Profiling users of the Vélo 'v bike sharing system

Albrecht Zimmermann, Mehdi Kaytoue, Marc Plantevit, Céline Robardet, Jean-François Boulicaut

Profiling users of the Vélo'v bike sharing system

Albrecht Zimmermann, Mehdi Kaytoue, Céline Robardet, Jean-François Boulicaut FIRSTNAME.NAME@INSA-LYON.FR
INSA-Lyon, CNRS, LIRIS UMR5205, F-69621, France

Marc Plantevit MPLANTEV@LIRIS.CNRS.FR
Université Lyon 1, CNRS, LIRIS UMR5205, F-69622, France

Abstract

Detecting and characterizing geographical areas that are attractive places for specific people, in specific contexts, is an important but challenging new problem. Mobility traces and their related circumstances can be modeled thanks to an augmented graph in which nodes denote geographic locations and edges are represented by a set of transactions that describe users' demographic information (e.g. age, gender, etc.) as well as the conditions of the movement (e.g. day/night, holiday, transportation mode, etc.). We propose to extract connected subgraphs that are related to some user profiles, and use it to understand the usages of the Vélo'v bike sharing system.

1. Introduction

The problem considered hereafter is how to detect and characterize geographical areas that are attractive places and routes for specific contexts. Such areas are frequently accessed together in certain conditions by users of similar profiles compared to all contexts and users. Starting from a relational database that gathers information on people movements – such as origin, destination, date and time of travel, means of transport, reasons for traveling, etc. – as well as demographic data, we adopt a graph-based representation that results from the aggregation of individual travels. In such a graph, the vertices are locations or points of interest (POI) and the edges stand for user's co-visitations. Travel information as well as user demographics are labels associated to the edges of the graph. Figure 1 (a) depicts an example of travels undertaken by users (denoted $u_1, \ldots, u_4$). For each user, we know her age and gender, the context of the move (day or night) and the set of movements, identified by a pair origin/destination, that occur in this context. Capital letters, from A to E, represent POI. This table can also be viewed as an edge-attributed graph where edges stand for movements and are labeled by the attribute values of the context. For instance, we have a directed edge (A, C) labeled by $(F, 20, Day)$ for the user u_1. Given a specific context, the edge-attributed graph can be transformed into an aggregate graph whose edges are weighted by the number of attributed edges that hold for the context. Three examples of aggregated graphs are given in Figure 1 (b),(c) and (d). The weights of the aggregated graph can be seen as the support of the context in the graph.

The problem is thus to identify the contexts and sub-graphs that are specific to one another. By specific, we mean that a large proportion of the weight of each sub-graph edge mainly corresponds to users that satisfy the context. The adequacy of a context to an edge is assessed by a χ^2 test and some novel quality measures that makes it possible to identify the so-called *demographic and contextualized specific areas (DCSA)*. Two *DCSA* patterns are presented in Figure 1 (c) and (d) (in bold): The first one identifies a sub-graph that is traveled during the *day* , mainly by people with *age greater than* 45. In the second sub-graph, bold edges are very specific to *male* persons' behavior, whatever the travel time.

2. Travel patterns in the VÉLO'V system

VÉLO'V is the bicycle sharing and renting system run by the city of Lyon (France) and the company JCDecaux.[1] The VÉLO'V dataset contains movement data collected between Jan. 2011 and Dec. 2012. Each movement includes both bicycle stations and timestamps for departure and arrival, as well as some basic demographics about the user of the bike. We aggregated all movements a user performed between any two stations for the entire time period. Hence, the VÉLO'V stations are nodes in the graph (342 in total),

Proceedings of the 2^{nd} International Workshop on Mining Urban Data, Lille, France, 2015.

[1] http://www.velov.grandlyon.com/

User	Gender	Age	Time	Travels
u_1	F	20	Day	(A,C),(B,A), (C,B)
u_1	F	20	Night	(D,C),(D,E),(E,A), (E,D)
u_2	M	23	Day	(A,B),(B,C),(C,A), (C,B)
u_2	M	23	Night	(A,B),(B,C),(C,B) (C,D),(D,C),(D,E), (E,D)
u_3	F	45	Day	(A,B),(B,C),(C,D), (D,A),(D,E),(E,D)
u_3	F	45	Night	(B,D),(D,B)
u_4	M	50	Day	(A,B),(B,C),(C,B), (C,D),(D,A),(D,E), (E,D)
u_4	M	50	Night	(A,C),(C,A)

