Data Mining Opportunities in Geosocial Networks for Improving Road Safety

Michael Fire, Dima Kagan, Rami Puzis, Lior Rokach, and Yuval Elovici
Telekom Innovation Laboratories and Information Systems Engineering Department,
Ben-Gurion University of the Negev, Beer-Sheva, Israel
Email: {mickyfi,kagandi,puzis,liorrk, elovici}@bgu.ac.il

***Abstract*—Traffic measurements, road safety studies, and surveys are required for efficient road planning and ensuring the safety of transportation. Unfortunately, these methods can be cumbersome and very expensive. In this paper we point out a source of transportation information that is based on collaborative community-based navigation applications, such as Waze. Partial and anonimized information publicly exposed by Waze through their application provides valuable information that can significantly ease the future of transportation studies. Moreover, we show that Waze user reports may expose locations plagued with accidents but in lacking police coverage. This knowledge may help police departments to improve road safety by relocating the police units to these locations. Lastly, the data discussed in this paper connects transportation and road safety research to location based services and social network platforms.**

***Keywords*-Accident prevention;Road Safety;Geosocial Networks;Data Mining;Waze.**

I. Introduction

In recent years, due to the increasing popularity of smart devices that contain positioning components, many geosocial networks have emerged. Geosocial networks, such as FourSquare[1], Facebook Places[2], Google Latitiude[3], Mobli[4] and Waze[5], already have millions of active users. These networks provide their users with the ability to associate content with geographic locations. Usually, these geosocial networks use locations reported by users in order to provide various location-based services, such as geotagging [1] and friend tracking [2].

In addition, geosocial networks can use their enormous dataset of collected user locations to offer their users crowdsourcing location based services, like traffic monitoring[5] and travel recommendation[6]. In this paper, we focus on the the Waze geosocial network which provides a community-based turn-by-turn navigation application. Based on the driving speed of its, users Waze infers the most up-to-date road conditions in order to recommend the best driving routes. Waze users can also report accidents, traffic jams, speed traps, and nearby police units. All this information is publicly available and can be obtained by web scrapping. Waze makes significant efforts to preserve the privacy of its users. Privacy concerns and possible implications are out of the scope of this paper. However, available information, in its aggregated form, can be very useful for many tasks, such as improving road safety and deployment of police units, just to name a few. In this study, we show how information created by the Waze's user community helps to identify dangerous intersections and locations that are plagued on a daily basis by reoccurring accidents. The "police nearby" reports may be used to study the deployment of police units and their effectiveness, measure the police response time to an accident, and more. We also show that there are in fact areas with a relatively high police presence yet without any reported accidents.

The remainder of this paper is organized as follows. In Section II, we give a brief overview of studies in the field of geosocial networks and transportation networks. In Section III, we describe the methods and experiments that were used in our study. In Section IV, we present the obtained results. In Section V, we present our conclusions from this study and lastly, In Section VI, we offer future research directions.

II. Related Work

A. Geosocial Networks

Geosocial networks are relatively new research field that has only been development in the last few years. In 2010, Lee and Sumiya [3] used data collected from Twitter to develop the geosocial event detection system. In 2011, Lindqvist et al. [4] examined the reasons on which people use location sharing applications. In the same year, Noulas et al. [5] studied the user behavior in the FourSquare geosocial network in an effort to better understand human mobility patterns. Cho et al. [6] also studied human mobility patterns by using data from BrightKite, Gowalla and cellphone location data. In 2012, Sadilek et al. [7] used 2.5 million geo-tagged Twitter messages and modeled the spreading of infectious diseases throughout a real-world population.

B. Transportation Networks

Aside from increased usage of location based services, the widespread use of cellular phones can provide us with data for a comprehensive study of travel behavior based on the

[1]https://foursquare.com
[2]https://www.facebook.com/about/location
[3]https://www.google.com/latitude
[4]http://www.mobli.com
[5]http://www.waze.com
[6]http://www.cityowls.com

mobility patterns of randomly selected mobile phones in the transportation system [8], [9]. The network in [9] was created for the National Israeli Transportation Planning Model. In such networks, hourly flows and congested travel times are usually obtained using traffic assignment models [10], [11], [12] from the road network topology and estimated Origin-Destination matrices. Using data from Waze, it is possible to obtain high quality flow and travel time data in order to calibrate and/or validate traffic assignment models.

III. Methods and Experiments

In order to conduct our study, we collected data from the Waze geosocial network by using a dedicated Web crawler which collected accidents reports, police units nearby reports, traffic jams reports, speed traps reports, and traffic data from the Waze application. The Waze geosocial network has more than 1.1 million active users in Israel and may provide good coverage of main areas in Israel [13]. Therefore, our Web crawler focused on user's reports in Israel and collected report data from 31 different days between the months of May and June 2012.

