Journal of Engineering and Applied Sciences 13 (2): 529-532, 2018
ISSN: 1816-949X

Social Media Big Data Based Transport Infrastructure Design

Jan Eermak
Department of Applied Informatics, Faculty of Civil Engineering,
Czech Technical University, Prague, Czech Republic

Abstract: Practical solutions to the problem of developing a transport infrastructure brings a lot of analysis before the proposal of the transport infrastructure. Initial analysis should cover it from what places and what ways is moving segmented group of people, the future user's infrastructure. Currently, the monitoring of displacement of the population using census and traffic surveys results. This study describes method how to use new data sources for transport infrastructure planning. The first one is using of geolocation data from social networks for transport infrastructure planning and compare it with the current transport infrastructure planning. The second data source is open data from municipalities. This data sources could be combined together. There are many advantages and disadvantages and this study discussed both pros and cons. Benefits include low costs of data from social network, data are continuously accessible and could be reached in real time. The disadvantages include small data population covering most users are from cities, coverage of villages is worst.

Key words: Big data, social networks, transport, infrastructure, open data, population

INTRODUCTION

This study deals how data from new data source could be helpful in sustainable construction of transport infrastructure. New data source means social networks and open data which provides municipalities and combine them in cooperation with current methods to create model for sustainable transport infrastructure.

Definition and problem description: Currently, transport infrastructure planning is based on many data inputs such as census and traffic surveys. Simplified for transport infrastructure design examines how many people are transported from place A to place B focused on defined area like city or part of city. Data are calculated using quantitative approaches based on statistical and mathematical techniques to model the operations and performances of transportation networks (Macario, 2010). Data inputs are based on the research of movement and map data like Geographic Information System (GIS) (Chattopadhyay *et al*., 2002). The old one technique traffic survey is still used.

As time progresses further forward still created new data sources that can be used as additional input data to the model for the design of roads. Every day more than 2.5 billion people using social networks. Many of them using social networks for posting own content for their followers.

This study will focus on the following socials networks: Foursquare, Facebook, Twitter and Instagram the biggest social networks for their large user base and for their API for consume user's shared content. Researcher choose social network Foursqaure because Foursqaure is the biggest geolocation based social network with many venues categorized spots on the map. Facebook and Instagram were chosen because these networks using the same geolocation service and for the large user base. Large user based has also social network Twitter. Twitter is social network for short messages which could contain geolocation (Anonymous, 2015).

Shared content could contains geolocation data where the content has been shared. Using information about user who posts the content and using timestamp we could aggregate each user's content into one timeline. This timeline will show the movement of specific user and detect which means of transport was used.

The beneficial aspects of new online technologies include the low cost of acquiring data from them, the constant timeliness and constant availability and ultimately the collection and analysis of the data done in real time.

The first problem is definition what is place A and place B and why is it important. It is necessary to define where the user is moving. For example our user is an employee and from Monday to Friday is moving from home to work and back to home. For this kind we should call place "A" as a home and place B as a workplace.

As a second example we should imagine an undergraduate student. During a week our student lives in a dormitory and is moving from dormitory to university and back to dormitory. We could call place a as a home and university as a workplace. But our student has "second" home it is a place where student is moving typically on several weekends. To define places A and B we should use models based on geolocation social media data.

The second problem for explanation is define where groups of people are during a day (day's visitors) and where the same groups of people are during a night (staying overnight). We should aggregate this data by users or by location.

MATERIALS AND METHODS

Access to data from social networks: Each social network mentioned in this study provide own API. API is documented interface where 3rd party application serve requests and API returns responses in specify format. Each social network has different requests and responses format, limits and privacy policy (Cermak *et al*., 2015).

For access to the data is needed to create own (3rd party) application on the social network developers dashboard. This application receives credentials. Credentials are using for requests authentication.

API limits means how many requests can application send to the API for specify time. Every request type has separately limits, e.g., Metadata about places has different limit than request which receives user content.

Types of request has own privacy policy. That means eg. place's Metadata are fully available and its return to the application but user content can be serves according post, user or social network policy. Post policy determines user and indicates how content has been shared if was shared publicly for users relations or for relations of user relations, etc. User policy determines user profile visibility, if is it visible for public or for user relations only. In both cases API reflects policy. API could contains another limitation, e.g., requests cannot be send as application but only as user which is authenticated through this application. In this case user and posts privacy are related to this authenticated user.

