Open Business Intelligence: on the importance of data quality awareness in user-friendly data mining

Jose-Norberto Mazón, Jose Jacobo Zubcoff, Irene Garrigós
University of Alicante
Alicante, Spain
{jnmazon,igarrigos}@dlsi.ua.es,
jose.zubcoff@ua.es

Roberto Espinosa, Rolando Rodríguez
University of Matanzas
Matanzas, Cuba
{roberto.espinosa,rolando.rodriguez}@umcc.cu

ABSTRACT

Citizens demand more and more data for making decisions in their daily life. Therefore, mechanisms that allow citizens to understand and analyze linked open data (LOD) in a user-friendly manner are highly required. To this aim, the concept of Open Business Intelligence (OpenBI) is introduced in this position paper. OpenBI facilitates non-expert users to (i) analyze and visualize LOD, thus generating actionable information by means of reporting, OLAP analysis, dashboards or data mining; and to (ii) share the new acquired information as LOD to be reused by anyone. One of the most challenging issues of OpenBI is related to data mining, since non-experts (as citizens) need guidance during preprocessing and application of mining algorithms due to the complexity of the mining process and the low quality of the data sources. This is even worst when dealing with LOD, not only because of the different kind of links among data, but also because of its high dimensionality. As a consequence, in this position paper we advocate that data mining for OpenBI requires data quality-aware mechanisms for guiding non-expert users in obtaining and sharing the most reliable knowledge from the available LOD.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining; D.2.10 [**Software Engineering**]: Design

General Terms

Design

1. INTRODUCTION

Citizens demand an increasingly transparent behavior of public institutions. Importantly, transparency implies that public data should be available with the aim of providing the greatest benefit to the wider society through an active participation of citizens. Therefore, public data should be freely available to be easily used, reused and redistributed by anyone, i.e. open data. Consequently, open data are generally shared as a raw data in formats such as CSV, XML or as HTML tables, without paying attention in structure nor semantics. Unfortunately, this fact prevents non-expert citizens from acquire actionable information from open data. Mechanisms that allow citizens to analyze and understand open data in a user-friendly manner are thus highly required.

To this aim, the concept of Open Business Intelligence (OpenBI) is introduced in this position paper. OpenBI facilitates non-expert users to integrate different open data sources and semantically annotate them, thus obtaining linked open data (LOD) [3], in order to (i) analyze and visualize LOD, thus generating actionable information by means of reporting, OLAP analysis, dashboards or data mining; and to (ii) share the new acquired information as LOD to be reused by anyone.

As a consequence, OpenBI requires the development of systematic approaches for guiding non-expert users in obtaining and sharing the most reliable knowledge from the available LOD. One of the most challenging issues is related to data mining, since non-experts (as citizens) need guidance during preprocessing and application of mining algorithms to obtain reliable knowledge.

According to the seminal work in [7], data mining is the process of applying data analysis and discovery algorithms to find knowledge patterns over a collection of data. Importantly, the same authors explain that data mining is only a step of an overall process named knowledge discovery in databases (KDD). KDD consists of using databases in order to apply data mining to a set of already preprocessed data and also to evaluate the resulting patterns for extracting the knowledge. Indeed, the importance of the preprocessing task should be highlighted due to the fact that (i) it has a significant impact on the quality of the results of the applied data mining algorithms [11], and (ii) it requires significantly more effort than the data mining task itself [9].

