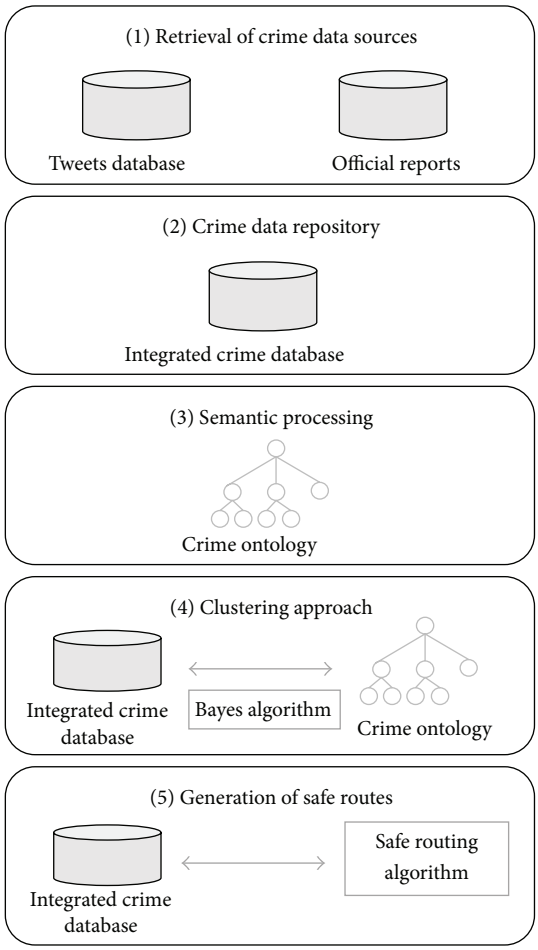

Figure 1: The general framework with its stages.

3. The General Framework

The framework is composed of five stages: (1) retrieval of crime data sources, (2) the crime data repository, (3) the semantic processing, (4) the clustering approach, and (5) generation of safe routes (see Figure 1).

Summing up, the approach consists of retrieving tweets that are related to crime events, taking into account attributes that define the time and location when an event occurs. These tweets are analyzed and integrated with crime data from official sources (government institutions). The integration process is automatically performed by using descriptions and spatiotemporal attributes of data. An application ontology is proposed to define candidates for the integration task. Later, the Bayes algorithm is used to classify data that cannot be automatically integrated. For the cases of synonymy names, the GeoNames web service is used to solve this conflict.

The categorization is carried out analyzing the spatiotemporal attributes and the description of words that appear in the tweets and/or official database record. For example, crime events could have occurred walking or by car, with or without violence. The description of these events is used to categorize each record or tweet. The crime data cannot be directly classified, because this information could be

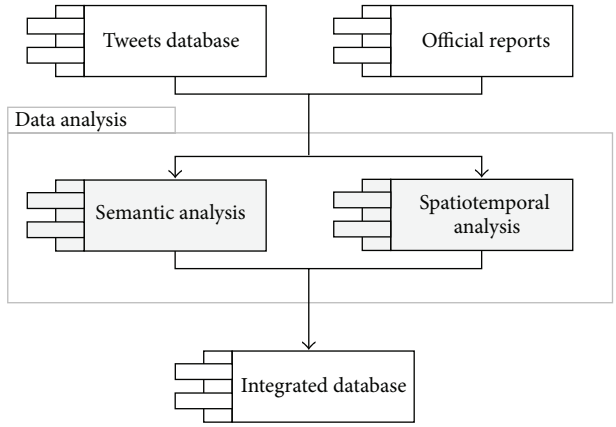

Figure 2: Hybrid approach: integrated crime database from the official and social sources.

imprecise or incomplete. For instance, if a tweet does not contain the address or any reference like points of interest, time definition, and crime or theft, then the tweet is not taken into consideration. Thus, information that is incomplete or with imprecise data will be classified by using the Bayesian algorithm. Moreover, the categorization allows us to know what type of crime/theft occurred in certain place or point (with or without violence). While the classification defines the confidence (security) level for an area or street, this is defined by the sum of all crime/theft points of an area or street. Thus, all the categorized and classified data are used as input parameters for the safe routing algorithm. The final result is a safe route that does not cross or contain points with high crime rates.

As a conclusion, the first stage presents the crime data sources to be processed: particularly relevant tweets and official reports. The second stage consists of building a crime data repository. It stores data that were integrated from Twitter and official sources. The third stage is in charge of classifying crime records by their description. The process is driven by an application ontology. The fourth stage gathers the events by crime/theft type, time, and location where they happened. Lately, classified areas are generated, according to their confidence level (how often a crime occurs in a place, date, and time) in order to be categorized like secure or insecurity clusters. The fifth stage performs an estimation based on the crime rates, which is processed by applying a Bayesian algorithm in order to obtain crime values with respect to points in safe routes. It visualizes the safe route on the mobile application, according to the following crimes: robbery with violence to passengers, theft without violence to passengers, car theft with violence, and car theft without violence.

