Knowledge Tier Platform for Graph Mining in (Smart) Cities

Miguel Nuñez-del-Prado Edgardo Bravo Miguel Sierra
Isaias Hoyos Miguel Canchay
Universidad del Pacfico
Av. Salaverry 2020
Lima - Peru
{m.nunezdelpradoc,er.bravoo,l.sierraflores,i.hoyoslopez,cacnayd}@up.edu.pe

Abstract

In the present effort, we present a knowledge tier platform to collect information from cities in a form of graphs. This platform enables people to share the information of the area where they live allowing them to inform about pollution, crime levels, traffic jams, streets topology, commerces, markets, etc. The main objective is to provide information, stored in Elastic about a city to find spatio-temporal patterns using Graph Mining techniques based on Apache Spark GraphX.

1 Introduction

In the last years, we have seen the explosion of data from on-line activity, user content generated, health, scientific computing, mobile phones activity, etc. This data increments due to the daily transaction of people in urban centers and still grows. By 2030, 60% of the worldwide population will live in cities appearing 27 megacities greater than 10 million inhabitants (Chourabi et al., 2012). One technique to solve this problem is to generate new instruments for gathering and combining information continuously (Hernández-Muñoz et al., 2011). Consequently, there is an increment of collaborative platforms to collect data. For instance, a platform, called *WebCar*, to collect GPS data from vehicles to estimate traffic in a city (Lo et al., 2008). In the field of human health, *Psychlog* (Gaggioli et al., 2013) is a mobile phone platform designed to collect users psychological, physiological, and activity information for mental health research relying on a self-report questionnaire. The last example developed an Internet site and implemented the collection of data for a multicenter study of ethical decision-making (Avidan et al., 2005).

In the present effort, we present a knowledge tire platform to collect information on cities in a form of graphs. This platform enables people to share the knowledge of the area where they live allowing them to inform about pollution, crime levels, traffic jams, streets topology, commerces, markets, etc. The primary objective is to provide information about the city to find spatio-temporal patterns using Graph Mining techniques.

The present paper is organized as follows. Section 2 introduce some basic concepts, while Section 3 describes the platform architecture. Sections 4 and 5 show some preliminary results and present the discussion about the platform. Finally, Section 6 concludes the paper and presents future works.

2 Basic Concepts

In the current section, we introduce some basic concepts, such as graph, knowledge tiers and Spark for describing the platform.

2.1 Graph

A graph is a mathematical structure composed of *vertices*, *nodes* or *points*, which are connected through *edges*, *lines* or *arcs* as depicted in Figure 1. A graph ($G = (V, E)$) is composed of a set of V vertices and E edges. in our context this structure allows us to represent street intersections as geo-referenced nodes and roads as edges.

2.2 Haversine distance

The Haversine distance (Shumaker and Sinnott, 1984) computes the shortest distance between two points represented by latitude and longitude in the earth's surface.

$$\begin{aligned} d_{lon} &= lon_2 - lon_1 \\ d_{lat} &= lat_2 - lat_1 \\ a &= (sin(\tfrac{dlat}{2}))^2 + cos(lat_1) \times \\ &\quad cos(lat_2) \times (sin(\tfrac{dlon}{2}))^2 \\ c &= 2 \times atan2(\sqrt{a}, \sqrt{1-a}) \\ d &= R \times c \end{aligned} \tag{1}$$

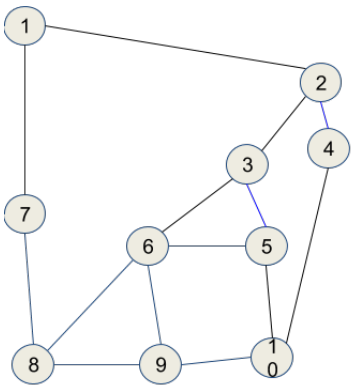

Figure 1: Example of a graph.

Where lat,lon and R are the latitude, longitude and radius of the Earth, respectively.

2.3 Knowledge Tiers

Since we are able to model street network of a city in the form of a graph. Note that each node and edge could have a weight representing different phenomena of a city, such as: (1) congestion, (2) crime, (3) pollution, (4) population density, (5) urban transportation, (6) subway network, etc. Thus, for each phenomenon, we have a graph modeling this particular fact. Finally, we can stack each node as depicted in Figure 2 to have a knowledge stack.

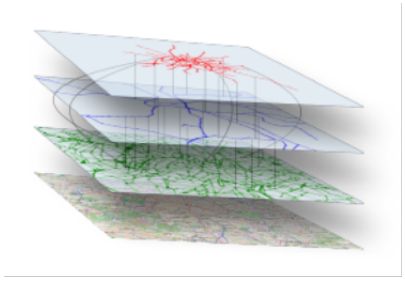

Figure 2: Knowledge tiers[1].

2.4 Apache Spark

Apache Spark is an open source cluster developed by the University of Berkeley. Then, the code was maintained by Apache Software Foundation. Apache provides distributed computation taking charge of task dispatching, scheduling, and basic I/O functionalities. These functionalities are available through Java, Python, Scala and R interfaces.