Importantly, when mining complex data such as LOD, the preprocessing task is even more time-consuming, because of the high dimensionality of complex data [11]. High dimensionality means a great amount of attributes difficult to be manually handled and making the KDD awkward for non-experts data miners. Specifically, high dimensionality implies several data quality criteria to deal with in the data

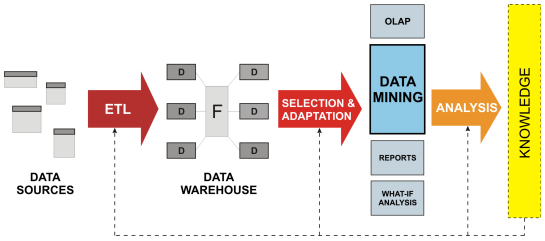

Figure 1: The KDD process: from the data sources to the knowledge

sources to ensure reliable knowledge discovery. Several statistical techniques have been proposed to deal with dimensionality reduction issue [8], such as PCA (Principal Component Analysis) or Regression Trees, among others. However, by using those techniques there are an important information lost: data structure cannot be considered. To overcome this situation, in [11] the definition of user-friendly data mining applications is suggested: data preprocessing should be automated, and all steps undertaken should be reported to the user or even interactively controlled by the user, at the same time that useful information is not lost.

Bearing these considerations in mind, in this position paper we advocate that data mining for OpenBI requires data quality-aware mechanisms for guiding non-expert users in obtaining and sharing the most reliable knowledge from the available LOD. To this aim, we propose to conduct a set of experiments to assess how different data quality criteria on LOD affect behavior of different data mining techniques, thus generating a knowledge base that can be used for guiding the non-expert users in the application of data mining techniques whilst reliable knowledge is obtained.

The remainder of this paper is as follows. Section 2 briefly describe some related work about data quality awareness mining. Section 3 defines our approach. Conclusions and future work are sketched in Section 4.

2. RELATED WORK

The KDD process (Figure 1) is summarized in three phases: (i) data integration in a repository, also known as preprocessing data or ETL (Extract/Transform/Load) phase in the data warehouse area, (ii) algorithms and attributes selection phase for data mining (i.e., the core of KDD), and (iii) the analysis and evaluation of the resulting patterns in the final phase.

Every phase of this process is highly dependent on the previous one. This way, the success of the analysis phase depends on the selection of adequate attributes and algorithms. Also, this selection phase depends on the data preprocessing phase in order to eliminate any problem that affects the quality of the data.

For the first phase of the KDD process there are some proposals that address the problem of data quality from the point of view of cleaning data: (i) for duplicates detection and elimination [5, 1], entity identification under different labels [14], etc.; (ii) resolution of conflicts in instances [15] by using specific cleaning techniques [16]; (iii) uncommon values, lost or incomplete, damaging of data [13], among others. Several cleaning techniques have been used to solve problems, such as heterogeneous structure of the data: an example is standarization of data representation, such as dates.

There are other approaches that consider data quality during the second phase of the KDD process. In [4], the authors propose a variant to provide the users all the details to make correct decisions. They outline that besides the traditional reports it is also essential to give the users information about quality, for example, the quality of the metadata. In [10], authors use metadata as a resource to store the results of measuring data quality.

However, data quality not only refers to cleansing procedures but a wider spectrum of criteria should be considered [17], for example complete, correlated and balanced data [6]. One of the main proposals in this sense was presented in [2] where Berti-Equille defined a method for measuring the quality of association rules obtained, with the aim of defining which are the best options to be applied.

Unfortunately, current related work overcomes the scenario of studying data quality issues when LOD are mining with the aim of guiding non-expert user in obtaining reliable knowledge when applying data mining techniques.

3. AN APPROACH TO GUIDE USERS IN APPLYING DATA MINING FOR OPENBI

This section describes our overall framework for guiding non-expert users in selecting the right data mining algorithm being aware of the data quality of the LOD sources. Our approach allows citizens to analyze and visualize LOD within the OpenBI scenario.

Our framework consists of two main stages (as shown in Figure 2): (i) conducting a set of experiments to analyze different data quality criteria on the LOD sources and how they affect results of data mining algorithms, thus creating a knowledge base; and (ii) using the knowledge base to give advice to non-expert users for selecting the most appropriate data mining algorithm to be applied on the available LOD.