3.1. Stage 1: Retrieval of Crime Data Sources. In this stage the data are retrieved and recollected from two sources: (1) crime information from the tweets dataset and (2) crime information from the official reports (see Figure 2). The following considerations for describing both processes are outlined as follows.

Table 1: Crime-related events from Twitter accounts covering Mexico City.

Twitter account	Location	Creation date	Followers	Number of tweets	Belongs to government	Website
SSPDFVIAL	Mexico City	07.14.2010	369,115	154.65	Yes	http://ssp.df.gob.mx/
PolloVial	Mexico City	01.31.2013	667	71.91	No	No website
Trafico889	Mexico City	05.14.2009	137,099	90.54	No	http://siempre889.com/trafico/
Alertux	Mexico City	10.16.2012	179,574	35.59	No	http://www.alertux.com/
072AvialCDMX	Mexico City	10.20.2010	83,535	134.71	Yes	http://www.agu.df.gob.mx/
RedVial	Mexico City	03.09.2010	63,702	44.81	No	http://rvial.mx/

Table 2: Record structure for the official database.

Event type	Year: 2012				Year: 2013			
	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
Robbery of passenger in public transport	159	175	154	111	102	134	115	86
Robbery of passer without violence	47	80	54	79	73	46	49	57

From the tweets dataset, it is important to know information represented by tweets that talk about crime-related events. The features were considered as follows: specialized and generic accounts, mention of crimes in popular streets, common abbreviations of crimes, popular places where some crimes were committed such as malls, entertainment sites, private and public locations, and historical monuments. This process determined the popular words and relevant concepts by using the ontology. Thus, the most common n-grams for each tweet were obtained, by sorting them according to the occurrence frequency. The repetitive n-grams were selected with a threshold of more than 100 mentions. Thus, from the most frequent unigram, bigram, and trigram lists, we have identified by hand 456 common crimes on popular streets, 150 common crime-related events, 135 common crime hashtags, 69 common nicknames, 65 common buildings, places, and monuments, 34 common abbreviations, and 26 common combinations of prepositions.

The dataset contains 450,250 tweets collected over a period of six months, from January 7, 2015, until June 24, 2015, without considering retweets and posts with blank spaces. Tweets are collected from reliable Twitter profiles that correspond to known services and institutions (see Table 1). The APIs that were used to retrieve crime tweets were the Search API and Twitter4j.

With respect to official reports, they were retrieved by the InfoDF system. The process consisted of requesting to the SSP and PGJ government agencies and the records with respect to crime information such as robbery of passerby and vehicle theft.

On the other hand, information retrieved from the SSP agency contains more details than information from the PGJ such as the type of offense, the colony or area where the event occurred, and the time. A processed record from the official police database (PGJ-DF) is shown in Table 2. It represents a crime/theft associated with a location or neighborhood, as well as the time by trimester.

3.2. Stage 2: A Crime Data Repository. In this stage the most relevant tweets are identified in order to carry out a semantic matching, which is driven by an ontology that was adopted from [23]. A fragment of this ontology for integrating semantically crime tweets with official reports is depicted in Figure 3.

The ontology is explored by Algorithm 1, and it uses the hyperonymy and synonymy as semantic relationships in order to find a matching between each tweet and crime report. It means that if a tweet is expressed by domains of time and location, their hyperons and synonyms are searched within the ontology for contextualizing them. So, synonyms and hyperons are stored into a vector. The process is repeated by each record of the official institutions; the obtained vector is compared and in case of a match, then, a term or concept is the same or has the same parent; thus, it is considered a candidate to be unified. In other words, if a match is found, then the data are mixed (tweets and databases from the PGJ and SSP).

3.3. Stage 3: The Semantic Processing. The applied process to tweets establishes the following tasks. (1) *Tokenizer*. It consists of separating and identifying the tweets, as well as filtering the tweets to remove the stop words from the token stream. This task was performed by Lucene system [25]. (2) *Processing Data*. This task involves analysis of tweets; they are identified by spatial, temporal, and description attributes in order to identify where, when, and how a crime event occurred. (3) *Categorization*. It uses the semantic matching by means of grouping the tweets, according to the crimes defined in the ontology. The semantic distance for each processed tweet is computed by using the weighted crime rate. The output of this stage is a set of tweets categorized by a crime type. Figure 4 shows a general diagram that describes the integration of these tasks. Thus, in order to illustrate the above, two different records from Twitter dataset at different levels of spatial granularity are presented as follows:

(1) @SSPDFVIAL: crime report in public transport, with a fire arm, Iztapalapa,

(2) @Alertux: theft of students in Ecatepec, streets 105 and 106 @RedVial.

