Linked Open Data Mining for Democratization of Big Data

Roberto Espinosa, Larisa Garriga
WaKe Research
Universidad de Matanzas `Camilo Cienfuegos`
Cuba
e-mail: {respinosa, larisa.garriga}@umcc.cu

Jose Jacobo Zubcoff, Jose-Norberto Mazón
WaKe Research
Universidad de Alicante
Spain
e-mail: {jose.zubcoff, jnmazon}@ua.es

Abstract—Data is everywhere, and non-expert users must be able to exploit it in order to extract knowledge, get insights and make well-informed decisions. The value of the discovered knowledge from big data could be of greater value if it is available for later consumption and reusing. In this paper, we present an infrastructure that allows non-expert users to (i) apply user-friendly data mining techniques on big data sources, and (ii) share results as Linked Open Data (LOD). The main contribution of this paper is an approach for democratizing big data through reusing the knowledge gained from data mining processes after being semantically annotated as LOD, then obtaining Linked Open Knowledge. Our work is based on a model-driven viewpoint in order to easily deal with the wide diversity of open data formats.

Keywords- big data; linked open data; data mining

I. INTRODUCTION

Nowadays, governments and organizations are worldwide generating open data for the sake of transparency. Open data philosophy encourages the value of reusing data through participation and collaboration among citizens, public institutions and private organizations. The promise of open data is to improve citizens' life through (i) development of applications (Web, smartphones, etc.) that reuse and add value to existing open data, and (ii) data analysis to get new insights and acquire knowledge that support daily decision making process. While the former is an opportunity for entrepreneurs to go for novel business models based on data, the latter represents a point of view in which open data is moving towards a big data scenario in which every stakeholder states that the more data one has, the more analysis one can perform, and then, the more informed decisions one can make, i.e., the increasing availability of these open data is a great opportunity for everyone to take advantage of their analysis.

Unfortunately, this could be only true for experts in data analysis (the so-called, data scientists) that can manage the great amount of heterogeneous open data or for those companies that may hire them; but, what about regular citizens with no abilities to analyze data? They should be able to exploit available data in order to extract knowledge, get insights and make well-informed decisions in their daily life. Otherwise, we are attending to the “big data divide” in which only a little percentage of experts are able to take advantage of big data, while regular people is missing this opportunity [1].

There are some available tools that try to bridge this big data divide, e.g., Google Fusion Tables 1 or Tableau Public 2. However, although they provide some initial data analysis capabilities (i.e., some kinds of charts and visualizations on maps), the knowledge discovery is somehow limited. Importantly, data mining is one of the most prominent techniques to discover implicit knowledge patterns, thus gaining richer insights into data. However, non-expert users may find complex to apply data mining techniques to obtain useful results, due to the fact that it is an intrinsically complex process in which (i) a great number of algorithms can be applied to solve the same problem with different outcomes, and (ii) correctly applying data mining techniques always requires a lot of manual effort for preparing the data sets according to their quality. Consequently, successfully applying data mining requires the know-how of an expert in order to obtain reliable and useful knowledge in the resulting patterns.

Democratization of big data therefore requires that data analysis techniques, such as data mining can be used by everyone. To do so, we envision an approach in which semantic information provided by Linked Open Data (LOD) is used to guide non-expert users in the application of data mining techniques, as LOD allows for helping in exposing, sharing, and accessing data by using URIs and RDF [2]. Also, traditionally, once someone has applying some data mining techniques to discover some knowledge from data, it cannot be incorporated to the body of LOD as new knowledge, thus hampering its reuse. In order to overcome this pitfall, metadata on how the knowledge can be

1http://tables.googlelabs.com/
2 http://public.tableausoftware.com
discovered from data sources (by using our previously defined user-friendly data mining approach [3]) must be incorporated as new data sources in the LOD sets (see Figure 1 for an overview of the process).

Our approach is based on well-known software engineering techniques (based on model-driven development [4]) in order to being able to make the process as automated and simple as possible. In the next section, more details on our model-driven approach are given.

II. A MODEL-DRIVEN PERSPECTIVE FOR BIG DATA DEMOCRATIZATION BY USING LINKED OPEN DATA

Our approach for user-friendly data mining methodology [3] was designed for enables “masses” to apply data mining. It is the foundation of our approach for democratization of big data by using LOD. This user-friendly approach is based on the construction of a knowledge base that contains information of previous data mining experiments in order to provide guidance to non-expert users to apply data mining techniques. To generate our knowledge base, a model-driven approach is defined, based on a scientific workflow, specifically using a Taverna workbench tool 3. Model-driven software development provides mechanisms to define models and metamodels, and transformations to obtain the corresponding code from models in an automatic way. Specifically, three main contributions are provided:

- A method to create a knowledge base to collect all the information about what an expert data miner considers relevant for applying data mining algorithms to data sources.
- A tool that provides the expert user to set the required parameters for the construction of a data-mining recommender system based on the existing knowledge base.
- A method to allow non-expert user to transparently use a recommender in order to get the data mining algorithm to apply in a user-friendly manner.

