Information Systems 57 (2016) 223–240

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

City data dating: Emerging affinities between diverse urban datasets

Gloria Re Calegari*, Irene Celino, Diego Peroni

CEFRIEL – Politecnico di Milano, via Fucini 2, 20133 Milano, Italy

ARTICLE INFO

Available online 10 August 2015

Keywords:
Smart city
Data diversity
Spatio-temporal data resolution
Mobile data processing
Correlation analysis
Regression analysis
Clustering analysis
Information fusion

ABSTRACT

Cities are complex environments in which digital technologies are more and more pervasive; this digitization of the urban space has led to a rich ecosystem of data producers and data consumers. Moreover, heterogeneous sources differ in terms of data complexity, spatio-temporal resolution and curation/maintenance costs. Do those diverse urban sources reflect the same picture of the city? Do distinct perspectives share some commonalities?

In this paper we present our data analytics/empirical experiments on a set of urban sources related to the city of Milano; our investigation is aimed at discovering "affinities" between datasets by means of different quantitative and qualitative correlation analyses. We also explore the influence of spatial resolution and data complexity on the dependence strength between heterogeneous urban sources, to pave the way to a meaningful information fusion.

1. Introduction

We live in the age of data and the digitization of cities has led to produce massive datasets and data streams related to the urban environment. This data deluge is caused by a number of different factors: the advent of the open data movement, with its call for transparency in public sector information; the increasing popularity of sensor technology and the Internet of Things; the maturity of location-based services and social networks, with the constant production of user-generated information, more and more frequently characterized by its spatio-temporal context. Information is also produced as collateral effect of other activities: for example, telecommunication operators collect call data records as needed by their business, but that kind of information can also be seen as a special representation of the urban space.

Urban data has thus become ubiquitous, and our cities can be described through the lenses of a multitude of information sources. Still, the collection, cleansing, curation and maintenance of specialized data sources can result in a complex and expensive process; this is the case of datasets requiring a manual intervention, like demographics data which requires a human-based census activity, or an error-prone (semi)automatic processing, like land use information, that starting from aerial or satellite imaging, characterizes the environment with reference to domain-specific classifications.

Our current investigation is aimed to answer to the following research question: would it be possible to use one or more "cheap" datasets as proxy for more "expensive" data sources? In other words, would it be possible to (semi) automatically generate or revise an outdated dataset, which otherwise would require a costly human work, on the basis of the content of other up-to-date information sources?

To realize such a goal, the first step is to analyze available urban datasets and to identify potential intrinsic dependence and inter-relationships between them. Since their

* Corresponding author.
E-mail address: gloria.re@cefriel.com (G. Re Calegari).

http://dx.doi.org/10.1016/j.is.2015.08.001
0306-4379/

heterogeneous provenance reflects specific and distinct perspectives on the city, we need to investigate whether and how to reconcile the possibly diverging "pictures" of the urban environment those sources convey. Using a human relations metaphor, we explore if diverse urban datasets "date each other" and show "natural affinities".

Moreover, since heterogeneous sources come with different spatio-temporal characterizations (spatial resolution and/or temporal reference), diverse datasets need to be pre-processed and transformed to become comparable. The multi-faceted nature of data complexity can therefore change when changing observation granularity, so we need to distinguish "love at first sight", i.e. possibly high correlation between data at a coarse-grained level, with "friendship at a deeper look", i.e. different affinities at a fine-grained resolution.

In this paper, we present the results of our investigation on a number of diverse datasets related to the city of Milano in Italy and our analysis is aimed at mining relations between those information sources. One of those datasets is a very large call data record set from a telecommunication operator and we specifically focus our exploration on the correlation between mobile data and the expensive-to-maintain information sources.

The remainder of the paper is as follows: Section 2 introduces the characteristics of urban datasets and the main challenges in their processing; Section 3 details the information sources about Milano used in this research, while peculiarities of call data records are provided in Section 4; Section 5 illustrates our "data dating" experiments based on a relation-seeking [1] approach and, specifically, correlation analysis (Section 5.1), regression analysis (Section 5.2) and clustering analysis (Section 5.3), with an increasing level of data complexity; related works are presented in Section 6, and Section 7 concludes the paper with some perspectives on our future work.

2. Availability of urban datasets and challenges in their analysis

Digital information about cities abound today. The sources of such information are constantly increasing, due to the pervasiveness of information and communication technologies in the so-called Smart Cities domain. In this section, without the claim of being comprehensive, we would like to give an overview of the possible urban-related datasets that can be found today and of some of the challenges in those datasets management, manipulation and analysis.

With the advent of the *open data* movement, with its call for transparency and knowledge sharing, a very large number of data sources has been made available on the Web, through a new generation of CMS systems able to give access to this wealth of information, often originating from public bodies and research activities. With special reference to urban information, local authorities have started publishing numerous datasets referring to the city environment: demographics and statistics from municipalities (e.g. distribution of population, family income, crime statistics), listing of local businesses from chambers of commerce, various levels of descriptions about the environment from an urban planning perspective (e.g. land use or land cover, cadastre information), and so on.

Moreover, the popularity of sensor technologies and the so-called Internet of Things (IoT) has led to the availability of massive *real-time and streaming information*, like climate sensors from environmental agencies (e.g. temperature, pressure, humidity and other ecosystem measures), smart meters and GPS traces from public utilities (e.g. energy consumption or public transportation position).

After the Web 2.0 boom, also *user generated information* about cities has become ubiquitous. Crowdsourcing initiatives like OpenStreetMap[1] have popularized the Volunteered Geographic Information paradigm of "citizens as sensors" [2] and have collected data about different kinds of points of interest in urban environments (e.g. monuments, restaurants, public services). Location-based social networks like Foursquare, Twitter, Flickr have also produced a stream of "check-ins" and geo-located information that represent the digital counterpart of human activities in the urban space.

It is important to note that while a large part of the aforementioned sources can be considered open or at least openly accessible, there exist also *closed data sources* produced and maintained by private businesses, which provide specific perspectives on what happens in our cities. Examples of this kind of datasets are public utilities information, including telecommunication operators: as collateral effect of the mobile networking services, telco companies collect data about the phone activity over time and also over space (due to the positioning of transceiver towers). This example of data is a strong indicator of people presence and movement in the urban environment.

Managing, processing and comparing those diverse urban datasets can be cumbersome. Besides common issues like dealing with data scale and improving data quality, we would like to highlight some challenges that emerge from the specific case of comparison between datasets referring to the same geospatial environment.

One issue arises from the *varying spatial resolution* of information sources: being produced by diverse actors for different reasons, it is quite common that datasets are heterogeneous in terms of the geospatial extent they refer to. For example, population statistics could be at municipality level, land use information from cadastre could be at building level, and smart meters measures could refer to individual points (identified by latitude and longitude). This means that those sources are not immediately comparable.

Information sources can also refer to *different time-frames*: population census is usually done every n years, while sensor information is potentially provided in real-time; some other private/closed data sources can be made available as historical dumps, while IoT data can have different frequency updates (every 10 min vs. once a day). Directly comparing those sources can lead to poor results, because connections and correlations could be traced between datasets that give different pictures about the environment. Moreover, because of the time-frame, as well as because of the data provider, data sources can differ in terms of *reliability*: again apart from data quality issues, in managing and processing different datasets it is important to take into account whether and to what extent

[1] Cf. http://www.openstreetmap.org/.