Input: Tweets
Result: Unified and categorized tweet
(1) Let $q[i]$ = *elements_of _tweet*
(2) $n = 0$
(3) **while** $n! = i$ **do**
(4) Parsing and identification ($q[i]$)
(5) node.start()
(6) **while** *node* != *null* **do**
(7) $j = 0, i = 0$;
(8) **if** *Hyperonomy or Similarity*(*concept_name*) **then**
(9) *conVec*[j] = *get_parent_and_children*(*node*)
(10) node.next()
(11) *CrowdVector*[j] = *eventType_search*(*conVec*[j])
(12) temporal_search(conVec[j])
(13) j++, k++, n++

Algorithm 1: The OntoExplore algorithm.

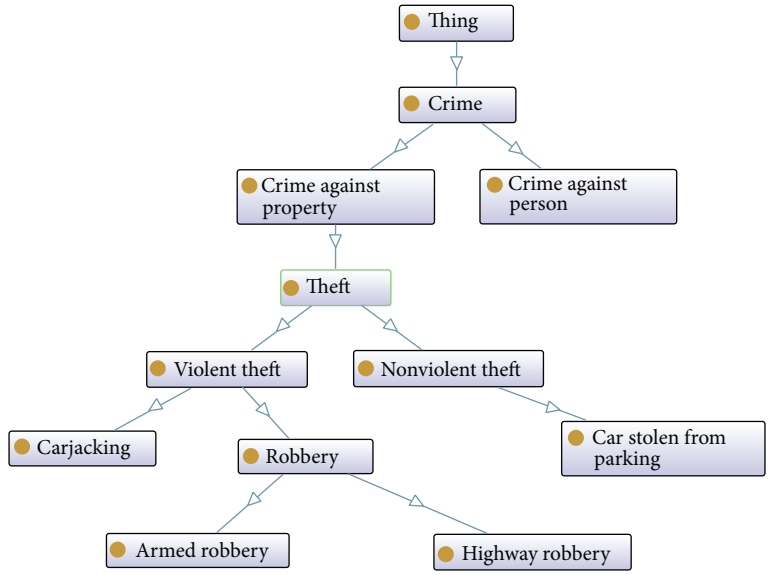

Figure 3: A fragment of the crime ontology to semantically match tweets and official information.

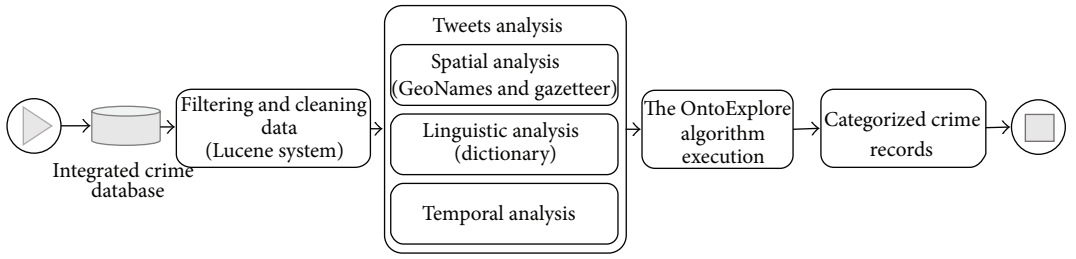

Figure 4: Tasks involved in the semantic processing.

The cleaning process includes the adaptation of Algorithm 2 for removing stop words, according to the definition presented in [26].

On the other hand, a tweet is associated with a record of the official database according to its relevance, by using the ontology and the hyperonymy relation. The relevance is measured by contextualizing the record and the tweet. This means that each semantic item of the tweet and record is identified, and then it is searched in the crime ontology in order to find matching terms, synonyms, or related concepts.

Input: An arbitrary stop word dictionary T, the set of schema trees QI
Result: A maximal set of stop words T'
(1) $T \leftarrow 0$
(2) W = the set of all words in the domain
(3) $D = T \cap W$
(4) while exist word $w \in D$
(5) for each interface $q \in QI$
(6) remove the stop word constraints for the lables of sibling nodes
(7) **if** *no stop word constraint is violated then* **then**
(8) $T' \leftarrow T \cup w$
(9) **else**
(10) remove antonymus of w appearing in D from T'
(11) return T'

Algorithm 2: Removing stop words.

For example, in the first tweet above, the location is not precise; only the name of neighborhood is given, but the streets are not defined. However, these features enrich the description of the event.

