Finding, Assessing, and Integrating Statistical Sources for Data Mining

Karin Becker[1], Xiaojie Tan[2], Shiva Jahangiri[3], and Craig A. Knoblock[3]

[1] Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
karin.becker@inf.ufrgs.br
[2] Nanjing University, Nanjing, China
tanxjnj@gmail.com
[3] University of Southern California, Los Angeles, California, USA
knoblock@isi.edu

Abstract As the knowledge discovery process has been widely applied in a variety of domains, there is a growing opportunity to use the Linked Open Data (LOD) cloud as a primary data source for knowledge discovery. The tasks of finding the relevant data from various sources and then using that data for the desired analysis are the key challenges. There is a striking increase on the availability of statistical data and indicators (e.g. social, economic) in the LOD, and the Cube ontology has become the *de facto* standard for their description according to a multi-dimensional model. In this paper we discuss a detailed scenario for using the LOD as a primary source of data for building analysis models in the Peacebuilding domain. Next, we present an approach to finding potentially relevant cube datasets in the LOD cloud, assessing their compatibility, and then integrating the compatible datasets to enable the application of data mining algorithms.

Keywords: LOD, Cube ontology, data integration, knowledge discovery

1 Introduction

As the knowledge discovery process has matured and is widely applied in a variety of domains, and advanced mining algorithms/tools become available, there is a growing opportunity to use the LOD as a primary data source for knowledge discovery. For instance, statistical data, enriched by other types of data, can be used to develop models that enhance awareness (e.g. rise of undernourishment), understanding (e.g. identifying contributing factors to the rise of undernourishment), and forecasting (e.g. predicting future undernourishment) on relevant aspects of society. Obtaining data from various sources and integrating them for a given analysis purpose are key tasks in this scenario.

The most recent LOD status report [8] reveals an amazing growth in the number of government statistical datasets. Most datasets adopt the Cube vocabulary [7], a W3C recommendation for publishing multidimensional data. The Cube vocabulary establishes that *datasets* contain *observations* about *measures*,

according to one or more *dimensions*; where a *data definition structure* explicitly describe the structure and semantics of the respective observations.The importance of the Cube ontology is such that different projects are focused on tools for using, publishing, validating and visualizing cube datasets [1,4,3,5,6].

In this paper we discuss a detailed scenario using the LOD as a primary source of data targeted at building analysis models in the Peacebuilding domain, and then we present an approach for dealing with the following issues:

- Existing work assume consumers have previously chosen the cube datasets to be queried, visualized and integrated. However, consumers might need support in finding the cube datasets that are relevant for their analysis purposes as a prior step. Such discovery must take into account that: a) datasets may have been modeled using different strategies despite using a common vocabulary, and b) relevant datasets may be distributed in the LOD cloud, i.e. accessible through different endpoints ;
- Data integration needs to consider additional issues within the context of multidimensional datasets. Measures can only be integrated in a coherent manner if they are subject to common dimensions (i.e. measures of the same granularity). Common dimensions represent the same real-world entity, though they may have different representations in the cloud, i.e. described using distinct ontologies or data types. Compatibility rules need to be established whenever resources are not linked explicitly or implicitly.
- Observations for entities of interest may be distributed in different cube datasets that need to be combined. For instance, observations for GDP may be published in several datasets (e.g. one per year), while observations for inflation may be published in a single dataset for a whole decade.

2 An illustrative analysis scenario

Peacebuilding is defined by the United Nations as the "range of measures targeted to reduce the risk of lapsing or relapsing into conflict by strengthening national capacities at all levels for conflict management, and to lay the foundation for sustainable peace and development". Different organizations contribute to awareness and monitoring of contributing factors for peacebuilding such as the Food and Agriculture Organization (e.g. food security indicators), the World bank (e.g. economic indicators), the Fund For Peace (e.g. Fragile States Indicators - FSI), among others. Each organization provide data in a proprietary format.

Assume an analyst wants to develop a predictive model for an FSI index (e.g. Poverty and economic decline) based on historical data. According to the FSI methodology, this indicator is valued based on factors such as inflation, unemployment, GDP, etc. The analyst would browse different portals seeking for relevant indicators, download files, and use preprocessing functions available in a mining framework or database software to integrate, clean, and select relevant data, an activity that is labor-intensive, time-consuming, and error-prone.

Fig. 1 depicts the approach proposed. Initially, the analyst provides (1): a) seed concepts that characterize the variable to be predicted (economic decline), as