[1]Fereshteh ASGARI, Inferring User Multimodal Trajectories from Cellular Network Metadata in Metropolitan Areas

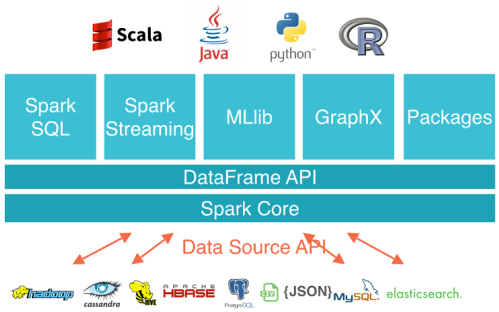

Figure 3: Spark framework.

As shown in Figure 3 Apache Spark provides at the top of its framework a tool for graph mining call *GraphX*[2]. This API allows parallel graph computation and integrates tools for extraction, transformation and load. More detail about the architecture as well as the capabilities of Spark is given in the next section.

3 System Overview

In the current section, we describe the architecture of our platform. As illustrated in Figure 4, our platform allows collecting data from Open Street Maps [3] (*OSM*) to build the graph representing streets and intersections in the form of a comma separated values *CSV* files. Then, these *CSV* files are stored in a NoSQL database. We use *Elasticsearch*[4] as NoSQL database due to its scalable, flexible and performant search and analytics engine (*c.f.,* Figure 5).

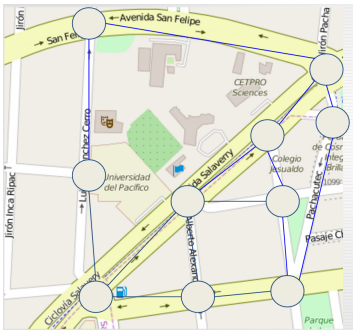

Figure 4: Example of a graph over streets.

Once data is saved in the NoSQL database, we are able to analyze the knowledge tiers represented and combined in form of graphs trough Spark

[2]GraphX: `http://spark.apache.org/graphx/`
[3]OSM: `https://www.openstreetmap.org/`
[4]Elasticsearch :`https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html`

GraphX as depicted in Figure 5. For instance, with this platform, we could optimize supply chain in cities minimizing cost, avoiding traffic jams and passing over low crime rate zones. We can also discover spatial patterns to understand common features of high crime rates areas in a city. All these analytics could be performed using programming languages such as: *Scala*[5], *Java*[6], *Python*[7] or *R*[8].

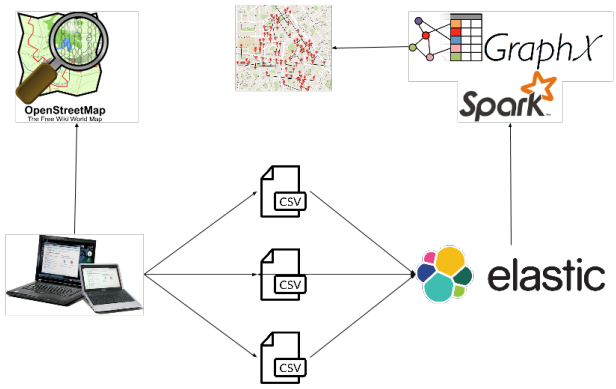

Figure 5: Overview of the Knowledge Tier Platform.

Finally, we implement a Python script to visualize the result of the pattern mining process using *Google Maps*[9]. In the next section, we present some preliminary visualization of graphs stored in the platform.

4 Preliminary results

In this section, we present some preliminary results, of the Knowledge Tier Platform, about data gathering, and visualizations.

Concerning the data collection, we have done two campaigns to collect data from streets and tweets in Lima, Peru. The former campaign was performed in the month of May collecting $\tilde{1}00\ 000$ and $\tilde{4}20\ 000$ nodes and vertices, respectively. The latter campaign was carried on between the months of April to Jun obtaining $\tilde{7}$,1 millions of geolocated tweets.

About visualization, the platform allows to plot a graph over a cartography, where the nodes are placed in the intersections of the streets and the

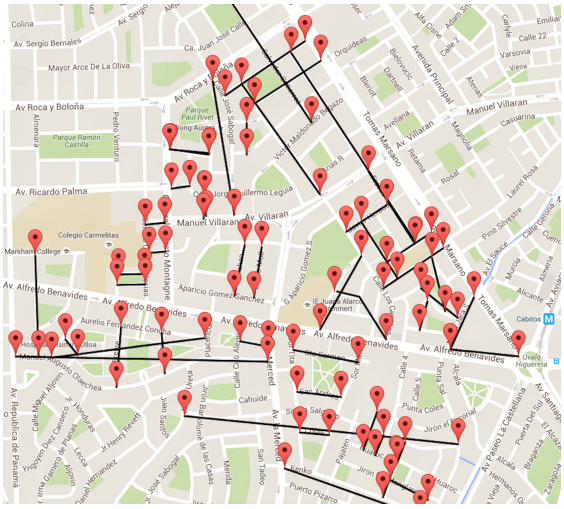

Figure 6: Visualization of the graph over the streets.

edges model the streets connecting nodes or intersections as shown in Figure 6.