Therefore, the tweet is contextualized and related to the concept "violence." Moreover, the second tweet has a precise location, but it is imprecise regarding the event description. This process generates a semantic classification matched to the official database categorization, in which each tweet is described according to few categories either Boolean or descriptors (e.g., theft, crime, and violence). The location is derived by the neighborhood, street name, or spatial relation (e.g., near and far). The time is computed by explicit data (at 11:00 am) or temporal adverbs (e.g., now and afternoon). Thus, a parsing that contains the attributes identified by the semantic classification (event type: crime and theft in car or walking) for evaluating the domains is generated. The parsing structure that was obtained by analyzing the tweet application is presented as follows.

@semantic_classification

Theft, Crime, Theft with violence, Theft walking, Theft in car, Theft without violence

@attribute day Mon, Tue, Wed, Thu, Fri, Sat, Sun

@attribute id_location 30339461, 30339462, 30339493, 30339495, 30339496, 30339671, ...

@attribute time 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24

@data

Mon, 30339461, 10, Theft

...

Fri, 30339671, 20, Crime

...

The label @semantic_classification represents the result of the semantic matching, @attribute represents the valid attributes and their domains, and @data denotes the classified data after syntactic and semantic-syntactical analysis.

3.4. Stage 4: The Clustering Approach. In this stage an approach based on Bayes algorithm is proposed for clustering the crime data that are stored in the repository. A bag of words is used in order to classify the crimes; it means that there is a set of words that describes each type of theft or crime. If a tweet or record contains these words, it is classified according to the type of theft or crime that these words represent in the database. So, the Bayes algorithm classifies tweets or records that do not contain these words or only one of these words. The goal is to classify tweets that are relevant and that the semantic processing cannot classify. Thus, a tweet that describes a theft or crime which does not have an exact location, time, or description will be classified by using the Bayes algorithm.

The Bayes algorithm takes into account certain features (words that compose a tweet or record from official database) that are identified and assigned to a particular cluster. In this task, many clusters can appear, and they will be filtered later. The Bayes theorem is useful not only to cluster data related to crime or theft but also to cluster data when they contain words that do not belong to a type of crime; so these clusters will be omitted. The use of Bayes algorithm had two specific goals: the first one is to classify data that cannot be semantically classified and the second is oriented towards performing an estimation. For example, a user requests a safe route in certain area, but there are not data of theft/crime reports, neither Twitter nor official database for this area. So, this approach finds the cluster which belongs to the area (e.g., a safe or unsafe cluster) and makes an estimation to this zone for determining what points are probably safer than others. Finally, weights are assigned for all points and these weights are sent as parameters to the safe route algorithm.

Equation (1) was applied to data and they were clustered. The next estimation is also computed for establishing a probability for each point that belongs to a route and a crime event. In other words, such equation defines the likelihood that an event occurs on a given day and time and at a specific location:

$$P(c \mid x) = P(x) \times \frac{P(X \mid C)}{P(x)}. \quad (1)$$

Input: $N = n_1, n_2, \ldots, n_k$ Set of nodes
n_s start node
n_f finish node
Result: $D = d_1, d_2, \ldots, d_k$ Set of distance values (weighted crime rate)
(1) Assign distance values $d(n_s) = 0$, $d(n_i) = \infty \, \forall \, n_i \neq n_s$
(2) Let $U = N - n_s$
(3) Let current node $n_c = n_s$
(4) **while** *exists*(n_c) **do**
(5) **foreach** *neighbor*(n_c) **do**
(6) Let be $n_j = neighbor(n_c)$
(7) **if** $d(n_j) > length(n_j, n_c) + d(n_c)$ **then**
(8) $d(n_j) = length(n_j, n_c) + d(n_c)$
(9) $U = U - n_j$
(10) **if** $n_f \notin U$ or $\min(length(n_i, n_j)) = \infty \, \forall \, n_i, \; n_j \in U$ **then**
(11) return $d(n_k) \, \forall \, n_k \in N$
(12) **else**
(13) $n_c = d(n_i) = \min(d(n_j)) \, \forall \, n_j \in U$

Algorithm 3: The safe routing algorithm.

