Smart Cities Symposium Prague 2017

Knowledge Discovery from Car Sharing Data for Traffic Flows Estimation

Alessio PAGANI, Francesco BRUSCHI and Vincenzo RANA

***Abstract*— The newly introduced car sharing services are an unexploited source of data that could be used to estimate the state of the road network as well as to provide new interesting analysis on urban mobility.**

In this paper we propose a Knowledge Discovery System that first gathers information from car sharing sites and applications, and then processes it to estimate interesting metrics such as travel time and vehicle flows in the urban areas at different times and in different days. We further argue that the information gathered can be processed in real-time, to estimate instant traffic, and can be exploited to perform deeper analysis, using historical data.

Finally, we analyze vehicle availability as a function of time in different zones and show how the results can be applied to travel time estimation, car stockout forecast and multimodal travel planning.

Keywords: Smart cities, Car sharing, Intelligent transport systems, Knowledge discovery from data, Data fusion, Cars flow estimation, Travel time estimation, Multimodal travel planning, Car sharing stockout

I. Introduction

An accurate knowledge of the *road network state* is fundamental to elaborate efficient strategies in urban transport management and planning.

Currently, the major suppliers reconstruct the road network state using dedicated data collection methods [1] such as GPS sensors (Floating Car Data [2]), mobile cell towers, access detectors, speed detection systems and, recently, crowdsourced data.

With **road network state** we refer to the state of the traffic on the urban streets, expressed for example as [3]:

1) Travel time: time needed to travel along a road segment of length L.
2) Speed: average speed in a road segment L completed in time TL.
3) Delay: difference between the effective travel time and the same travel with no traffic.

In this paper we propose an approach that exploits a *Knowledge Discovery from Data* technique to extract new information from car sharing data available online (gathered from the major providers available in Milan). This new information is used to estimate the road network state without requiring probes or sensors. Moreover, we show some interesting analysis, for example of the availability of the cars in the different parts of the city during the day, of travel time and of the average vehicle speed and fuel consumption. Eventually we propose an example of how these data can be elaborated, using data processing techniques [13], to estimate the travel time exploiting the historical information, and we introduce a new approach to detect the car stockout and to design a multimodal real time travel planner.

A. Pagani, F. Bruschi and V. Rana are with Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Milano, Via Ponzio 34/5, 20133 Milano, Italy (e-mail: <alessio.pagani, francesco.bruschi, vincenzo.rana>@polimi.it).

Knowledge Discovery from Data (KDD) is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [9].

[8], [10] outlines nine steps major steps in KDD:

1) Define purpose of process.
2) Generate subset data point for knowledge discovery.
3) Removing noise and expired data and handle missing data fields.
4) Find useful properties to present data depending on the analysis context.
5) Map purposes to a particular data mining methods.
6) Choose data mining algorithm and method for searching data patterns.
7) Research patterns in expressional form.
8) Returning any steps 1 through 7 for iterations also this step can include visualization of patterns.
9) Use information directly, combining information into another system or simply enlisting and reporting.

The study focus the Milan Metropolitan Area, that is the City of Milan and the surrounding areas organically connected where the car sharing services are available. The services considered are *Car2go*, *Enjoy* and *Twist*.

The car sharing data are gathered, using the available API or through data scraping techniques, from the websites of the operators. A dedicated server recorded the position of each car over the course of several days.

II. Rationale

Current services that estimate the state of urban road network use external on purpose instruments such as probe vehicles or passage / speed sensors. In this paper we propose an approach to estimate such state without using any external instrument but only exploiting data already available online, produced by the car sharing providers. This information is processed to provide an estimation of the state of the road network in the cities where other services (with dedicated resources) are not available and can be used, as an additional source, to improve the accuracy of state of the road network in the cities where other services are already available.

978-1-5386-3825-5/17/$31.00

III. State of the art

The major providers currently involved in the road state reconstruction are described below.