User can be authenticated for this application by authentication flow. All mentioned social networks using open authorization in second version, called OAuthz (Leiba, 2012) protocol or oauth2 like for user authentication. User is redirect to social network login page where accepts or denied authentication request and scopes. Scope determine which user data can application receives related to this user.

User's content (also known as post, Tweet, check-in) is entity which is created by user explicit interaction to the social network. Each social network has own rules which elements are required which are optional and has own privacy options. For example social network Instagram (Seltzer *et al*., 2015) has as a required element the image and privacy is determine if user has publicly profile or if users profile is locked and can be viewed by accepted relations only. Content could contains attachments like an image, video, external URL link, sound, etc. Place is localization entity. Should have own Metadata like name, category and localization coordinates.

Social data description: There are differences on data which can be reach from the APIs. The difference can be either form as well as their scope. The more data is acquired, the overhead can be made larger set of modeling. Not all data are, however, necessary to roads modeling. For example images. Of course other data can infer missing values required for entry into the model (Kaiser, 2014, 2012). For example, the means of transport which is used by the user can be identified according to locations between which the user is moving and how long it takes to move between these points. Type of transport can also detect text user's content.

For the initial data input is needed to determine how many people are moving from point A to point B. To determine these data sufficient timestamp when the user is at point A and timestamp when the user is at point B. To move between these points, users can make transfer points public transport interchange station or junction depends of type of transport. To relate the data to the demographic distribution of the data needed by the users themselves such as their age, gender, marital and social status.

RESULTS AND DISCUSSION

Determination of means of transport from the data: To determine the means of transport that the user is used to move from point A to point B can serve Metadata about where the user is moving. Metadata means not only geolocation coordinates as well as the categorization of the place. If the points A and B are two metro stations, it is likely that the user has used the subway to move between these points. But if the first point is a parking and the end point is the subway station, it's not clearly to define the means of transport. User could move the car between two points or leave the car and go on the subway. A parking could be detected using "parking" category in a venue (the better way) or combine coordinates of the user's interaction with the municipality open data parking locations. To detect the means of transport in this case may help point between the start and end points. When it's landing on the road, it is likely that the user used the car. If is the subway is likely that

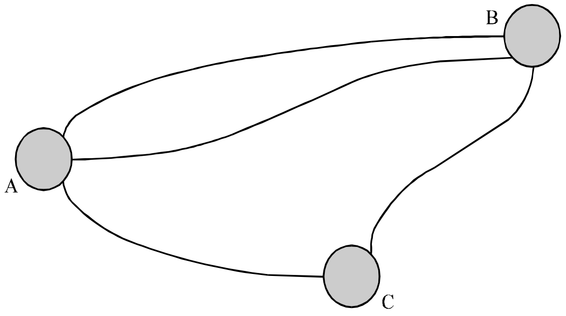

Fig. 1: Transportation routes between points A and B

the user used this means of transport. In the case that there is neither category of place can be used to detect the means of transport text of the shared content (Fig. 1).

To determine what are the correct points A and B where user is moving we should use geolocation, timestamp and repetition of user interaction. For example where user lives we could determine if user is active on this place every day (from Monday to Friday) in the night. Of course there are many exceptions for unemployed, shift work's employees, commercial travelers, etc. We determined place A as we should called it as a home. For determine of a home place we should use location Metadata from social networks. If the place has a category "Home" or something like this, we should use this place. To determine place B place where is user during a day. It should be a workplace, school etc. When user is not at home should be in a work or on the road to the workplace. To recognize home from work if we are not be able to used interaction timestamps we could use the other location interaction. For example when user is at work he could go to the restaurant (cafeteria, student's canteen, buffet) to eat. We are studying which way user using to transport between places A and B. If user is using the same place every day, we should mark this place and calculate with them. In the next figure is point A as a home, point B as a workplace. Between points A and B are many routes but user often using specific point, we could mark this point as C.

To know about moving large group of people we should use present population metrics. Present population splits groups to two daily visitors and staying overnight people. In this metrics we are focusing on this two groups and moving between a day in a specific location. In a location there are 100% people and we are watching how many percents of this groups are in this place during a day. We should combine many nearby location and observe from which location to the other is group moving and when. Figure 2 and 3 show the difference between working days and weekends. There are little bit difference between working days from Monday, Thursday and Friday.