The clustering approach obtains some probabilities that represent the estimation for specific crimes that were previously classified semantically. These probabilities define possible *hot spots* that represent events, which are directly associated with streets. In addition, the method classifies incoming data generated by the categorization task, searching patterns that were defined as words associated with a crime. It generates some predictive spatiotemporal patterns according to the indicated parameters. Thus, the goal is to search the crime probability for specific locations in a given route for having some criminal events when a user travels on that route. The probabilities are based on the following combinations:

let $P(\text{coordinate} = [x, y] \mid \text{event})$ be the probability of an event directly related to its location and let $P(day\_r\_time \mid \text{event})$ be the probability of an event, taking into account its temporariness. So, the computation of such probabilities is presented as follows;

$P(c \mid x)$ represents a *subsequent probability* and it is defined by the probability that an event occurs considering past events, in which $P(x)$ represents the *total probability*, which denotes the number of times that some given attributes appear in the events (e.g., theft);

$P(x \mid c)$ represents the *conditional probability*, which is denoted by the number of times when a target appears in each attribute (e.g., car), taking into account the total number that the target appears in all attributes;

$P(c)$ is defined as an *a priori probability*, which represents the number of times that a target appears in all events.

Events, attributes, and locations are identified and classified by the Bayes algorithm. The clusters are generated according to the defined patterns that were used in the training process. Particularly, the places are described as trusted or untrusted. The probability is used as a weighted value and it is normalized in a range of $[0, 1]$, where 0 represents the lowest and 1 the highest likelihood. Thus, these values are used in the route generation. An example of the clustering computation with the obtained likelihood is presented as follows.

> Classified_Event ("12/05/2014", "3", Monday, 1,2255928567); {"Event": "Theft", "ID_Coordinate": "30339694", "latitude": 19.4038744, "longitude": 99.150226, "Probability": "0.0009295401"}, {"diagnosed weight: 0.89"}

3.5. Stage 5: Generation of Safe Routes. The safe routes are visualized in the mobile mapping application, which was developed as a client, by using REST (Representational State Transfer). It is a web service technology that generates requests and the data parsing is received by the server. The spatial feature is supported by Open Street Maps [12]. The safe route is obtained by the adaptation of the Dijkstra algorithm, in which the nodes in the network are assigned as an average weight that was obtained from the number of crimes for a specific point or geographic area. The values generated from the Bayes algorithm reflect the probability of having or not an event such as "theft/crime."

Let n_s be the starting node called initial node and let d be the distance of node n_k from the initial node to n_1. Safe route algorithm assigns some initial distance values based on the weighted crime rate for these points. So, the weighted crime rate is computed by the number of complaint crime/theft occurrences in a specific point and time. Algorithm 3 describes the process for obtaining the safe routes.

Thus, this algorithm uses data that were sent by the mobile application (origin and destination points) in order to return a route that avoids locations where crime events have occurred. The confidence level is a metric that is computed by considering the number of crime incidents that have occurred in a specific point and time. This is used as a

weighted value for the safe routing algorithm. The displayed route is marked with coordinates that were returned by the algorithm, in order to visualize the route on the mobile mapping application.

The weights can be also relaxed by spatial and temporal values such as date (day, month, and year), location (point or area), and time (hour or period). In the interface level, users can configure the route search process by modeling some parameters as follows. If the route is generated by using either social, official, or integrated data source, then the search process also makes a difference between different transportations (e.g., walking or by car).

4. Experimental Results and Evaluation

The mobile application was implemented in Android 4.0, and the tests were performed in mobile devices. In this section, the results based on information of Mexico City are presented. The repository is composed of 5,441 events, which were recollected and processed from the tweets and correlated with official reports. In this dataset, a frequency table is generated, which indicates the number of times that each attribute appears, given events such as theft and crime.

The frequency values are defined as weights when deriving a safe route. All values are assigned to corresponding nodes in the network. In addition, the probability that an event occurs at some location and specific date is computed and stored in a vector table. This also allows a comparison of crime probabilities at different places. The sort of summarization provides a support to evaluate the probability of an event to occur at a specific location and/or particular time (e.g., the probability that a crime or theft can occur in "Avenue Eje Central on Monday"). Then, the probability that an event occurs at a given place and time is defined in the following example: "On Monday at 9:00 am in Iztapalapa."

$P(x) = P(\text{Monday}) = 220/2149$

$P(x \mid c) = P(\text{Monday} \mid \text{Theft}) = 18/218$

$P(c) = P(\text{Theft}) = 218/2149$

$P(c \mid x) = P(\text{Monday} \mid \text{Theft}) =$

$P(\text{MONDAY})$ times $P(\text{THEFT} \mid \text{Monday})/P(\text{Monday}) = 0.0818$

(1) The probabilities of some classified crime events are computed.

(2) The selected combinations (e.g., theft at a particular location and time) are defined. However, this problem is NP-hard; thus a semantic classification has been applied to restrict the search space and decrease the computational complexity.

(3) This provides the probabilities for all events that occur for each particular domain value.

An example to compute the probability for crime events, such as the probability that "a theft occurs at a given location on Friday, or a crime occurs on Tuesday at noon," is presented as follows.

