World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:8, 2018

Road Traffic Accidents Analysis in Mexico City through Crowdsourcing Data and Data Mining Techniques

Gabriela V. Angeles Perez, Jose Castillejos Lopez, Araceli L. Reyes Cabello, Emilio Bravo Grajales, Adriana Perez Espinosa, Jose L. Quiroz Fabian

***Abstract*—**Road traffic accidents are among the principal causes of traffic congestion, causing human losses, damages to health and the environment, economic losses and material damages. Studies about traditional road traffic accidents in urban zones represents very high inversion of time and money, additionally, the result are not current. However, nowadays in many countries, the crowdsourced GPS based traffic and navigation apps have emerged as an important source of information to low cost to studies of road traffic accidents and urban congestion caused by them. In this article we identified the zones, roads and specific time in the CDMX in which the largest number of road traffic accidents are concentrated during 2016. We built a database compiling information obtained from the social network known as Waze. The methodology employed was Discovery of knowledge in the database (KDD) for the discovery of patterns in the accidents reports. Furthermore, using data mining techniques with the help of Weka. The selected algorithms was the Maximization of Expectations (EM) to obtain the number ideal of clusters for the data and k-means as a grouping method. Finally, the results were visualized with the Geographic Information System QGIS.

***Keywords*—**Data mining, K-means, road traffic accidents, Waze, Weka.

I. Introduction

ACCORDING to Mexican Institute on Statistics and Geography (INEGI) in 2015 were reported 12,321 road traffic accidents, causing 315 deaths in Mexico City (CDMX), placing this phenomenon in one of the five main causes of mortality in the city. The CDMX is one of the most populous cities in the world, according to the INEGI, has 8,918,653 population and more than 5 million registered cars in circulation [1], which generates a phenomenon of high daily mobility, long transportation times, vehicular congestion, environmental and road accidents, among others [2]. Nowadays the generation of data by crowd-sourcing is a way of obtaining massive information, through which data analysis can be done in the short time lapses and propose alternative solutions to negative transport problematics at low cost and in real time.

Gabriela V. Angeles-P., Jose Castillejos, Araceli L. Reyes-C. and Emilio Bravo-G. are with the Science and Technology College, Autonomous University of Mexico City, Mexico City, Mexico (e-mail: viridiana.angeles@estudiante.uacm.edu.mx, josecaslop@gmail.com, liliana.reyes@uacm.edu.mx, emilio.bravo@uacm.edu.mx).

Adriana Perez-Espinosa and Jose L. Quiroz-F. are with the Electrical Engineering Department, Metropolitan Autonomous University, Mexico City, Mexico (e-mail: pea@xanum.uam.mx, jlqf@xanum.uam.mx).

Project funded by SECITI/116/2017

Center for Research and education in transport and mobility CITMA S.C. Mexico

There are different studies on the behavior of traffic in several countries of the world for example, in Israel was carried out an research allowing identified some dangerous intersections and some locations in which accidents were concentrated on a daily in 2012 [3]. Another research was carried out in Buenos Aires City, Argentine which used Twitter and Waze data and comparing the traffic incidents based on an intelligent approach [4].

Researches like the aforementioned are carried out thanks to the ease with which the information is obtained. Nowadays, the social networks helps us to develop studies of different areas due to the large amount of information collected from users. In recent years, social networks such as Facebook, Twitter and Google+, among others, have become in sources of information for data mining due to its high growth [5], for example Twitter has more than 41 million users since June 2009 and continues its growth [6].

In [7] was demonstrated that the use of data mining algorithms, mainly grouping algorithms, allow identifying areas with strong vehicular traffic conflicts.

In this paper we present the descriptive study on the information of road accident reports of the Waze social network, which was performed using the Expectation Maximization (EM) and K-means algorithms and the Weka and QGIS tools. Section II shows the methodology used for data mining Knowledge Discovery in Database (KDD). Section III presents the results of the analysis performed with Weka and visualizations using maps elaborated in QGIS [8]. And finally, Section IV presents the conclusions.

II. Methodology

The methodology used was KDD for the treatment and analysis of the data. Following, the stages that conform the methodology are described [9]:

A. Objective

The main objective is to identify the area and thoroughfares of the CDMX, as well as the period of time in which a greater number of accidents were registered during 2016.