The collected data included 5,369 accident reports and 29,789 nearby police reports. For every report, we collected the report location and the time of the report. To to avoid duplicate reports of the same event, we partitioned the monitored geographical region into a grid of cells of around 0.25 square kilometres per areal unit. Next, for each areal unit, we counted how many accidents and nearby police reports were reported. Then, for each area unit, we gave a score of between 0 and 31, according to the number of different days in which accidents occurred at the areal unit. We also repeated the same process and gave similar scores according to the number of police nearby reports in each areal unit. We used the Google Earth software[7] to visualize heat maps of the areal units that received high scores and were considered more dangerous than other area units. Using the heat maps, we manually identified each one of these area units and checked whether the area unit contained of an intersection. We also used the heat maps to identify area units that received a relatively high score due to reported police nearby user reports and which considered to be more monitored areas. Moreover, by using plot chart and linear regression techniques, we also identified area units with anomalies. For example, we identified area units with relatively high police activities but without a single reported accident.

In this study, we also checked how many area units with reported accidents did not contain any police nearby reports. Additionally, we also evaluated the average response time in which police units arrive to the scene of reported accidents. These calculations were carried out by randomly choosing 1,000 reported accidents and checking whether if, after the initial report of the accident, a police nearby reported followed. We then calculated the time elapsed starting from the initial accident report until the first police nearby report in a radius of 0.25 kilometres from the accident reports. In order to avoid accidents that were already reported, we only randomly chose accident reports in locations where no other accident was reported in at least one hour before in a radius of 0.25 kilometres from the location of the reported accident.

[7]http://earth.google.com

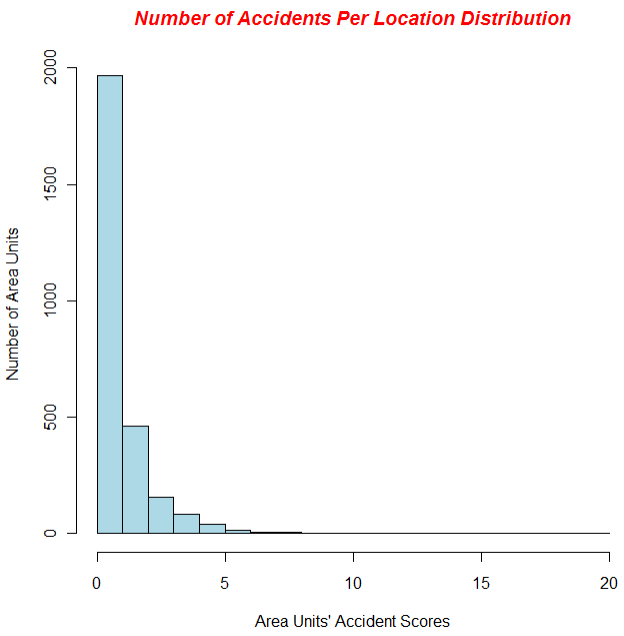

Figure 1. Distribution of the accidents score across area units only a few area units have many reoccurring accidents on a daily bases.

IV. Results

In the end of the process, we discovered 2,743 area units with at least one accident, 312 area units with a score of equal or greater than 3, and 19 area units with a score equal or greater than seven (see Figure 1). Moreover, as a result of the data analysis, we also identified 579 different locations which had at least five reoccurring accidents during the 31 days (see Figure 1). These 579 locations were responsible for 5,156 reported accidents, more than half of all the reported accidents. We also identified 3,555 locations where police nearby were reported at least 15 times.

By using heat maps together with the Google Earth software, we identified geographic locations which received the highest reoccurring accidents and police nearby scores (See Figure 2). By manually reviewing the locations of the reported accidents, we noticed that at least 75% of the twenty area units which received the highest score due to reoccurring accidents were intersections. Moreover, 40% percent of the top 20 area units which received the highest

Figure 2. Satellite image from Google Earth combined with accident and police report heatmap - area units with high accident scores (marked as red and white ellipses) and high police scores (marked as purple circles). The area units with high accidents score are mainly intersections.

score due to police nearby reports were also intersections. These results are consistent with known statistics about traffic accidents which evaluate that around 50% of urban crashes and 30% of rural crashes occur at intersections [14].

Next, we used plot chart and linear regression methods to find the correlation between the number of accident reports and the number of police nearby reports in each area unit. Using linear regression, we deduce the following regression equation

$$score_{accidents} = 0.132 score_{police} + 0.19$$

where $R^2 = 0.198$, and $p-value = 2.2e-16$. Furthermore, using the plot chart (see Figure 3), we noticed area units with anomalies between their police score and their accidents score. These anomalies assisted in identify the geographic areas with a relatively high number of accidents and a low number of police nearby reports, or areas with many police nearby reports but without any accidents. We also calculated the average reported time between accident reports and the police nearby reports in the same area unit. The results indicate that out of 1,000 reported accidents, only 321 were followed by police nearby reports and the average police response time were 1,720 seconds with $Median = 1,172$.

