Knowledge-Based Systems 000 (2016) 1–15

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Statistically-driven generation of multidimensional analytical schemas from linked data

Victoria Nebot*, Rafael Berlanga

Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Campus de Riu Sec, 12071 Castellón, Spain

ARTICLE INFO

Article history:
Received 8 January 2016
Revised 4 July 2016
Accepted 6 July 2016
Available online xxx

Keywords:
Linked data
RDF
Multidimensional models
Statistical models

ABSTRACT

The ever-increasing Linked Data (LD) initiative has given place to open, large amounts of semi-structured and rich data published on the Web. However, effective analytical tools that aid the user in his/her analysis and go beyond browsing and querying are still lacking. To address this issue, we propose the automatic generation of multidimensional analytical stars (MDAS). The success of the multidimensional (MD) model for data analysis has been in great part due to its simplicity. Therefore, in this paper we aim at automatically discovering MD conceptual patterns that summarize LD. These patterns resemble the MD star schema typical of relational data warehousing. The underlying foundations of our method is a statistical framework that takes into account both concept and instance data. We present an implementation that makes use of the statistical framework to generate the MDAS. We have performed several experiments that assess and validate the statistical approach with two well-known and large LD sets.

1. Introduction

The vision of the Semantic Web (SW) is to create a common framework that allows data to be shared and reused at different levels (i.e., between applications, enterprises and communities). The most tangible realization of the SW is the Linked Data (LD) cloud, which contains around 85 billion triples over 10 thousand different datasets (as of September 2015)[1] and it is expressed using the Resource Description Framework (RDF) [23] modeling language. Lately, communities from different areas as well as governments and public organizations have published large volumes of interlinked data in the LD cloud following the publication guidelines and providing the basis for creating and populating the Web of Data [14].

Given the explosive growth both in data size and also schema complexity and heterogeneity, LD sources are becoming increasingly difficult to understand and use, which limits the exploration and the exploitation of potential information they contain. This has brought to the fore the need for new tools able to explore, query, analyze and visualize these semi-structured, semantically-enriched and heterogeneous data sets [10]. While several different tools such as graph-based query builders, semantic browsers and exploration tools [4,5,7,15] have emerged to aid the user in querying, browsing and exploring LD, these approaches have limited ability to summarize, aggregate and display data in the form that a scientific or business user expects, such as aggregation tables and graphs. Moreover, they fall short when it comes to provide the user an overview of the potential data that may be of interest from an analytical viewpoint.

In this paper our hipothesis is that LD constitutes a valuable source of knowledge worth exploiting from a multidimensional (MD) perspective. Specially the Business Intelligence (BI) field can benefit enormously from adding and aligning the new knowledge uncovered from LD sources with the existing corporate data warehouses to make better and more informed decisions. BI uses the MD model to view and analyze data in terms of dimensions and measures, which seems the most natural way to arrange data. BI has traditionally been applied to internal, corporate and structured data, which is extracted, transformed and loaded (ETL) into a predefined and static MD model. The relational implementation of the MD data model is typically a star schema. The dynamic and semi-structured nature of LD poses several challenges to both potential analysts and current BI tools. On one hand, manual exploration of the datasets using the available browsers and tools to find MD patterns is cumbersome due to the heterogeneity and incompleteness of LD and the lack of support for obtaining informed summaries. Moreover, as the datasets are dynamic, their structure may change or evolve, making the one-time MD design approach unfeasible.

This paper approaches the exploration and discovery of potential analytical data from LD sources in a radical and innovative way. We propose a statistical framework to automatically discover can-

* Corresponding author.
E-mail addresses: romerom@uji.es (V. Nebot), berlanga@uji.es (R. Berlanga).

[1] http://stats.lod2.eu/.

http://dx.doi.org/10.1016/j.knosys.2016.07.010
0950-7051/

didate MD patterns hidden in LD sources. We call these patterns multidimensional analytical stars (MDAS) . A MDAS is a MD star-shaped pattern at the class level that encapsulates an interesting MD analysis [26]. The main innovations and contributions of this research to the LD and BI community are stated below:

- We define the concept of MDAS as a mapping of the MD model to a statistical layer on top of LD sources. This statistical layer is able to deal both with heterogeneity and incompleteness common in LD sets and feeds itself from the RDF schema graph elements and the instance data. We characterize and identify the facts and potential dimensions and measures that compose a MDAS in terms of classes and properties, whose underlying pattern has been inferred in the statistical layer.
- We develop a statistical framework to approach the problem of discovering MDAS in a foundational and automatic way. The automation of the process relieves the analyst from the burden of having to explore the dataset to become acquainted with it. Moreover, the discovery of the MDAS is driven both by the implicit semantics of the data and the statistical arrangement of the triple instances.
- We overcome the long-term known issues of data heterogeneity and incompleteness in the LD world. Both issues arise from the very same nature of LD and, in particular, the RDF modeling language, which does not impose a hard schema on the instance data. This modeling approach provides more flexibility than traditional modeling approaches based on hard schemas and constraints but introduces the above-mentioned new challenges. The statistical nature of the developed approach is able to deal with both heterogeneity and incompleteness of the datasets and discovers different configurations of MDAS that capture such heterogeneity.
- The statistical nature of the approach allows us to use well-known sampling techniques to build the statistical model instead of using the complete dataset, which may not always be available (e.g., SPARQL endpoints) or is too large to be processed (e.g., Big Data).
- We provide an implementation that makes use of the statistical model to generate MDAS. The algorithms provide the analyst with all the pieces to compose and configure MDAS while ensuring the population with instance data. On one hand, we automatically generate the *bases* of the stars (i.e., the nucleus), where a ranking of properties according to their relevance for the star is presented so that the user can select properties aided by the statistical indicators. Moreover, we also generate dimension types to enrich the base of the star with potential dimensions and measures. Dimension types are organized into groups to alleviate heterogeneity among dimensions that are expressed syntactically different but are semantically similar.
- We present several experiments with two different and well-known LD sets and assess the quality of our statistical model to generate MDAS.

The structure of the paper is as follows. In Section 2 we motivate our approach with an example. Section 3 presents preliminary concepts. Section 4 presents the main foundations that underlie our approach. That is, we present a model for MDAS over LD sources. Section 5 defines the statistical model developed to approach the problem of generating MDAS and the implementation. Section 6 presents the experiments and results. In Section 7 we review the literature related to the problem of analyzing LD and Section 8 gives some conclusions and future work.

2. Motivating example

We will now illustrate the need for an automatic and statistical approach to infer MD patterns (i.e., MDAS) from LD sources. We use the Enipedia[2] dataset for the examples. Enipedia is an initiative aimed at providing a collaborative environment through the use of wikis and the SW for energy and industry issues. They provide energy data from different open data sources structured and linked in RDF. Fig. 1 shows an "ideal" RDF knowledge base that models information about `Powerplants`, which have a `Country`, the type of `Fuel` and the `carbonemmissions-23kg` associated (see the upper schema part). The lower part of the figure has descriptions about resources, e.g., powerplants (&r1, &r3), countries (&r4, &r5) and fuel type (&r2, &r7). The resources are connected to other resources or literals by an arc if there is an `rdf:property` relating them. For example, the fact that *&r1* is connected to *&r4* by a `state` property means that the resource represented by *&r1* is located in the resource represented by *&r4*. In addition, resources are connected to resource classes using the `rdf:type` property, which indicates that the resource is a member of the class it is connected to. For example, resource &r1 is a `Powerplant`. In the schema, classes and properties may also be related by `rdfs:subclassOf` and `rdfs:subpropertyOf` properties, respectively, indicating a hierarchy on classes and properties. Also, the domains and ranges of a property may be defined using the `rdfs:domain` and `rdfs:range` properties.

In the previous ideal scenario, where all resources comply with a well-defined schema, it is easy to see that, by exploring the schema, we can find interesting associations between classes that resemble MD patterns. For example, from the previous schema, the class `Powerplant` is a good candidate for being the fact, the classes `Country` and `Fuel` can act as dimensions and the `carbonemmissions-23kg` property can act as the measure.

Unfortunately, the real LD world could not be further from the previous ideal scenario. The flexibility of the RDF model and the decentralized approach for creating and publishing LD has resulted in very heterogeneous datasets, where not only different datasets modeling the same domain objects are annotated with different schemas, but also, within datasets, the schemas are usually very poor (if non-existent) and the annotation of resources is incomplete and incoherent. Therefore, the exploration and use of these datasets for analytical purposes becomes increasingly difficult. Fig. 2 shows an example of the heterogeneity in the Enipedia dataset, where four resources of type `Powerplant` are presented. We observe that different properties are used to refer to the location of the powerplant (i.e., `country`, `state`, `county`), with no defined relation among them neither domain and range definition at the schema level. It is also frequent that a property is used with different domains and ranges. The class representing the location also varies from `Country` to `OECDMember` with no relation between them. The same occurs with classes `Fuel` and `Energy_concept`, which represent apparently the same class. Also, not all powerplant resources have attached the same data properties (i.e., &r4 and &r7 have `carbonemmissions-23kg`, whereas &r10 has also the property `carbonemmissionsnextdecade-23kg` and &r1 has none).

In the current LD scenario, approaches that rely only on exploiting the schema to discover interesting associations and patterns in the data become insufficient, as in most of the cases, the only schema that we find of top of LD sets are the semantic types of the resources, that is, the classes. Even in the cases where data have a larger semantic layer on top, this schema is usually heterogeneous both in its shape and usage. Therefore, in this paper we approach the problem by proposing a solution that combines both the semantics provided by the existing schema and statistical information derived from the instance data.

[2] http://enipedia.tudelft.nl/wiki/Main_Page.

Please cite this article as: V. Nebot, R. Berlanga, Statistically-driven generation of multidimensional analytical schemas from linked data, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.07.010