$P(\text{Theft}) = 0.104$

$P(\text{Crime}) = 0.065$

$P(\text{Crime}(w)) = 0.017$

$P(\text{Crime}(p)) = 0.812$

...

$P(\text{Theft} \mid \text{day} = \text{Monday}) = 0.032$

$P(\text{Crime} \mid \text{day} = \text{Monday}) = 0.13$

$P(\text{Crime}(w) \mid \text{day} = \text{Monday}) = 0.017$

$P(\text{Crime}(p) \mid \text{day} = \text{Monday}) = 0.813$

$P(\text{Theft} \mid \text{hour} = 9) = 0.1$

$P(\text{Crime} \mid \text{day} = 9) = 0.05$

$P(\text{Crime}(w) \mid \text{hour} = 9) = 0.05$

$P(\text{Crime}(p) \mid \text{hour} = 9) = 0.8$

$P(\text{Theft} \mid \text{hour} = 1245612) = 0$

$P(\text{Crime} \mid \text{day} = 1245612) = 0.2$

$P(\text{Crime}(w) \mid \text{hour} = 1245612) = 0$

$P(\text{Crime}(p) \mid \text{hour} = 1245612) = 0.8$

$\Rightarrow$

$P(\text{Theft} \mid \text{Time} = 9 \,\&\&\, \text{day} = \text{Monday} \,\&\&\, \text{coordinate} = 1245612) =$

$P(\text{Theft}) \times P(\text{Theft} \mid \text{day} = \text{Monday}) \times P(\text{Theft} \mid \text{time} = 9) \times (\text{Theft} \mid \text{coordinate} = 1245612) = 0.104 \times 0.032 \times 0.1 \times 0 = 0$

The example was based on taking into consideration the defined points from a given area, it means that we have the estimation that an event occurs in x point at 11 am or in x point in several hours of the day. The probability that a theft occurs is increasing/decreasing depending on the hour and the day of week. In that case, the user can also ask for all possible combinations near his location. For instance, for y point, he wants to know the probability of suffering a theft at different hours of a day. So, he can know what event (type of crime) is the most common to occur for an x point at 6 pm.

Figures 5 and 6 show safe routes between two points for transient users, where the circles on the map represent untrusted points. The route was generated by avoiding these points; although this means that the path could be longer, it is the safer route. Moreover, Figure 6 depicts the generated safe route for the same points indicated in Figure 5, not only when the user is not a transient but also when the user is a driver. In this case, the route is different because the estimation is carried out only processing the event types related to drivers.

Finally, it is possible to generate routes considering the data reports of different periods of time (e.g., thefts and crimes occurred from April to July in 2015). The safe routing algorithm generates a specific route that avoids the points where theft events have occurred in the past. Nevertheless, the generated route can change if the user increases the time period for taking into account (e.g., from January to March in 2015) or including all temporal data available (e.g., from 2013 to 2015). It allows us to know what location is safer than others at specific day and hour in the same geographic area.

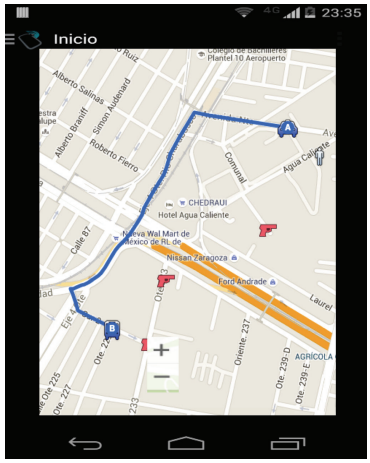

Figure 5: Suggested safe route for a transient.

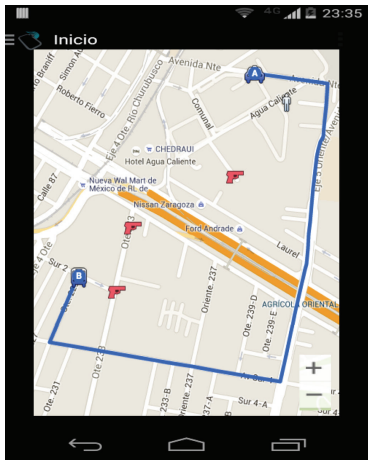

Figure 6: Suggested safe route for a user in a car.

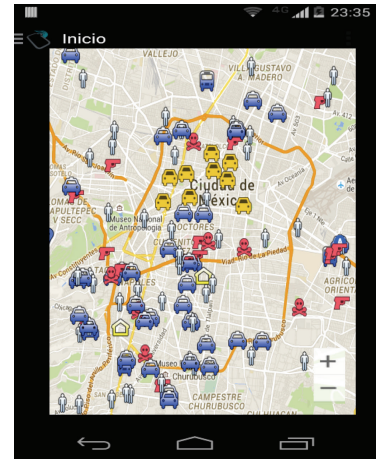

Figure 7: All events occurred from 2013 to 2015.