V. Conclusions

In this study, we presented our initial methods and results in a study of Geosocial networks with the aim of improving road safety. According to our results, it is possible to identify the intersections and areas which are more likely to suffer from accidents by analyzing user accidents reports in Waze. As expected, most of the area units that received the highest accident and police nearby scores were intersections. Moreover, we discovered 579 locations that were responsible for more than half of the reported accidents. We believe that improving road safety and increasing law enforcement in these locations can save lives. Furthermore, by analyzing the user police nearby reports, and by using a plot chart combined with linear regression, we could identify the area units in which there are many accidents and not enough police coverage, or high police coverage but without any reported accidents. Identifying these problematic areas can assist in improving the police unit deployment and help prevent future accidents.

Using the presented methods, we also estimated that the percentage of accidents that did not include police intervention is 67.9%, and the average response time of the police to an accident is 28.66 minutes. Using the presented methods of timing police arrival time to an accident can assist in identifying problematic unit areas in which police take a long time to respond to.

VI. Future Directions

This study is an initial study which has many possible future directions. A possible direction is to measure the

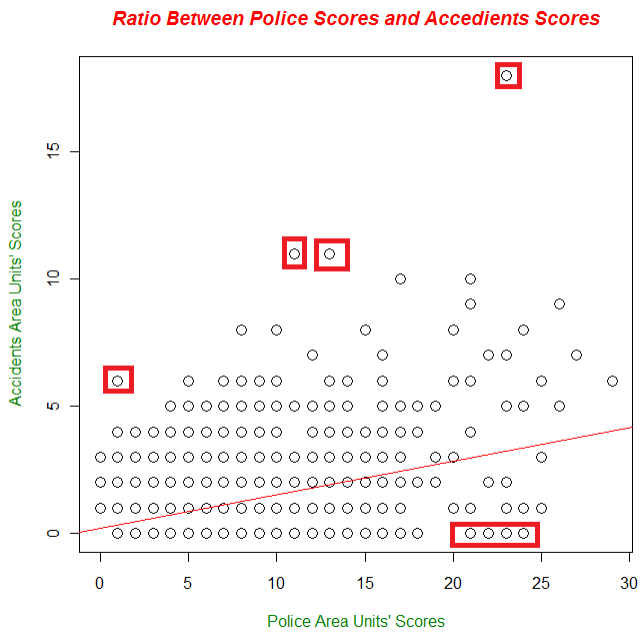

Figure 3. The ratio between an accident's area unit score and the police area unit score. The circles marked in red indicate area units with anomalies. For example, the four lower marked circles indicate area units with a high police nearby score, but without any accidents.

police influence on accident rates for different types of roads. Other possible direction is to measure the influence of speed cameras on accident rates. Additional possible direction is to use anomaly detection algorithms in order to automatically identify area units that have many reported accidents but insufficient police coverage. Another possible future direction is to use machine learning techniques to understand the causes that make some intersections more dangerous than others.

REFERENCES

[1] M. Egenhofer, "Toward the semantic geospatial web," in *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*. ACM, 2002, pp. 1–4.

[2] S. Consolvo, I. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge, "Location disclosure to social relations: why, when, & what people want to share," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 81–90.

[3] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*. ACM, 2010, pp. 1–10.

[4] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, "I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application," in *Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM, 2011, pp. 2409–2418.

[5] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in foursquare," *ICWSM?11*, 2011.

[6] E. Cho, S. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[7] A. Sadilek, H. Kautz, and V. Silenzio, "Modeling spread of disease from social interactions," in *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.

[8] S. Bekhor, Y. Cohen, and C. Solomon, "Evaluating long-distance travel patterns in israel by tracking cellular phone positions," *Journal of Advanced Transportation*, 2011.

[9] Y. Gur, S. Bekhor, C. Solomon, and L. Kheifits, "Intercity person trip tables for nationwide transportation planning in israel obtained from massive cell phone data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2121, no. -1, pp. 145–151, 2009.

[10] D. Bertsekas and E. Gafni, "Projection methods for variational inequalities with application to the traffic assignment problem," *Nondifferential and Variational Techniques in Optimization*, pp. 139–159, 1982.

[11] R. Jayakrishnan, W. Tsai, J. Prashker, and S. Rajadhyaksha, "Faster path-based algorithm for traffic assignment," *Transportation Research Record*, pp. 75–75, 1994.

[12] R. Puzis, Y. Altshuler, Y. Elovici, S. Bekhor, Y. Shiftan, and A. Pentland, "Augmented betweenness centrality for environmentally-aware traffic monitoring in transportation networks," *Journal of Intelligent Transportation Systems (to appear)*.

[13] T. Hoffman, "Waze navigates to 20 million," http://www.globes.co.il/serveen/globes/docview.asp?did=1000763278&fid=1724, 2012.

[14] N. C. H. R. Program, "A guide for addressing unsignalized intersection collisions," 2008.