TomTom produces traffic data using information gathered from public sources and probe vehicles that use TomTom devices [17]. This system provide reliable estimations on the highways but it is not very accurate in the urban areas due to the limited number of probe vehicles.

Infoblu [18] aggregates the information about the highways in the north of Italy, similarly to TomTom they use probe vehicles to estimate the traffic in the highways and in the major extra urban streets.

Google Traffic estimates the road traffic calculating the speed of users along the streets [11]. To do this, Google analyzes GPS-determined locations exploiting data available from a large number of mobile phone users, processing the incoming raw data about mobile phone device locations. Google Traffic has a good accuracy even in the urban areas and provides travel time estimations using historical data. With the acquisition of **Waze** [19], Google Traffic now integrates also crowdsourced data.

Inrix is the company exploiting the major number of different sources: it uses fixed probes, probe cars and crowdsourced data. Inrix provides information to several different public administrations [16].

On the application side, *Petrovska and Stevanovic* in their work [12] provide an automated and interactive visualization tool for congestion analysis in real time that uses live traffic congestion data from Google Maps traffic layer, with the aim of reducing the traffic congestion on roads and of providing important data which can help road traffic management.

Our approach estimates traffic, travel time and cars flow using a dataset publicly available online. Differently from the current services, our work does not require any dedicated hardware (e.g., GPS sensors, speed detection systems) or a large number of users (e.g., crowdsourced data) to work properly. This system can be used to estimate traffic congestion where Google Traffic and other services are not available and to improve the accuracy of road network tools like the one previously described.

IV. Data gathering

Data gathering is one of the more complex and elaborated phases of the entire KDD process, since information must be extracted from different sources using different techniques depending on the specific provider.

Generally the data sources are divided in three types [5], [6]:

- **Structured data** is data already tagged and sorted.
- **Semistructured data** has fixed fields but contains separate data elements, that make it easy to tag and structure them.
- **Unstructured data** is the most difficult to analyze because it has not fixed fields or path.

Depending on the type of data and the services exposed by the car sharing provider, the data can be gathered using public or private API or through a data scraping process. The data originating from heterogeneous sources is then merged to a common structure using Schema Matching and Data Transformation techniques [4] typical of data fusion processes.

In the first phase not only data has to be gathered from heterogeneous data sources, but it has also to be merged, normalized and stored in an unique consistent database. This step includes outliers detection and removal (outliers usually are fake cars used by some providers for testing purposes), and fixing wrong or imprecise data (e.g., car reserved but not used).

A. Data sources

We have identified the three main car sharing provider in Milan, and for one week in May 2015 we stored the data of the cars in a database. For each provider we gathered every 15 seconds the GPS position of the cars available in that moment, in addition to some related information such as the plate number, the fuel level and the status of the car.

The identified providers are:

- **Car2Go** [20]: it provides a specific API with JSON structured data.
- **Enjoy** [21]: it is accessible throught private http GET request to an internal server, authentication and city selection with cookies.
- **Twist** [22]: it provides unstructured data in a JavaScript file.

Data are gathered every 15 seconds using different techniques: *Car2Go* data is requested using the provided public API while the other 2 providers does not offer any API and then data is gathered by means of data scraping techniques.

B. Current implementation

Scripts for the information gathering are written in **CoffeeScript** [23] using **Node.js** [24], an event-driven, non-blocking I/O model that makes it lightweight and efficient, perfect for data-intensive real-time applications that run across distributed devices and **Express.js** [26], a minimal and flexible Node.js web application framework that provides a robust set of features for web and mobile applications. The data are then stored in a database according to the following format:

- GPS position (latitude and longitude).
- Fuel level.
- Plate.
- Status of the car (interior and exterior).

Plus some information added by the scraping script:

- Name of the provider.
- Timestamp.

This data includes only the information of the car available (rentable) at that time, while no information about the currently rented car is available.