As shown in Figure 1, the input of the user-friendly data mining is an arff file (well-known data input format for data mining), and the output is the information obtained in the mining process.

Although big data today comes in all types of formats, there are a lot of actions to do in order to analyze it and reusing the extracting knowledge. In order to describe our vision for obtaining knowledge, we need two sub-processes that allow us to complement the user-friendly data mining approach: the first that we have identified as pre-process and the second named post-process. In the pre-process phase the required transformations are implemented, in order to standardize the input data, to apply the proposed user-friendly data mining approach. While in the post-process phase, the mechanism to reusing the acquired knowledge is introduced.

Our approach is based on a model-driven viewpoint in order to easily deal with the wide diversity of open data formats. Each of those formats has its own data structure. Understanding each of them is quite difficult for non-expert users that try to acquire the right knowledge in order to make informed decisions. Fortunately, we also found common elements in the diversity of open data formats, in such a way that we can propose a homogeneous model to represent data independently of its format. To this end, the data metamodel designed is shown in the Figure 2.

In the pre-process stage we point out how an input open data source is transformed towards the data model according to the metamodel previously presented. Taking into account the possible schemes of the open data sources (RDF, JSON or CSV), we have implemented the correspondent grammars. Grammars were implemented by using the XText framework 4 within the Eclipse Modeling Framework 5. Once the grammar was defined we can obtain the correspondent models.

For each one of the possible input schemas of data sources, ATL transformations (Atlas Transformation Language) [5] are implemented to obtain the corresponding model in accordance with our data metamodel. Through this process, the uniformity of data in the mining process is guaranteed. Then, a model-to-text transformation is required to transform the data models to obtain a correct arff file as the input for our user-friendly data mining approach.

All the information that is generated in the applied data mining process [3] is stored in a model that conforms to the metamodel of Figure 3 (named as DMKB metamodel).

The question to solve is: how a semantically annotated RDF file can be obtained that contains the information included in the returned model? To address this problem in the post-process phase, a metamodel according to RDF files is presented, in order to apply ATL transformations between the metamodel containing all the data mining metadata (the

3 Taverna is an open source and domain-independent Workflow Management System, a suite of tools used to design and execute scientific workflows. Available at http://www.taverna.org.uk/

4 http://www.eclipse.org/Xtext/

5 http://www.eclipse.org/modeling/emf/
DMKB metamodel) and the proposed RDF metamodel (See Figure 4). The execution of this transformation returns a RDF model which can be serialized in an RDF document containing the description of the EMF objects. The RDF metamodel defines a simple model to describe relationships among resources in terms of designated properties and values.

The main advantage of our proposal consists of the potential for reusing the acquired knowledge in the data mining process. The contribution of our proposal is given by the generation of a RDF file semantically annotated with all the knowledge obtained in the data mining process. This output is very innovative because, in this way, useful information is shared and used by another user (see Figure 5).

The main benefit of using a metamodel that represents RDF files is to generate the result that is in a DMKB model in a RDF file to include like linked open knowledge. The obtained knowledge is part of open data philosophy, being published and ready to be reused. The obtained knowledge forms a new layer on the same data for future analyses. The cycle is closed since this information can be examined and enriched insofar as the data is analyzed by the user-friendly data mining proposal.

Users can then reuse the LOD files published on the Web. While semantically annotated these may be associated with other LOD files. When these are analyzed by our proposal, they become Linked Open Knowledge sources. The Linked Open Knowledge sources can also be associated to other Linked Open Data sources, thus enriching the available big data.

### III. CONCLUSIONS

Nowadays, it is essential that non-expert users can exploit the vast amount of heterogeneous available data in order to extract knowledge and make well-informed decisions in daily life. Also, the value of the previous discovered knowledge could be of greater value if it is available for later consumption for other users. In this paper, we present an infrastructure that allows non-expert users to apply user-friendly data mining techniques in Linked Open Data files to bridge the bid data divide. The main contribution of this paper is the concept of reusing the knowledge gained from data mining processes after been semantically annotated in the RDF file (Linked Open Knowledge). A model driven approach is used in order to maintain a standard structure having in account the diversity of the data formats.

### ACKNOWLEDGMENT

This work has been partially funded by Open.Mind project (GV/2014/098) from Generalitat Valenciana (Spain).

### REFERENCES