Figure 8: Crime map (made by volunteers).

Figure 7 depicts the events that occurred to transients and drivers from official and social sources; the icons represent theft of transients and drivers with and without violence, as well as theft of house.

On the other hand, the obtained results in Figures 5, 6, and 7 were compared with the Official Crime Map System (http://www.mapadelincuencial.org.mx/). Figure 8 only depicts events that volunteers marked as points where a crime or theft with violence occurred; the data difference is evident and the possibilities to configure the system are very limited in comparison with the proposed mobile information system (see Figure 9).

Thus, Figure 10 shows the hybrid map view from the web version. The events in yellow color were reported by the social network source and events with blue color represent the official source.

5. Conclusions and Future Work

In this paper, a hybrid approach for finding safe routes using semantic processing and classification algorithms, with data provided from a social network and the official crime reports, is presented. As a case study, a mobile information system was developed. It generates safe routes based on crime reports of Mexico City from large tweet repository and official databases. The data are semantically classified to determine whether the tweet describes a crime event or theft; in case of tweets which cannot be identified as crimes, they are evaluated by Bayes algorithm, which clustered them according to the contained description. Thus, the clusters are used to make prediction regarding the possibility that a crime can occur in a specific place and hour. The spatiotemporal analysis determined the location where the crime events occurred. Moreover, the confidence level of a location was defined and it was used as a parameter for computing the safer route.

The main contributions of this work are as follows: (1) the design of a hybrid approach based on semantic processing to retrieve crime data from a social network source; (2) the integration of crowd-sensed data with official government sources; (3) the validation of a tweet performed by comparing the sources, using k-fold cross validation; (4) the estimation

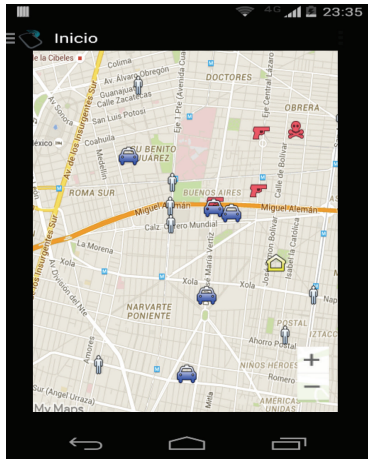

Figure 9: Theft/crime events that occurred in a specific time from the mobile information system.

Figure 10: The hybrid map view from the web version.

model based on the Bayes algorithm to obtain safe routes with data that were provided by the mobile device; and (5) the design of a mobile information system to generate safe routes.

According to the results of the estimation. the certainty degree is around 75% of effectiveness. It was tested by comparing areas with crime data, but the records were intentionally removed and original copy was kept. Thus, with the results of the estimation, a comparison with the original copy was performed. So, we found that the estimation has a performance of 75% for all the points of the data sample.

In addition, a metric to measure the confidence level or security for certain points and areas of Mexico City has been proposed. It allows finding safe routes, according to paths with a low crime rate. Moreover, the mobile application gathers long-term statistical data with almost real information from citizens, which are acting as sensors in the city. The results of the mobile system have been tested and compared with the Crime Map System.

Future works are oriented towards evaluating the cognitive perception of people, taking into consideration points or geographic places for finding comfortable routes. The sentiment analysis will be treated in order to incorporate this feature as a parameter in the computation of routes. Additionally, we are proposing the integration of our mobile application with the Mexican CCTV camera systems for sensing the dynamic of certain areas in the city. This contribution is focused on developing mobile information systems for routing and urban planning in city environments. Mobile applications are increasingly becoming essential for analyzing the urban dynamic of big cities. Thus, the appearance of the next generation of mobile information systems will be devised in real-time road network conditions. In addition, this generation is oriented towards improving the quality of human life for increasing the sustainability of the smart cities.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was partially sponsored by the Instituto Politécnico Nacional (IPN), the Consejo Nacional de Ciencia y Tecnología (CONACYT), and the Secretaría de Investigación y Posgrado (SIP) under Grants 20162006, 20161899, 20161869, and 20161611.

References

[1] H. Zhang, Y. Xu, and X. Wen, "Optimal shortest path set problem in undirected graphs," *Journal of Combinatorial Optimization*, vol. 29, no. 3, pp. 511–530, 2015.

[2] W. Templeton-Steadman and R. Williams, "Information delivery system and method for mobile appliances," US Patent App. 11/562,054, 2006.