As main database we used MongoDB [25], a document-oriented database that is proven to be efficient with this kind of data and effective even for big data problems [15].

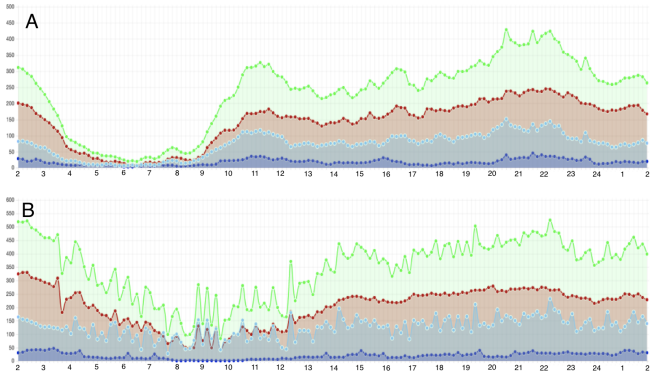

Fig. 1. Average cars hired per hour, during the weekday (A) and during the weekend (B).

V. Data aggregation

As explained in [7] aggregation tools have the ability to organize data in a way that it can be quickly searched, analyzed and efficiently utilized.

All the analysis and process are executed using Pandas [27], a python library providing high-performance, easy-to-use data structures and data analysis tools.

We consider three types of aggregation: per provider, per time and per zone.

1) Aggregation per provider: To analyze the preference of clients and the usage of each provider, the usage data is grouped by provider. Enjoy and Car2Go are widely used and their customer show similar usage patterns, with Enjoy being the provider with the major number of cars hired most of the times (also due to the fact that at the time of the study it was the provider with the highest number of cars). Twist is much less widespread, and its vehicles are almost always available but not used.

In each of the following graphs, the highest line is the sum of the three providers while the other three lines represent the three providers: the red line is for Enjoy cars, the azure one is for Car2Go cars and the blue one for Twist cars.

2) Aggregation per time: The data is firstly divided in two groups, **weekday data** (form Monday to Friday) and **weekend data** (Saturday and Sunday), and then aggregated per hour. The hired vehicles are calculated, for each provider, subtracting the available vehicles from the total of the vehicles.

Results are shown in Figure 1. During the weekdays it is easy to identify two usage peaks, one form 10:00 am to 12:00 am (more than 300 car used at the same time) and the other from 8:00 pm to 22:00 pm (more than 400 car used at the same time). On the other hand there are just a few vehicles used (less than 60 car used at the same time) during the night, especially from 4:00 am to 8:00 am. During the weekend, the car are used more uniformly, even during the night, with an average of 200 cars used at the same time between 5:00 am and 11:00 am and an average of 420 cars used at the same time during the rest of the day. The information about the hired is correlated with the traffic on the city roads, thus it is the first indicator for the reconstruction of the road network state.

The data of the two previous groups are divided in three subgroups, based on the GPS of the car, as shown in Fig. 2 and listed here:

- **City center**, the so called *cerchia dei Bastioni*, that delimits the historic center of Milan.
- **Circumvallation**, the area between the city center and the so called *circonvallazione filoviaria* that delimits the historic urban area of Milan.
- **Outskirts**, all the new urban area of Milan outiside the *circumvallation*.