3.1 Experiments for obtaining a knowledge base

A set of experiments to assess how different data quality criteria on LOD affect behavior of different data mining techniques should be conducted. The aim of these experiments is generating a knowledge base that can be used for guid-

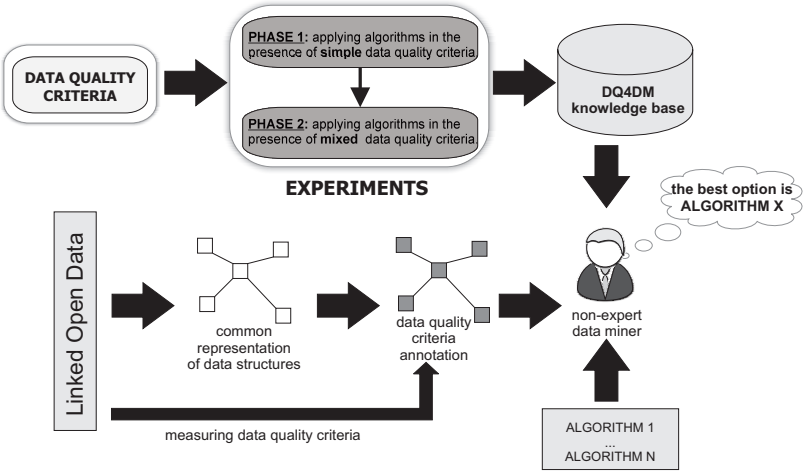

Figure 2: Overview of our approach for achieving data quality-aware mining in OpenBI

ing the non-expert users in the application of data mining techniques whilst reliable knowledge is obtained.

Data quality means "fitness for use" [14] which implies that the data should accomplish several requirements to be suitable for a specific task in a certain context. In KDD, this means that data sources should be useful for discovering reliable knowledge when data mining techniques are being applied. Our hypothesis is therefore, data quality criteria should be measured to avoid discovering superfluous, contradictory or spurious knowledge. This is specially true for high dimensional data as LOD, since a non-expert data miner without knowing in detail the domain of data can apply a data mining technique that provides misleading results. There are several data quality criteria which should be measured to determine the suitability of data for being used [6]. For example, if some attributes are selected as input for a classification algorithm (being some of them strongly correlated), the resulting knowledge pattern, though correct, will not provide the useful expected value. Therefore, those data quality criteria that may affect the result of data mining techniques should be determined in order to prevent non-expert user from using them in some scenarios.

Our method for preparing the knowledge base starts by using a initial and representative sample of LOD. This sample is manually "cleaned" to avoid data quality problems. From this initial dataset we will introduce some data quality problems in a controlled manner. This allows us to to test the incidence of data quality in the LOD sources. Then, an exhaustive analysis allows us to get some conclusions about the behavior of data mining algorithms handling a set of LOD sources with different data qualities. Finally, a knowledge base for optimal data mining on LOD can be obtained. Four main steps are proposed, namely:

1. Input data: apart from the LOD sources, our experiments take the user profile as input data. The user profile includes the data quality criteria to assess.
2. Data preparation: in this stage some dataset tests according to the user profile are created. Two kind of datasets are defined: the first one includes each of the data quality criteria individually, while the second one combines several data quality criteria.
3. Application of the experiments: the experiments are applied according to the type of techniques selected for the users in the first stage. First each simple data quality issue is considered individually, and in a second phase, a mixed set of data quality criteria is considered.
4. Knowledge base: results of experiments are included in a knowledge base.

Once a knowledge base is obtained, it can be used in OpenBI for a non-expert user to be aware of data quality when mining LOD.

3.2 Data quality aware mining of LOD

This section describes our approach for guiding user in applying mining algorithms on LOD. It consists of two steps: (i) creating a common representation of the LOD, and (ii) measuring data quality criteria of LOD sources to add them to the common representation. As shown in Fig. 2, our approach aims to give advice to data miners for selecting the most appropriate data mining algorithm for the available data sources.