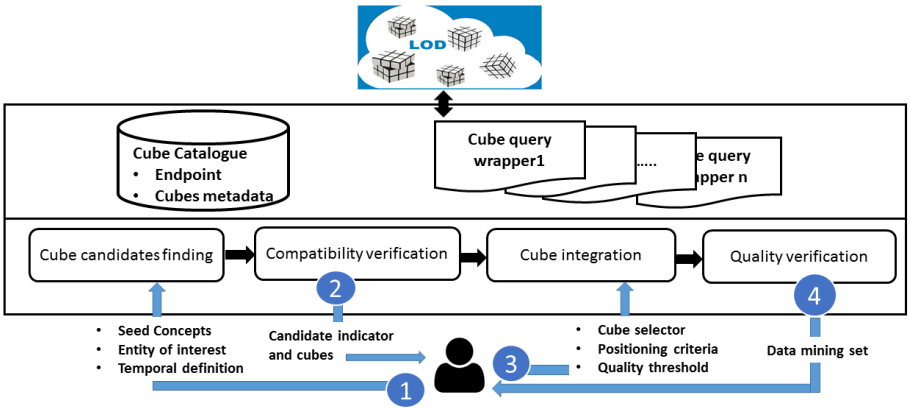

Figure 1. Cube discovery and integration framework

well as the type of features he is willing to consider (e.g. inflation, unemployment, GDP); b) the entity of interest (e.g. Country) c) the temporal dimension definition, including the unit and range (e.g. year 2004-2014). As a response, the system will provide a set of recommendations (2), which include indicators related to the concepts provided, together with the corresponding cube datasets and information further describing them (e.g. label, description). An important aspect is that these datasets are compatible with each other, and thus can be integrated. The user then inspects these recommendations, and selects a subset of them (3). He also provides new parameters, namely how measures and dimensions are to be disposed in the mining dataset as rows/columns, and quality thresholds (e.g. % of missing values). The system then retrieves the actual cube observations, constructs a mining dataset that organizes the features extracted, refines it by applying the quality thresholds, and outputs the resulting mining dataset (4). Finally, using an existing data mining framework, the analyst develops the remaining tasks for constructing the predictive model (e.g. feature selection, transformation, algorithm execution).

In our scenario, the recommendations might be:

- cubes of the World Bank Indicators endpoint [1], for indicators such as "Inflation, consumer prices (annual %)", "Inflation, GDP deflator (annual %)", "Unemployment, total (% of total labor force) (national estimate)", etc. Each indicator corresponds to one cube dataset, dimensioned by time and country.
- cubes of the FSI endpoint, for indicator "poverty and economic decline". Each dataset encompasses other measures in addition to this one, and they are dimensioned by country and year. However, each cube dataset correspond to a specific year or set of years (dataset 2006-2012, dataset2013, dataset2014);
- a cube of the Foodsecurity endpoint, for indicator "GDP". This cube contains also several other measures, and is dimensioned by country and year.

[1] http://worldbank.270a.info

The user would then select some of the indicators/cubes suggested. For example, there are more than 20 unemployment indicators (per age, sex, economic sector, etc.), and the user might select only a subset of these. He also selects the threshold of 80%, defines that country/year constitute the examples in the rows, and indicators are to be distributed in the columns. Finally, the system constructs the mining dataset.

3 Approach

In this section we describe our approach to finding relevant Cube datasets, assessing the compatibilty of those datasets, and then integrating them such that existing data mining algorithms can be applied over the resulting set (Fig. 1).

Finding candidate Cube datasets In order to locate potentially relevant Cube datasets, we start with a catalogue of Cube endpoints. The catalogue stores a set of access endpoints to cubes available in the LOD or to other accessible triple stores. Each endpoint is also related to graph information, as well as metadata about the cubes to which it provides access. This catalogue enables to search for all distributed sites, and makes it possible to discover potentially interesting data without requiring previous knowledge of the Cubes. The endpoints in our current Cube Catalogue were extracted from the Mannheim Catalogue [8].

The next step is to iterate over all entries of the catalogue to search for potentially interesting indicators. Metadata associated to each catalogue entry makes it possible to choose the appropriate Cube Query Wrapper for exploring the cubes available in each entry. A cube is relevant if it contains an indicator "similar" to a seed concept. At the current stage of the research, the similarity is based on labels and descriptions of measures, dimensions and/or related concepts. In the future, an ontology-based approach such as [2] may be considered. The result is RDF representing the relevant indicators, and the Cubes (data structure definition and datasets) in which they take part.

Query component wrappers aim at dealing with the different modeling styles of Cube datasets. The Cube ontology does provide a standard vocabulary, but there are many degrees of freedom that result in different styles of multidimensional modeling, possibly influenced by diverse backgrounds (e.g. BI, statistics). A qb:DataStructureDefinition (DSD) defines the structure of one or more datasets, in particular the dimensions and measures used, along with qualifying information (e.g. normalization). A qb:Dataset conforms to the structure of a DSD, and contains qb:Observations. A DSD specify components that are related to qb:DimensionProperty and qb:MeasureProperty resources. The semantics of dimensions and measures can be associated in different ways: from simple labels and description properties using a popular vocabulary (e.g. rdfs, skos), to a qb:concept property related to a skos:Concept. In our scenario, the FSI and the Foodsecurity use different styles: the former makes explicit the semantics of dimensions and attributes by relating their definitions to concepts, whereas FoodSecurity merely associate labels/descriptions to them. Other styles exist (e.g. World Bank). Thus, when looking for relevant measures/dimensions, these different styles need to be

taken into account, and the query wrappers enable to abstract from modeling idiosyncrasies.