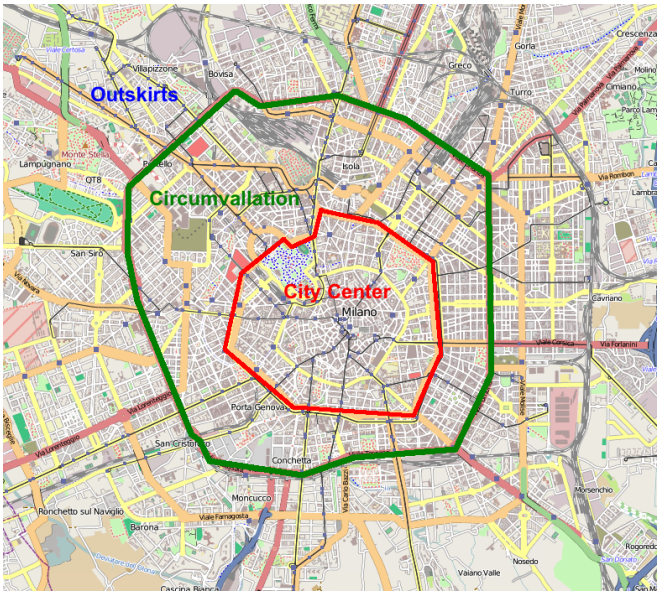

Fig. 2. Division of the city in the three major areas: city center, circumvallation and outskirts

3) Aggregation per zone: To interpret the data, we clustered the city in three areas, characterized by their urban function: the city center contains the majority of the offices and the shops, the area between the city center and the circumvallation includes shops, offices and houses, while the outskirts are the more popular area and includes the majority of the residential zones.

In Figure 3 the availability of cars in the city center during the weekdays (a) and the weekend (b) is shown: in the weekdays there is good availability of cars (at least 30 cars) between 7:00 am and 5:30 pm, while there is scarce availability during the rest of the day. Weekend situation is similar, with a good availability during the day (from 10:00 am to 5:00 pm) and low for the rest of the time.

In Figure 4 the availability of cars in the circumvallation during the weekdays (A) and the weekend (B) is shown: both in the weekdays and in the weekend there is a good availability of cars, with more than 200 cars always available. This prove that many people moves in this area of the city.

In Figure 5 the outskirts, the biggest area of the city, are considered, during weekdays (A) and the weekend (B). the majority of the cars is obviously here with the highest

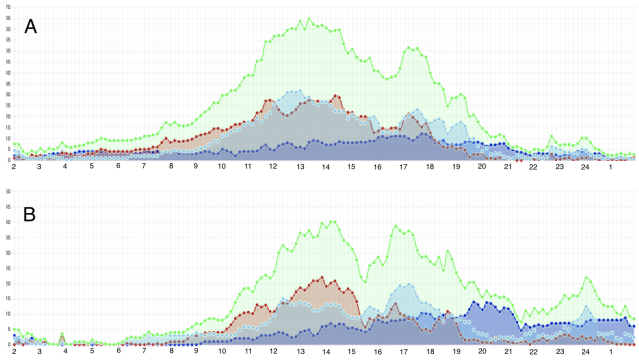

Fig. 3. Average number of cars available in the city center during the weekdays (A) and the weekend (B)

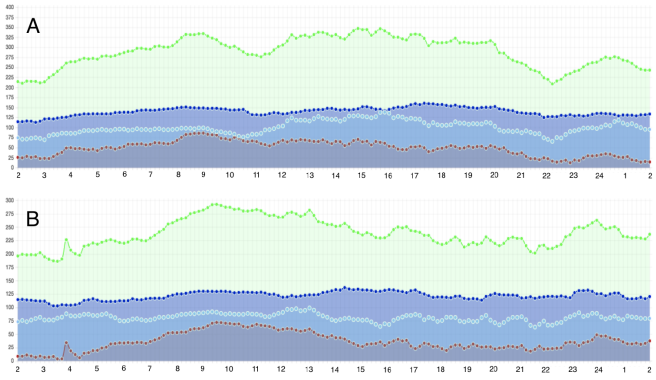

Fig. 4. Average number of cars available in the circumvallation during the weekdays (A) and the weekend (B).

concentration during the night (more than 900 cars available from 10:30 pm to 9:30 am during the weekdays and form 4:30 to 12:30 during the weekends), when people came back from work or from the bars and restaurants in the city center.