B. Data Collection

For the analysis of road accidents in the CDMX, the accident report data generated by Waze users was used. To

download the data, a Python script is used which it connects every 10 minutes time intervals with the Waze website [10]. The information was obtained in the JavaScript Object Notation (JSON) format, which was subsequently stored in a relational database consisting of six datatables. Particularly, the information by the accident reports was stored in the REPORTS datatable, where each record corresponds to a Waze report, which contains geographic coordinates, time, date, city to which it belongs, type and subtype of report, among others.

C. Extraction, Transformation and Loading of Data

From the datatable of REPORTS the attributes shown in Table I were taken from those reports whose type was ACCIDENT and whose coordinates belong to the CDMX in 2016. For this cleaning of the coordinates a cut was made with the help of the System of Geographic Information QGIS, using the WGS 84 (World Geodetic System 84) coordinate system with UTM projection area 14, which corresponds to the CDMX.

TABLE I
ATTRIBUTES OF THE ACCIDENT REPORTS USED FOR THE ANALYSIS

Attribute	Description
x	Longitude.
y	Latitude.
type	Report type.
subtype	Report subtype.
time	Time when the report was registered. (hh:mm:ss)
date	Day on which the report was recorded. (YYY:MM:DD)
city	Borough in which the user was at the time of the accident report.

To identify the reports corresponding to the same accident, a Python script was made to obtain a representative for each set of reports associated with the same accident. According to *Road Safety Theory*, it was considered that two reports belonged to the same accident if they were of the same sub-type, they were at a distance of no more than 150 meters and in a range of 20 minutes. We used (1) in order to obtain the distance between two reports with coordinates (ϕ_1, λ_1), (ϕ_2, λ_2) respectively. This equation is known as the Haversine equation [11].

$$\begin{gathered}\Delta\phi = \phi_2 - \phi_1 \\ \Delta\lambda = \lambda_2 - \lambda_1 \\ a = \sin^2(\frac{\Delta\phi}{2}) + \cos(\phi_1)\cos(\phi_2)\sin^2(\frac{\Delta\lambda}{2}) \qquad (1) \\ r = 637100 \text{ meters (radius of the Earth)} \\ d = 2 * r * \arcsin(\sqrt{a})\end{gathered}$$

Using the Python script, two datatables were created for each month, one of them stores the representatives of each accident and the other stores the associated reports. Once these datatables were stored, we run a query to each table of representatives to obtain the month with the highest number of accidents, which was the month of June. Fig. 1 shows comparative chart between the number of reports and the representatives of the reports per month.

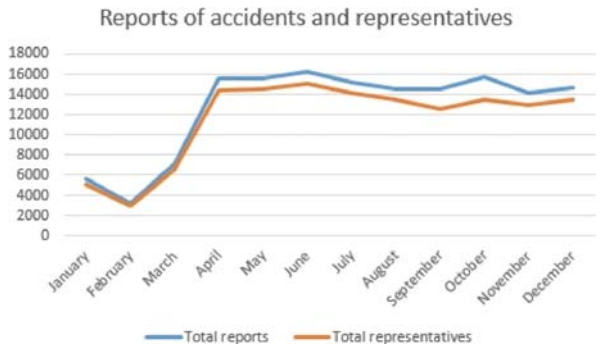

Fig. 1 Number of reports of accidents and representatives per month (2016)

Finally, in the datatable of representatives of June, two columns were added: one to indicate the day in which the accident was reported and the other to indicate the hour without contemplating minutes and seconds, with the purpose of performing the analysis in time intervals.

D. Data Mining

In this stage, the selection and application of algorithms was carried out to obtain the patterns that help to make decision through the Weka tool and a script in R.

We used the K-means clustering algorithm [12], [13] which generated the groups that have characteristics in common with the purpose of explaining the behavior of the data within the study area and thus identify the areas and roads with the highest number of accidents. The ideal number of seeds indicated to the K-means algorithm was calculated using the Expectation Maximization (EM) algorithm [13].

E. Interpretation and Evaluation

To visualize the results, the K-means algorithm programmed in R was executed and a new column is added to the datatable of representatives which indicates the cluster number to which the representative belongs.

With the information obtained in the previous process, QGIS was used to generate the map and then identify the zones and roads with the greatest number of accidents in the CDMX.

III. RESULTS

The number of clusters obtained by the EM algorithm for 9,946 accident representatives in June 2016, is shown in Table II.