[3] D. Reis, A. Melo, A. L. V. Coelho, and V. Furtado, "Towards optimal police patrol routes with genetic algorithms," in *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23-24, 2006. Proceedings*, vol. 3975 of *Lecture Notes in Computer Science*, pp. 485–491, Springer, Berlin, Germany, 2006.

[4] V. Ceikute and C. S. Jensen, "Routing service quality—local driver behavior versus routing services," in *Proceedings of the IEEE 14th International Conference on Mobile Data Management (MDM '13)*, vol. 1, pp. 97–106, IEEE, June 2013.

[5] Safety Apps, October 2015, http://www.csun.edu/police/safety-apps.

[6] E. Galbrun, K. Pelechrinis, and E. Terzi, "Urban navigation beyond shortest route: the case of safe paths," *Information Systems*, vol. 57, pp. 160–171, 2016.

[7] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Learning to detect patterns of crime," in *Machine Learning and Knowledge Discovery in Databases*, vol. 8190 of *Lecture Notes in Computer Science*, pp. 515–530, Springer, Berlin, Germany, 2013.

[8] 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, Part II, Lecture Notes in Computer Science, pp. 174–185, Springer, Berlin, Germany, 2014.

[9] L. Scott and N. Warmerdam, "Extend crime analysis with arcgis spatial statistics tools," *ArcUser Magazine*, 2005.

[10] M. Leitner, *Crime Modeling and Mapping Using Geospatial Technologies*, vol. 8, Springer, Dordrecht, The Netherlands, 2013.

[11] J. Kim, M. Cha, and T. Sandholm, "Socroutes: safe routes based on tweet sentiments," in *Proceedings of the 23rd ACM International Conference on World Wide Web (WWW '14)*, pp. 179–182, Seoul, South Korea, April 2014.

[12] M. Haklay and P. Weber, "Openstreetmap: user-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.

[13] J. M. Sánchez Bernabéu, J. V. Berná Martínez, and F. Maciá Pérez, "Smart sentinel: monitoring and prevention system in the smart cities," *International Review on Computers and Software*, vol. 9, no. 9, pp. 1554–1559, 2014.

[14] Wiki crimes, 2015, http://www.wikicrimes.org.

[15] T. Moon, S. Heo, and S. Lee, "Ubiquitous crime prevention system (UCPS) for a safer city," *Procedia Environmental Sciences*, vol. 22, pp. 288–301, 2014.

[16] H. Su, K. Zheng, J. Huang, H. Jeung, L. Chen, and X. Zhou, "Crowdplanner: a crowd-based route recommendation system," in *Proceedings of the 30th IEEE International Conference on Data Engineering (ICDE '14)*, pp. 1144–1155, IEEE, Chicago, Ill, USA, April 2014.

[17] V. Arnaboldi, M. Conti, and F. Delmastro, "Implementation of CAMEO: a context-aware middleware for opportunistic mobile social networks," in *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '11)*, pp. 1–3, Lucca, Italy, June 2011.

[18] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav, "Spatio-temporal-thematic analysis of citizen sensor data: challenges and experiences," in *Web Information Systems Engineering—WISE 2009*, vol. 5802 of *Lecture Notes in Computer Science*, pp. 539–553, Springer, Berlin, Germany, 2009.

[19] N. Powdthavee, "Unhappiness and crime: evidence from South Africa," *Economica*, vol. 72, no. 287, pp. 531–547, 2005.

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

[21] P. Gupta, G. N. Purohit, and A. Dadhich, "Crime prevention through alternate route finding in traffic surveillance using cctv cameras," *International Journal of Engineering and Advanced Technology*, vol. 2, no. 5, pp. 414–418, 2013.

[22] H. Chen, T. Cheng, and S. Wise, "Designing daily patrol routes for policing based on ANT colony algorithm," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. II-4/W2, pp. 103–109, 2015.

[23] D. Dzemydiene and E. Kazemikaitiene, "Ontology-based decision support system for crime investigation processes," in *Information Systems Development*, pp. 427–438, Springer, New York, NY, USA, 2005.

[24] Y. Chabot, A. Bertaux, C. Nicolle, and T. Kechadi, "A complete formalized knowledge representation model for advanced digital forensics timeline analysis," *Digital Investigation*, vol. 15, pp. 83–100, 2015.

[25] E. Hatcher, O. Gospodnetic, and M. McCandless, *Lucene in Action*, Manning Publications, Greenwich, Conn, USA, 2004.

[26] F. Mata-Rivera, M. Torres-Ruiz, G. Guzmán, M. Moreno-Ibarra, and R. Quintero, "A collaborative learning approach for geographic information retrieval based on social networks," *Computers in Human Behavior*, vol. 51, pp. 829–842, 2015.