These analysis are used to determine the cars flow, that is, as expected, in the morning from the outskirts to the center and in the evening from the center to the outskirts. To determine the **cars flow** accurately in the state of the cars network we did a similar analysis dividing the city in more zones considering the main city roads and the major points of interest.

VI. Data processing

The car sharing data, stored in the form of *car entry* is processed to detect *travels*. A **travel** is defined as a pair of car entries: each **car entry** include the car plate, GPS position of the car, the timestamp, the fuel level and other informations. A travel is created each time a vehicle (identified by its plate) disappears and reappears after some time, excluding cars that reappear in the same position (these are considered cars booked but not used, thus excluded from the travel list).

A car disappears from the list of the available vehicles (updated each 15 seconds) when it is booked or rented by a user and it reappears when the user completes the trip.

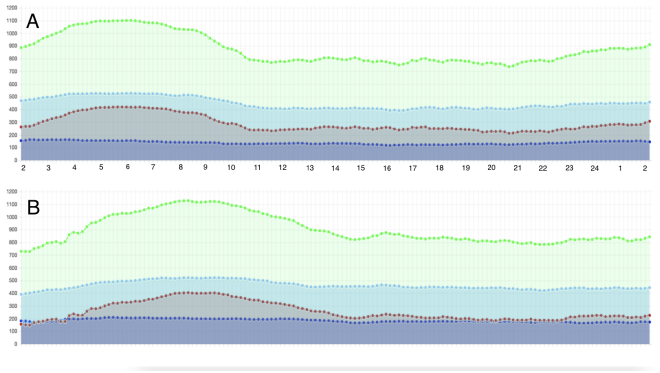

Fig. 5. Average number of cars available in the outskirts area during the weekdays (A) and the weekend (B).

For each travel we then extract some meaningful informations, such as:

- **travel time**: difference from when a car is booked / rented and released.
- **average speed during the day**: difference between the estimated distance traveled (we do not know the exact path but we estimate it using the best route proposed by a trip planner) and the travel time.
- **fuel consumption per km during the day**: the fuel consumed per km during the different hours of the day.

1) Travel time: The **travel time** is calculated as the difference from when a user book / rent a car and when the user complete the travel. The calculated travel time thus include also the time from when the car is booked and when the travel actually starts.

Figure 6 shows the cumulative travel time: approximately 10% of the travels ends in 10 minutes, half of the travels are concluded in half an hour, 75% of the travels end in 1h. The majority of the travels (about 40% of the total) last between 20 and 40 minutes.

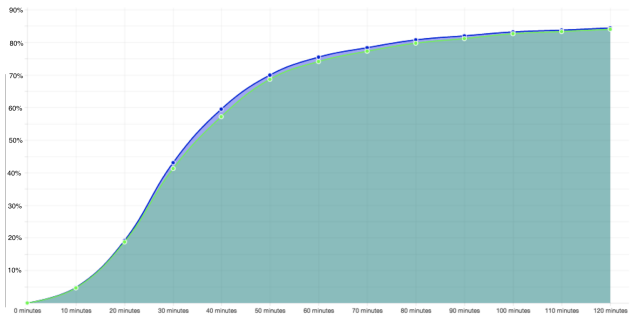

Fig. 6. Cumulative travel time.

2) Average speed and fuel consumption: To compute average speed and fuel consumption of each travel we firstly estimate the *distance traveled* as the length of the best route proposed by a route planner (we have no information about the real routes). We then define the **average speed** as the difference between the distance traveled and the travel time

and the **fuel consumption per km** as the difference between the fuel consumption and the distance traveled.

Even if the results are rough, due to the approximations and the quality of the data (especially for the fuel), Figure 7 are shows clear trends in the average speed and fuel consumption: the blue line shows the averages for the weekdays, while green line shows the averages for the weekends. It is clear that the average speed is much higher during the night (triple fo the day hours), the average speed remains high until 6:30am during the week days and until 9:00am during the weekends. On the contrary, the fuel consumption is higher during the day, with 2 peaks at 13:00pm and one at 17:00pm.