Figure 7: Visualization of the heatmap of tweets over the cartography.

Another possibility of visualization are *Heatmaps*. In our case, *Heatmaps* are generated based on nodes weight. For example, Figure 7 presents a *Heatmap* of collected tweets in the platform. It is worth noting that tweets are affected to the nearest node relying on latitude and longitude of both nodes and tweets. We use as distance function the *Haversine* function (*c.f.*, Subsection 2.2). In the next section, we argue about the platform, and we present our vision of its application to research on Smart Cities.

5 Discussion

We firmly believe in the potential of this project as the cornerstone to enable new research directions. Graphs have been widely used to model different kinds of phenomena ranging from: urban street

[5]Scala: `www.scala-lang.org`
[6]Java: `www.java.com`
[7]Python: `www.python.org`
[8]R: `www.r-project.org`
[9]Google Maps: `/maps.google.com`

network (Jiang and Claramunt, 2004), urban and regional models (O'Sullivan, 2001), macroscopic model of city traffic (Prasanna et al., 2009), model city evacuation plan (Yamada, 1996), to plan strategy for vehicular ad hoc network in a city environments (Lochert et al., 2003) to mobility models (Mogre et al., 2007). In this project, we plan to use this graph model representing streets and intersections to study:

Supply chain from a transportation point of view. When cities have more nanostores than retailers, it is more complicated to transport products to small spare stores.

Multi-modal transportation is a problem in urban context where individuals need to optimize their movements within a city by using different massive transportation mode.

Crime patterns could be extracted by combining different features from the graph model.

Pollution dispersion could be modeled by a street and intersection models to represent and forecast particles of matter dynamic in a city.

Social network activity levels could be represented in the urban graph to detect social activity for extracting the hot spots in a city.

Privacy perception to understand how people consider privacy and what are the real dangers and risks.

The aforementioned list of possible research directions is not limited to these topics. There are many issues related to smart cities still opened.

6 Conclusions

In the present work, we have detailed the architecture of the Knowledge tier platform. The novelty of this platform is to gather diverse kind of data from different knowledge layers to extract spatio-temporal patterns for smart cities applications. We have shown the potential of this platform as the stone corner for many research question in the near future.

References

Alexander Avidan, Charles Weissman, and Charles L Sprung. 2005. An internet web site as a data collection platform for multicenter research. *Anesthesia & Analgesia*, 100(2):506–511.

Hafedh Chourabi, Taewoo Nam, Shawn Walker, J Ramon Gil-Garcia, Sehl Mellouli, Karine Nahon, Theresa A Pardo, and Hans Jochen Scholl. 2012. Understanding smart cities: An integrative framework. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2289–2297. IEEE.

Andrea Gaggioli, Giovanni Pioggia, Gennaro Tartarisco, Giovanni Baldus, Daniele Corda, Pietro Cipresso, and Giuseppe Riva. 2013. A mobile data collection platform for mental health research. *Personal and Ubiquitous Computing*, 17(2):241–251.

José M Hernández-Muñoz, Jesús Bernat Vercher, Luis Muñoz, José A Galache, Mirko Presser, Luis A Hernández Gómez, and Jan Pettersson. 2011. Smart cities at the forefront of the future internet. In *The Future Internet Assembly*, pages 447–462. Springer.

Bin Jiang and Christophe Claramunt. 2004. A structural approach to the model generalization of an urban street network. *GeoInformatica*, 8(2):157–171.

Chia-Hao Lo, Wen-Chih Peng, Chien-Wen Chen, Ting-Yu Lin, and Chun-Shuo Lin. 2008. Carweb: A traffic data collection platform. In *The Ninth International Conference on Mobile Data Management (mdm 2008)*, pages 221–222. IEEE.

Christian Lochert, Hannes Hartenstein, Jing Tian, Holger Fussler, Dagmar Hermann, and Martin Mauve. 2003. A routing strategy for vehicular ad hoc networks in city environments. In *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, pages 156–161. IEEE.

Parag S Mogre, Matthias Hollick, Nico d'Heureuse, Hans Werner Heckel, Tronje Krop, and Ralf Steinmetz. 2007. A graph-based simple mobility model. In *Communication in Distributed Systems (KiVS), 2007 ITG-GI Conference*, pages 1–12. VDE.

David O'Sullivan. 2001. Graph-cellular automata: a generalised discrete urban and regional model. *Environment and Planning B: Planning and Design*, 28(5):687–705.

UR Prasanna, M Srinivas, and L Umanand. 2009. Macroscopic model of city traffic using bond graph modelling. *International Journal of Engineering Systems Modelling and Simulation*, 1(2-3):176–183.

BP Shumaker and RW Sinnott. 1984. Astronomical computing: 1. computing under the open sky. 2. virtues of the haversine. *Sky and telescope*, 68:158–159.

Takeo Yamada. 1996. A network flow approach to a city emergency evacuation planning. *International Journal of Systems Science*, 27(10):931–936.