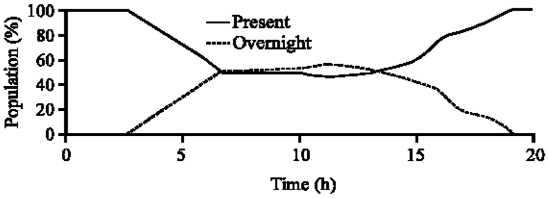

Fig. 2: Example of present population in working days

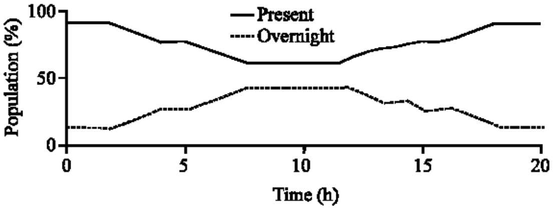

Fig. 3: Example of present population in weekends

Figure 2 and 3 show how many people stays in specific area during a day. The y-axis represent percentage of population groups in the area from 0-100% of population groups in total. The x-axis represent the hour during a day in 24 h format from 0-23. The total amount of both groups of population is always 100% in the every time during a day. Figures show the differences between working days and weekends.

Using opendata from municipalities and combine it with social media data: Some municipalities provides geolocation data about their area as open data. Open data means machine-readable data which are available in the Internet network and it's shareable without license limitation. These data can be used to supplement the Metadata about the places that were taken from social networks and can be used to map data on network traffic roads. These data may be used for the mapping of exists from public transport, stops clarify the position of but also complement the entrances and exits to the highway.

From the data that publish cities like London (Media Endpoints, 2015) or Prague (Opendata Praha, 2015) can be used map data in the form of polygons. List of parking lots will serve the mapping points on the Metadata about the parking while public transportation stops data can be used to specify the position of public transportation.

Using new data sources from social networks has potential, its advantages and disadvantages. Benefits include low costs of data from social network, data are continuously accessible and the data done in real time. The disadvantages include that coverage of the data is in cities mainly in towns and villages at least and not at all of users share their content with privacy level which cannot be reached by API.

CONCLUSION

Despite the drawbacks mentioned data can be new sources of quality data input for sustainable construction and predicting the transport infrastructure. As the optimal solution would be a combination of existing analytical tools supplemented by inputs from the optimization models of new sources of data from social networks. These resources should over time, how they grow, complement the final model.

ACKNOWLEDGEMENTS

The results presented in the study were partially supported by the Research, Development and Innovation Council an advisory body to the Government of the Czech Republic (project SGS16/200/OHK1/3T/11).

REFERENCES

Anonymous, 2015. Statistics and facts about social media usage. Statista Database Company, Hamburg, Germany. http://www.statista.com/topics/1164/social-networks/.

Cermak, J., L. Horak, J. Kaiser, M. Sura and T. Vanicek *et al*., 2015. Artificial intelligence methods in building industry II. Master Thesis, Czech Technical University in Prague, Prague, Czech Republic.

Chattopadhyay, M., R.S. Resmi and A.S. Promodhlal, 2002. Application of remote sensing and geographic information system in infrastructure development. J. Indian Soc. Remote Sens., 30: 143-147.

Kaiser, J., 2012. Algorithm for missing values imputation in categorical data with use of association rules. ACEEE. Intl. J. Recent Trends Eng. Technol., 6: 111-114.

Kaiser, J., 2014. Dealing with missing values in data. J. Syst. Integr., 5: 42-51.

Leiba, B., 2012. Oauth web authorization protocol. IEEE. Internet Comput., 16: 74-77.

Macario, R., 2010. Critical issues in the design of contractual relations for transport infrastructure development. Res. Transp. Econ., 30: 1-5.

Media Endpoints, 2015. Instagram developer documentation. Media Endpoints, USA. https://instagram.com/developer/endpoints/media/.

Opendata Praha, 2015. Hadat data. Opendata Praha, Czech Republic. http://opendata.praha.eu/.

Seltzer, E.K., N.S. Jean, K.E. Golinkoff, D.A. Asch and R.M. Merchant, 2015. The content of social media's shared images about Ebola: A retrospective study. Public Health, 129: 1273-1277.