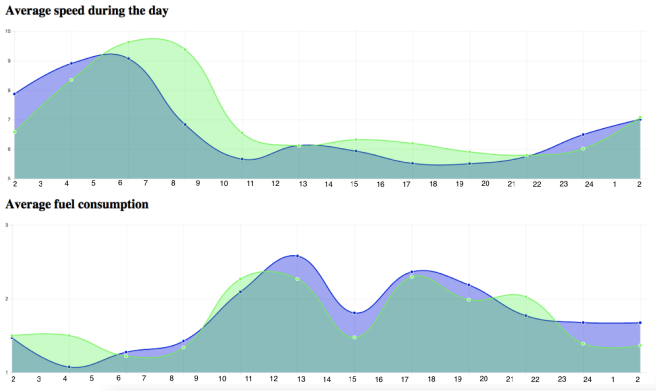

Fig. 7. average speed (A) and average fuel consumption (B) during the day.

A. Estimated travel time

Next, we define a tool that estimates the travel time at a given hour between two points of the city using real time and historical data. This tool is used to improve the estimated road network state. All the travels started within a given radius r from the starting point and ended within a radius r from the destination point are grouped and the estimated travel time is calculated as the average travel time of each travel (Sec. VI-.1). The data considered can include all the data of the travels occurred in the same day of the week at the same hour (used for statistical analysis) or only the travels in the current day (used for real-time traffic estimation).

In Figure 8 is shown an example of estimation of a travel time between two points of the city: the tool looks for the cars that traveled from point A to point B at that hour and calculates the average travel time.

B. Cars stockout detection

Another interesting application would be the detection of areas with a high request of cars but with no cars available (stockout). This would allow to detect zones where the number of vehicles is decreasing even though they are still very requested. This information would be very useful for the providers as they would be able to take action by moving the fleet according to the requests with an increase of profits: for example by relocating the cars into these specific areas or offering discounted price to users moving towards less populated, but requested, areas.

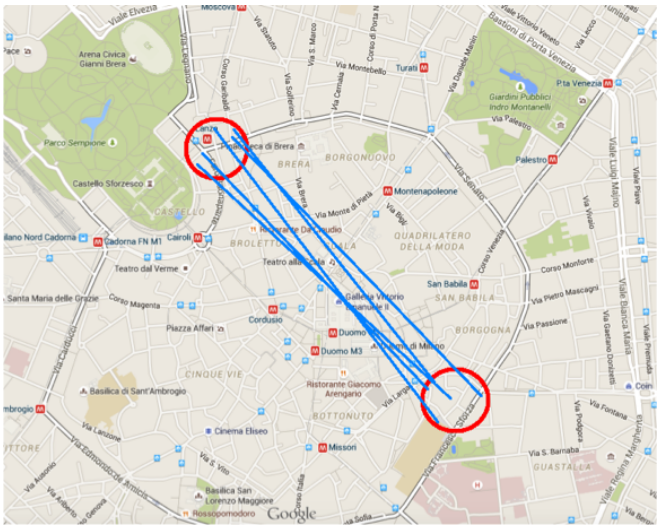

Fig. 8. To estimate the travel time from point A to point B we calculate the average time of similar travels selecting all the travels that start inside a circle of radius r with center in A and end in a circle with radius r and center in B that occurred at the same time in the same day in the last weeks.

C. Multimodal travel assistant integration

Another possibility would be to integrate the information extracted into existing travel planners. In particular, this would make it possible to improve the accuracy of existing multimodal real-time travel assistants. For instance, the approach described in [14] shows how to develop a platform for travel planning in real time using publicly available data. Integrating car sharing information, such as the position of the available cars and the state of the road network (cars flow and travel time), with the public mobility state enable the possibility to design a multimodal travel assistant that uses both massive public transportation (e.g., bus, undergrounds) and private public transport (car sharing services) using only data publicly available online.