For a better grouping of data, the clusters with 1, 2 and 3% were eliminated, since statistically these clusters do not represent a significant value of the data set. Thus, the number of clusters provided to the K-means algorithm was 10. Tables III and IV show the results obtained by Weka.

In Table III, we can see that of the 9,946 reports of accident representatives, the largest number of these is located in the borough of Miguel Hidalgo and the most common type of accident was ACCIDENT_MINOR. As shown in Table IV,

TABLE II
NUMBER OF CLUSTERS OBTAINING BY EM ALGORITHM

Clustered Instances		
0	319	(6%)
1	420	(8%)
2	61	(1%)
3	340	(7%)
4	589	(11%)
5	35	(1%)
6	644	(13%)
7	149	(3%)
8	536	(10%)
9	26	(1%)
10	235	(5%)
11	82	(2%)
12	148	(3%)
13	357	(7%)
14	134	(3%)
15	595	(12%)
16	454	(9%)

TABLE III
FULL DATA

Attribute	Full Data (9946.0)
x	-99.1678
y	19.3967
city	Miguel Hidalgo
subtype	ACCIDENT_MINOR
day	15
hour	14

the borough of Miguel Hidalgo is in cluster 2 with 1485 data and in cluster 5 with 1706, both with the same type of accident, also we can observed that the peak traffic time is at 12 a.m.. This borough has a total of 3191 representatives for ACCIDENT_MINOR.

Likewise, the results obtained show that the second conflict zone is the borough of Benito Juárez with a total of 1451 representatives for ACCIDENT_MINOR.

Fig. 2 shows the results obtained after applying K-means using the R script, where a new column was added. This column indicates the cluster number to which each accident representative belongs. This information was mapped in the QGIS system.

Subsequently, the density of accidents representatives of each polygon is calculated, where the sum of kilometers of the sections of roadways that each cluster has is considered. This calculation was made in the QGIS tool. As can be seen in Table V, Cluster 5 is the one with the highest density in terms of accident representatives per kilometer of roads.

In order to identify the roads with the largest number of accidents representatives, we focused in the cluster 5, which cover some parts of the boroughs: Miguel Hidalgo, Azcapotzalco, Cuauhtémoc, Álvaro Obregón and Benito Juárez.

In Cluster 5, we reproduce the same process and we obtained the ideal number of clusters which was five (see Table VI) that consist of 2559 records. However, we decide remove the cluster smaller or equals to 9%, and execute the K-means algorithm with the rest of clusters (3 clusters). Tables VII and VIII shows the result of the execution.

TABLE IV
DENSITY OF ACCIDENT REPRESENTATIVES PER CLUSTER

Final cluster centroids:

Attribute	Cluster 0 (791.0)	Cluster 1 (1000.0)	Cluster 2 (1485.0)
x	-99.1224	-99.1588	-99.1888
y	19.4186	19.3424	19.4069
city	Venustiano Carranza	Coyoacán	Miguel Hidalgo
subtype	OTHER	OTHER	ACCIDENT_MINOR
day	21	7	26
hour	20	12	12

Attribute	Cluster 3 (369.0)	Cluster 4 (867.0)	Cluster 5 (1706.0)
x	-99.1303	-99.2055	-99.1911
y	19.4508	19.3618	19.4114
city	Gustavo A. Madero	Álvaro Obregón	Miguel Hidalgo
subtype	ACCIDENT_MAJOR	ACCIDENT_MAJOR	ACCIDENT_MINOR
day	6	16	7
hour	21	19	12

Attribute	Cluster 6 (1451.0)	Cluster 7 (644.0)	Cluster 8 (836.0)
x	-99.1685	-99.0688	-99.1598
y	19.3802	19.3593	19.4174
city	Benito Juárez	Iztapalapa	Cuauhtémoc
subtype	ACCIDENT_MINOR	ACCIDENT_MAJOR	ACCIDENT_MAJOR
day	17	18	18
hour	16	20	9

Attribute	Cluster 9 (797.0)
x	-99.1213
y	19.4607
city	Gustavo A. Madero
subtype	ACCIDENT_MINOR
day	13
hour	10

Clustered Instances (percentage split)

0	418	(8%)
1	557	(11%)
2	773	(15%)
3	163	(3%)
4	442	(9%)
5	876	(17%)
6	717	(14%)
7	356	(7%)
8	409	(8%)
9	413	(8%)