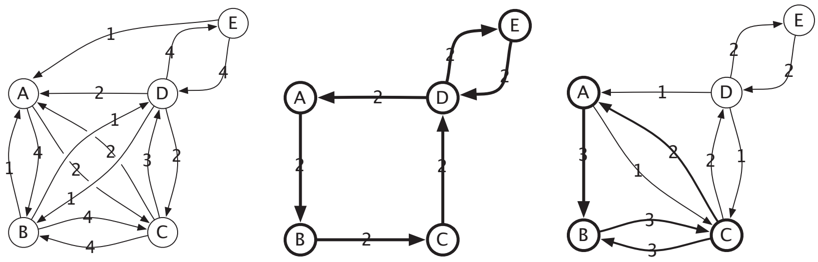
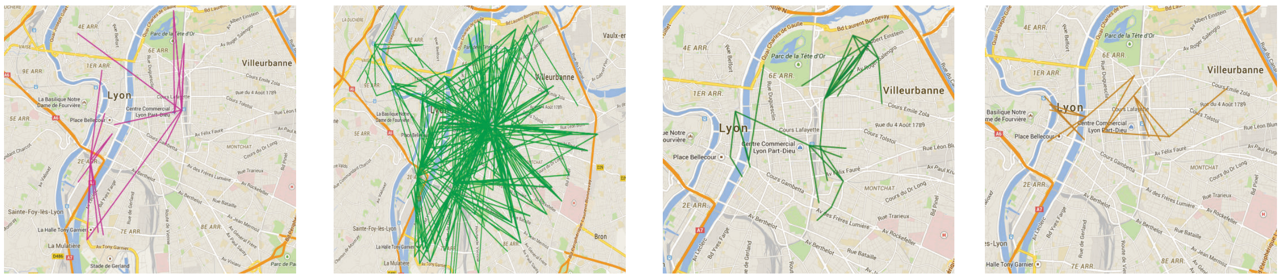


Figure 1. Example of contextualized trajectories: (a) Transactional view; (b) Aggregate graph w.r.t the most general context $\star = (Age \in [20, 50],\ Gender \in \{F, M\},\ Time \in \{Day,\ Night\})$; (c) Aggregate graph w.r.t. context $(Age \in [45, 50], Time = Day)$; (d) Aggregate graph w.r.t. context $(Gender = M)$;



(a) YoB $\geq$ 1968, ZIP $=$ 42400 **(b)**YoB $\geq$ 1962, CAT $=$ OURA **(c)**YoB $\geq$ 1980, TYP $=$ std **(d)**YoB $\geq$ 1992, ZIP $=$ 69003

Figure 2. *DCSA* discovered from VÉLO'V

and edges link two stations if a VÉLO'V customer checked out a bicycle at the first station and returned it at the second one. We treat the edges as undirected. Customers are described by nominal attributes such as gender, type of membership card, ZIP code and country of residence, as well as a numerical one: year of birth. There are a total $50,601$ customers. The data set comprises around 2 million contextualized edges.

Given the characteristics of different users, we aim to identify populations that use the rental bicycles in a particular manner. Figure 2 shows 4 different *DCSA* from VÉLO'V. Pattern (a) identifies people born after 1968, living in a city (Saint Chamond) located approximately 50km from Lyon. It is therefore not surprising that the edges involve the two main train stations of Lyon: Perrache (south-west) and Part-Dieu (center), from which users take bicycles to areas that are not easily reached by metro or tram, such as the 1^{st} and 4^{th} arrondissements. The edges of pattern (b) radiate from all of Lyon's train stations, not only the major ones. Its description refers to holders of a regional train subscription, and the pattern notably involves 200 nodes, almost 60% of the stations. It is very likely that this pattern identifies commuters. Pattern (c) involves users born after 1980 and we can identify three main areas: the scientific campus in the north, the Presqu'ile and its pubs, and the shopping area in the city center. It is notable that several of the long edges correspond to very comfortable cycling routes. Pattern (d) does not seem to be very exciting: young people that live in the 3^{rd} arrondissement use VÉLO'V bicycles to move around in their area. At a second glance, however, this is the closest that we will come to a ground truth in real-world data: the ZIP code of users aligns with the area where the bicycles are used!

3. Conclusion

The problem of finding *DCSA* in edge-attributed graphs has many applications in location based social networks and recommendation systems. It allows to find connected components highly characteristic of a given category of users. The proposed inductive approach is solved thanks to an efficient data mining algorithm that avoids materializing all contexts/induced-graph pairs and benefits from pruning and upper bound computations techniques. Its use for the analysis of the bicycle sharing system Vélo'v demonstrates its capabilities to provide new valuable insights. [2]

[2]This work has been partially supported by the projects GRAISearch (FP7-PEOPLE-2013-IAPP) and VEL'INNOV (ANR INOV 2012).