VII. Conclusions

We proposed a Knowledge Discovery from Data approach to extract useful information from data produced by the car sharing providers and publicly available online. We detected the zones of the city with the highest and lowest concentration of cars hour by hour during the weekdays and the weekends, revealing some interesting correlations: for example we confirmed that during the day there are a lot of cars in the city center, where there are the majority of offices and shops while during the evening the cars move to the suburbs. As expected, we showed that the speed is lower during the day than during the night and the fuel consumption has the opposite trend.

We designed a technique to estimate the road network state (travel time and traffic flows) that does not require any dedicated sensor or probe, but only uses information (historical and in real-time) about the car sharing travels during the day. This technique can be used to reconstruct the road network in the cities where other services (with dedicated resources) are not available and can be used, as an

additional source, to improve the accuracy of road network state in the cities where other services are already available.

Finally we proposed further possible applications where these data and the process described can be useful, such as car stockout detection and multimodal travel assistance.

REFERENCES

[1] M. Rahmani, H. N. Koutsopoulos, A. Ranganathan, Requirements and Potential of GPS-based Floating Car Data for Traffic Management: Stockholm Case Study, International Conference on Intelligent Transportation Systems ITSC, IEEE, 2010

[2] Shawn M. Turner, William L. Eisele, Robert J. Benz, and Douglas J. Holdener, Travel Time Data Collection Handbook, Research Report 07470-1F, Texas Transportation Institute The Texas A&M University System College Station, Texas, 1998

[3] C. A. Quiroga, Performance measures and data requirements for congestion management systems, Pergamon, Transportation Research Part C: Emerging Technologies, 2000

[4] F. Naumann, A. Bilke, J. Bleiholder and M. Weis, Data Fusion in Three Steps: Resolving Inconsistencies at Schema-, Tuple-, and Value-level, IEEE Data Eng., vol. 29, no. 2, pp. 21?31, 2006

[5] C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012

[6] S. Singh and N. Singh, Big Data Analytics, 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, 2011

[7] P. A. Prakashbhai and H. M. Pandey, Inference Patterns from Big Data using Aggregation, Filtering and Tagging- A Survey, Confluence The Next Generation Information Technology Summit (Confluence), IEEE, 2014

[8] S. Sargiroglu and Duygu Sinanc, Big Data: A Review, Collaboration Technologies and Systems (CTS), IEEE, 2013

[9] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, AI Magazine, Fall 1996, pp. 37- 54

[10] E. Begoli and J. Horey, Design Principles for Effective Knowledge Discovery from Big Data, Software Architecture (WICSA) and European Conference on Software Architecture (ECSA) Joint Working IEEE/IFIP Conference on, Helsinki, 2012

[11] K. Subramanian and R. Srikanth, Now, Apps for Live Traffic Feed, The Hindu, 2016

[12] N. Petrovska and A. Stevanovic, Traffic Congestion Analysis Visualisation Tool, International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2015

[13] C. French, Data Processing and Information Technology (10th ed.). Thomson. p. 2, 1996

[14] A. Pagani, F. Bruschi, V. Rana, M. Restelli, Reconstruction of public transport state, International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2016

[15] MongoDB, Big Data: Examples and Guidelines for the Enterprise Decision Maker, MongoDB White Paper, 2016

[16] http://www.inrix.com/publicsector.asp

[17] http://livetraffic.tomtom.com/

[18] http://www.infoblu.it/?q=en/RealTimeTraffic/

[19] https://biz.world.waze.com/

[20] https://www.car2go.com/

[21] https://enjoy.eni.com/

[22] http://twistcar.it/

[23] http://coffeescript.org/

[24] https://nodejs.org/

[25] https://www.mongodb.com/

[26] http://expressjs.com/

[27] http://pandas.pydata.org/