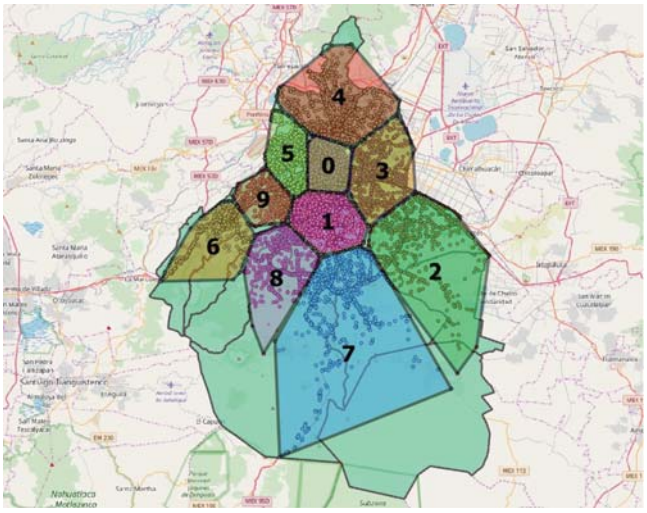

Fig. 2 Distribution of accident representatives in the ten clusters using QGIS

TABLE V
DENSITY OF ACCIDENT REPRESENTATIVES PER CLUSTER

Cluster	Number of accidents representatives	Kilometers of roads	Density
0	2122	630.95	3.3
1	2215	1104.22	2.0
2	911	2550.98	0.4
3	1546	1356.89	1.1
4	1371	1777.07	0.8
5	2559	635.94	4.0
6	630	680.23	0.9
7	1350	2245.16	0.6
8	1257	1049.83	1.2
9	1109	603.99	1.8

TABLE VI
NUMBER OF THE CLUSTER OBTAINED BY EM ALGORITHM FOR CLUSTER 5

Clustered Instances		
0	231	(27%)
1	302	(35%)
2	80	(9%)
3	20	(2%)
4	238	(27%)

TABLE VII
FULL DATA

Attribute	Full Data (1688.0)
x	-99.1927
y	19.4225
city	Miguel Hidalgo
subtype	ACCIDENT_MINOR
day	15
hour	13

TABLE VIII
RESULT OF K-MEANS ALGORITHM FOR THE CLUSTER 5 USING WEKA

Final cluster centroids:

Attribute	Cluster 0 (599.0)	Cluster 1 (533.0)	Cluster 2 (556.0)
x	-99.1937	-99.1895	-99.1948
y	19.424	19.4146	19.4266
city	Miguel Hidalgo	Miguel Hidalgo	Miguel Hidalgo
subtype	ACCIDENT_MINOR	ACCIDENT_MINOR	ACCIDENT_MINOR
day	8	24	15
hour	10	11	20

Clustered Instances (percentage split)

0	307	(35%)
1	275	(32%)
2	289	(33%)

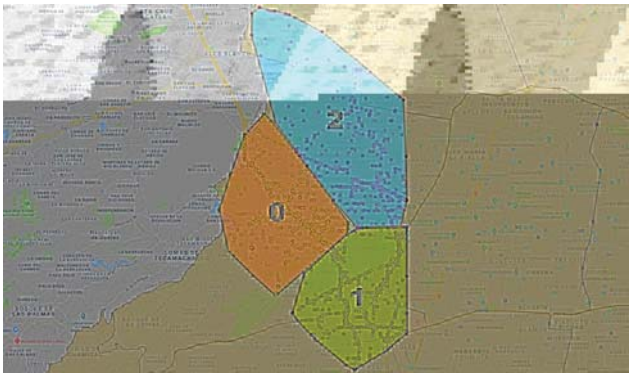

Fig. 3 Distribution of the accident representatives in the three clusters for Cluster 5, using QGIS