Assessing Candidate Dataset Compatibility Measures can only be integrated in a coherent manner if they are subject to common dimensions (i.e. represent same granularity). The first issue is to verify the similarity between the user input and the DSD dimensions, resulting in a set of candidate dimensions. Different properties might qualify a dimension, such as labels, descriptions or concepts. For instance, WorldBank use a dimension labeled "Calendar Year" without a range defined, whereas FoodSecurity and FSI use a dimension labeled "Year" and range rdfs:gYear. The second issue is to verify the compatibility among the candidate dimensions for a same concept, e.g. different dimension representing "year". To this end, several strategies can be employed, including explicit linkage between concepts (e.g. owl:sameAs between concepts in different vocabularies or between resources) to the deployment of ontology alignment techniques [9]. The final verification is that the DSDs contain the same number of dimensions, and that they are compatible. As an initial restriction, we are assuming just two dimensions, corresponding to the entity of interest and time definition provided by the user. It is also necessary to develop compatibility rules that transform the values of compatible but not similar dimensions to a common representation. The result is a set of recommended indicators, together with the respective DSDs and datasets.

Cube Dataset Integration Given a (sub)set of indicators and their respective cubes, the disposition criterion for rows/columns and a quality threshold, the final step is to integrate the compatible datasets. That integration needs to be both horizontal (join of datasets with different DSDs) and vertical (union of datasets of a same DSD). Using pairs of values for the two dimensions as joining criteria, the goal is to produce a table that contains, for each of pair, the indicators selected. In our example, a column will be created for Country, another for Year, and there will be one column for each selected indicator. Then, observation data will be retrieved and the table filled accordingly. The query components need to care for normalization definitions within the DSDs for correct retrieval of observations [7].

The vertical integration refers to observations that are distributed in different cubes, but refer to the same DSD. For example, FSI observations are grouped in three datasets due to publishing criteria. For the period 2004-2014 of the scenario, union must be applied to the observations in these three datasets.

4 Related Work

LOD2 Statistical Workbench [1,5], OpenCube [3] and OLAP4LD [4] are platforms that support using, publishing, validating and visualizing Cube datasets. The LOD extension [6] provides a set of operators for the Rapid Miner mining framework to deal with the LOD, including the retrieval of cube datasets, and linkage of examples with knowledge available in the LOD. None of these works deal with cube discovery or integration of cubes. An approach to find concepts

in an ontology given a seed concept, used as basis to generate mining datasets, is presented in [2], but it does not address multidimensional data. Thus our approach is complementary to these works.

5 Conclusions

In this paper we presented and illustrated an approach to finding potentially relevant cube datasets in the LOD cloud, assessing their compatibility, and integrating them to generate a mining dataset. We are currently developing a survey on the state of the practice of the Cube adoption in the LOD, and experimenting this approach in the Peacebuilding domain. We are particularly focused on the following aspects of the approach: automatic generation of query wrappers, based on cube examples of different modeling styles; formalization of dimension compatibility verification algorithms; the use of more sophisticated similarity functions for suggesting indicators [2]; ranking functions for prioritizing suggestions; wider quality assessment features; integration with mining frameworks; among others.

Acknowledgments This research is sponsored in part by a gift from the Underwood Foundation and by CAPES - Brazil.

References

1. Janev, V., Mijovic, V., MiloSevic, U., Vranes, S.: Supporting the linked data publication process with the lod2 statistical workbench. Semantic Web Journal (2014)
2. Janpuangtong, S., Shell, D.A.: Leveraging ontologies to improve model generalization automatically with online data sources. In: Proceedings of the 27th Conference on Innovative Applications of Artifitial Intelligence, Austin, USA, Jan. 2015
3. Kalampokis E. et al: Exploiting linked data cubes with opencube toolkit. In: Proceedings of the ISWC 2014 Posters & Demonstrations Track, Riva del Garda, Italy, Oct. 2014. pp. 137–140 (2014)
4. Kämpgen, B., Harth, A.: OLAP4LD–A Framework for Building Analysis Applications Over Governmental Statistics. In: The Semantic Web: ESWC 2014 Satellite Events, pp. 389–394. Springer International Publishing (2014)
5. Mader, C., Martin, M., Stadler, C.: Facilitating the exploration and visualization of linked data. In: Linked Open Data – Creating Knowledge Out of Interlinked Data, pp. 90–107. Springer International Publishing
6. Paulheim, H.: Exploiting linked open data as background knowledge in data mining. In: Proceedings of the International Workshop on Data Mining on Linked Data (DMoLD), Aachen, Germany, 2013 (2013)
7. Reynolds, D., Cyganiak, R.: The RDF Data Cube vocabulary. Tech. rep., W3C Recommendation (Jan 2014), `http://www.w3.org/TR/2014/REC-vocabdata-cube-20140116`
8. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: 13th International Semantic Web Conference, Riva del Garda, Italy, Oct. 2014. 245–260 (2014)
9. Taheriyan, M., Knoblock, C.A., Szekely, P., Ambite, J.L.: A graph-based approach to learn semantic descriptions of data sources. In: Proceedings of the 12th International Semantic Web Conference (ISWC 2013) (2013)