3.2.1 Creating a common representation

The task of creating a common representation of LOD is based on model-driven development. One candidate for this purpose is the *Common Warehouse Metamodel* (CWM) [12]. CWM is a standard for representing data sources metadata, consisting of a set of metamodels that allow up to represent data structures and related information. Therefore, LOD can be extracted into a model which will be useful for being annotated with some measures calculated from data quality criteria.

3.2.2 Data quality criteria annotation

Once the common representation of LOD is contained in a model, data quality criteria are measured and added to it.

This annotated model is used for guiding non-expert users to choose the right data mining algorithm being aware of the data quality of the LOD sources.

3.3 Implementation

The model-driven process of obtaining a common representation from LOD can be implemented by using Java in the Eclipse Modeling framework (EMF)[1]. The EMF project is a modeling framework and code generation facility for building tools and other applications based on a structured data model. Eclipse has been conceived as a modular platform that can be extended by means of plugins in order to add more features and new functionalities. In that way, we have designed a set of modules encapsulated in a single plugin that provides Eclipse with capabilities for supporting our approach:

Data source module. It implements a common metamodel for data (e.g. CWM).

LOD integration module. Metadata should be obtained from LOD. From this metadata, the corresponding data model is obtained by using the previous module.

Data quality module. It implements each approach for measuring and storing each useful data quality criteria in the corresponding data source model.

4. CONCLUSIONS

In this position paper, an approach based on model-driven engineering is proposed for automatically measuring data quality criteria in order to support non-expert users in selecting the most adequate data mining algorithm for LOD sources. This work intends to be a first step towards considering, in a systematic and structured manner, data quality criteria for supporting non-experts data miners in obtaining reliable knowledge on LOD sources.

5. REFERENCES

[1] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *VLDB*, pages 586–597. Morgan Kaufmann, 2002.

[2] L. Berti-Equille. Measuring and modelling data quality for quality-awareness in data mining. In F. Guillet and H. J. Hamilton, editors, *Quality Measures in Data Mining*, volume 43 of *Studies in Computational Intelligence*, pages 101–126. Springer, 2007.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[4] R. H. L. Chiang, T. M. Barron, and V. C. Storey. Extracting domain semantics for knowledge discovery in relational databases. In *KDD Workshop*, pages 299–310, 1994.

[5] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16, 2007.

[6] R. Espinosa, J. J. Zubcoff, and J.-N. Mazón. A set of experiments to consider data quality criteria in classification techniques for data mining. In B. Murgante, O. Gervasi, A. Iglesias, D. Taniar, and B. O. Apduhan, editors, *ICCSA (2)*, volume 6783 of *Lecture Notes in Computer Science*, pages 680–694. Springer, 2011.

[7] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, pages 82–88, 1996.

[8] I. K. Fodor. A survey of dimension reduction techniques. *LLNL technical report*, (June):1–24, 2002.

[9] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[10] M. Jarke and Y. Vassiliou. Data warehouse quality: A review of the dwq project. In D. M. Strong and B. K. Kahn, editors, *IQ*, pages 299–313. MIT, 1997.

[11] H.-P. Kriegel, K. M. Borgwardt, P. Kröger, A. Pryakhin, M. Schubert, and A. Zimek. Future trends in data mining. *Data Min. Knowl. Discov.*, 15(1):87–97, 2007.

[12] Object Management Group. Common Warehouse Metamodel Specification 1.1. `http://www.omg.org/cgi-bin/doc?formal/03-03-02`.

[13] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[14] D. M. Strong, Y. W. Lee, and R. Y. Wang. 10 potholes in the road to information quality. *IEEE Computer*, 30(8):38–46, 1997.

[15] D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Commun. ACM*, 40(5):103–110, 1997.

[16] O. G. Troyanskaya, M. Cantor, G. Sherlock, P. O. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[17] X. Zhu, T. M. Khoshgoftaar, I. Davidson, and S. Zhang. Editorial: Special issue on mining low-quality data. *Knowl. Inf. Syst.*, 11:131–136, February 2007.

[1] `http://www.eclipse.org/modeling/emf/`