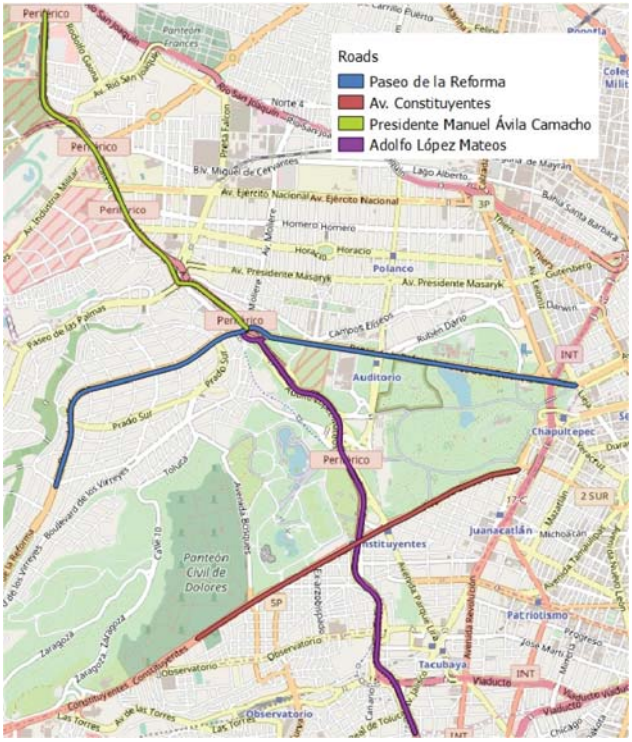

Fig. 4 Roads with the major amount of accidents

Fig. 3 shows the three clusters in the area with largest amount of accidents representatives.

Watching the Cluster 5 in QGIS with the help of plugin (s) OpenStreetMap and Google Maps we identify some avenues and streets with the major amount of accidents, as shown in Fig. 4.

IV. CONCLUSION

In this paper, we identify the areas and roads that present conflicts of accidents within CDMX, also we can see that the appropriate use of techniques of data mining and a methodology allows obtain useful information for the decision making. The results show that the roads with the largest number of accident representatives in the CDMX are: Paseo de la Reforma, Presidente Manuel Ávila Camacho, Av. Constituyentes and Adolfo López Mateos, i.e. these roads present the major amount of accidents which coincide with the reality due they have the highest vehicular traffic. It is worth mentioning that the accident reports were processed with distance algorithms allowing to determinate when two or more reports belong to the same event (accident) in order to eliminate data repeated.

The results obtained in this paper could be contribute in the decision making of experts in the transport areas, whose will decide the appropriate actions to be carried out in the zones with vehicle accidents conflicts.

REFERENCES

[1] INEGi. http://www.inegi.gob.mx., September 2016.
[2] Manjarrez, P. L., Vadillo, I. G. R., & Grajales, E. B. (2000). *Transporte urbano, movilidad cotidiana y ambiente en el modelo de ciudad sostenible: bases conceptuales.* Plaza y Valds, SA de CV.

[3] Fire, M., Kagan, D., Puzis, R., Rokach, L., & Elovici, Y. (2012, November). *Data mining opportunities in geosocial networks for improving road safety.* In Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of (pp. 1-4). IEEE.
[4] Caimmi, B., Vallejos, S., Berdun, L., Soria, Á., Amandi, A., & Campo, M. (2016, June). *Detección de incidentes de tránsito en Twitter.* In Biennial Congress of Argentina (ARGENCON), 2016 IEEE (pp. 1-6). IEEE.
[5] Mining, D., & Kulikov, O. (2009). *Data Mining Social Networks*.
[6] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). *What is Twitter, a social network or a news media?*. In Proceedings of the 19th international conference on World wide web (pp. 591-600). ACM.
[7] R. F. Estrada-S, A. Molina, A. Perez-Espinosa, A. L. Reyes-C, J. L. Quiroz-F, and E. Bravo-G, *Zonification of Heavy Traffic in Mexico City.* in Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2016, p. 40.
[8] QGis, D. T. (2011). Quantum GIS geographic information system. Open source geospatial Foundation project, 45.
[9] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *The KDD process for extracting useful knowledge from volumes of data.* Communications of the ACM, 39(11), 27-34.
[10] Waze Web. https://www.waze.com/es-419/livemap
[11] Shumaker, B. P., & Sinnott, R. W. (1984). *Astronomical computing: 1. Computing under the open sky. 2. Virtues of the haversine.* Sky and telescope, 68, 158-159.
[12] López, J. M. M., & Herrera, J. G. (2006). *Técnicas de Análisis de Datos Aplicaciones Prácticas utilizando Microsoft Excel y Weka.* Universidad Carlos III de Madrid. Pag, 99, 125.
[13] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.