From Requirements to Data Analytics Process: An Ontology-Based Approach

Madhushi Bandara[1(✉)], Ali Behnaz[1], Fethi A. Rabhi[1], and Onur Demirors[1,2]

[1] University of New South Wales, Sydney, Australia
[2] Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey
{k.bandara,ali.behnaz,f.rabhi,o.demirors}@unsw.edu.au

Abstract. Comprehensively describing data analytics requirements is becoming an integral part of developing enterprise information systems. It is a challenging task for analysts to completely elicit all requirements shared by the organization's decision makers. With a multitude of data available from e-commerce sites, social media and data warehouses selecting the correct set of data and suitable techniques for an analysis itself is difficult and time-consuming. The reason is that analysts have to comprehend multiple dimensions such as existing analytics techniques, background knowledge in the domain of interest and the quality of available data. In this paper, we propose to use semantic models to represent different spheres of knowledge related to data analytics space and use them to assist in analytics requirements definition. By following this approach users can create a sound analytics requirements specification, linked with concepts from the operation domain, available data, analytics techniques and their implementations. Such requirements specifications can be used to drive the creation and management of analytics solutions, well aligned with organizational objectives. We demonstrate the capabilities of the proposed method by applying on a data analytics project for house price prediction.

Keywords: Analytics process · Requirements · Ontology

1 Introduction

Analytic projects are complex with large investments being made on data preparation, tools and knowledge workers. Data analytics processes differ from traditional repeatable processes, requiring frequent intervention from knowledge-workers and flexibility to adapt the process when new insights emerge. To engineer correct analytics solutions that match the respective business objectives is challenging [11]. From the high-level requirements declared by management to the final analytics model adopted there is a complex process involving different decision making such as selecting suitable tools, algorithms, data sets and how to generate results and report them accurately. If the outcome is not accurate enough, nor considered the appropriate context, nor incorporated the correct datasets, nor satisfied the stakeholders' requirements, their time, resources and money spent on the study are wasted [10].

F. Daniel et al. (Eds.): BPM 2018 Workshops, LNBIP 342, pp. 543–552, 2019.
https://doi.org/10.1007/978-3-030-11641-5_43

As a solution to these issues, we propose an approach to express analytics requirements accurately via semantic models in order to support decision making and drive the process composition. In semantic modelling, ontologies are used to capture the domain knowledge, to evaluate constraints over domain data, to prove the consistency of domain data and to guide domain model engineering [2]. As ontologies provide a representation of knowledge and the relationship between concepts, they are malleable models good at tracking various kinds of software development artifacts ranging from requirements to implementation code [9]. Though there is research devoted to utilizing semantic web technologies in requirements engineering, most of it concentrates on specific requirements artifacts such as goals and use-cases and does not support reasoning over relations between all concepts [12].

To study this further, we conducted a systematic mapping study [3] to identify how existing literature leverages semantic web technologies to realize different stages of the data analytics process. The findings reveal a gap between defining analytics goals and requirements and linking them to the actual process realizations. The goals or requirements defined in existing literature are either at a very high level, not linked to operational level or they are expressed at the query level, limiting their capabilities for declarative analysis.

In this paper, we discuss the potential and limitations of using semantic models to capture data analytics requirements, relating them to different domain knowledge spheres, and how such an approach can lead to a requirements driven process composition in data analytics. Those requirements can be used as a communication tool, an artifact that enables traceability between requirements and the analytics solution implementation and a knowledge-base to guide semi-automated analytics process composition. Section 2 of the paper discusses the background and related work. Section 3 presents our proposed solution- a system that uses ontologies to drive the requirements capturing of the analytics process. Section 4 presents the system capabilities via an application to a predictive analytics process and paper concludes in Sect. 5.

2 Background and Related Work

In this section some background on the space of requirements related to data analytics is discussed in details, followed by the related work on how different existing systems capture or manage requirements to support data analytics.

2.1 Analytic Process Requirements Space

The data analytics requirements an organization should define can be identified at strategic, operational or tactical levels [14]. Taylor [13], in his work on framing requirements for predictive analytics projects, presents multiple dimensions that need to be captured in order to define requirements of an analytics project.

1. Performance Measures - metrics or business objectives
2. Decision requirements - What decisions are aided by the analysis, which in return can affect the performance measures, metrics or business objectives
3. Business context - Details of the processes and systems impacted by analysis, organizational units and roles involved, measures and metrics that may impact the scope of analysis
4. Data requirements - Input data and information useful for the analysis
5. Knowledge requirements - External regulations, internal policies, organizational practices and existing analytics insights

Among those performance measures, decision requirements and business context fall under strategic level requirements while data and knowledge requirements fall under the operational requirements. Wiegers provides a set of guidelines to define strategic requirements around an analytics project [14]. Brijs, in his work [7], presents a list of dimensions to express data requirements with respect to the data source, data storage, extraction, management and governance.

When linking the strategic level requirements to the operational level, another dimension we need to consider in detail is the type of analysis necessary to support certain decisions. There is no complete classification of analytics needs or problems and it may vary by context.

In addition, there are non-functional requirements that need to be specified for an analytics process at the operational level such as the model training time, expected accuracy and memory footprint which may impact the strategic level requirements as well.

2.2 Role of Requirement Models in Engineering Analytics Processes

Semantic web technologies have been used to engineer different stages of the data analytics process [11] as well as to compose the analytics process. In the systematic mapping study mentioned earlier we found that data analytics related literature use different ontologies to capture four concept classes: domain related, analytics specific, service related and intent concepts. Domain concepts described application specific information such as health care and sensor data. Analytics concepts were focusing on representing data pre-processing, mining and statistical algorithms. Service related concepts captured the implementation details such as data importing or computing services and work-flow composition. Intent concepts were the category that captures data analyst's requirements or goals.

We identified 10 studies that used intent concepts. 8 of them are used to express requirements at the execution level such as the Analytica Queries (AnQ) model in [8] that facilitates expressing user queries that need to be performed on data. Two studies were focusing on representing the high-level goals of the analysts: the Scientist's Intent Ontology [10] and the Goal Oriented Model [6]. They facilitate the modelling of strategic level user goals such as the desired outcomes of analytics tasks yet fail to link that to the operational level requirements or tasks related to an analytics process.

Moreover, we observed that existing techniques only focus on facilitating the selection of data providers, web services, and computational software modules. Hence to a large extent, analytics requirements are still a part of the mental model of the developer or the analyst who performs these tasks. In practice, several iterations of data cleansing, reformatting, the model selecting and process composition may be required in order to optimally serve the analytics problem. This may result in less effective data analytics solutions whose performance is likely to degrade with time.

As the data analytics community is extending wider into different industries and organizations and with analytics contexts and requirements changing rapidly, it is necessary to explore techniques that consider all dimensions such as business requirements, context and constraints. Hence a potential research area is a study of how high-level user goals and context can be represented and incorporated into data analytics solution engineering through data integration, process construction, and result interpretation. Approaches that link the user intentions and context into analytics processes have the potential of changing static analytics models deployed today into dynamic and adaptable analytics models that change the behavior, responding to changes in user goals or the operational context.

As discussed throughout Sect. 2, there are multiple levels of requirements associated with an analytics process. Capturing them accurately via models, linking them to existing analytics knowledge can improve the analytics solution design, enabling consistency and constraint checking as well as the requirement-driven design of data analytics processes.

3 Proposed Solution

We propose a system for managing data analytics requirements supported by a semantic knowledge-base. Figure 1 illustrates the main components of the proposed system organized into three layers: user interface (UI), business logic and data. Each component is described in more detail below.

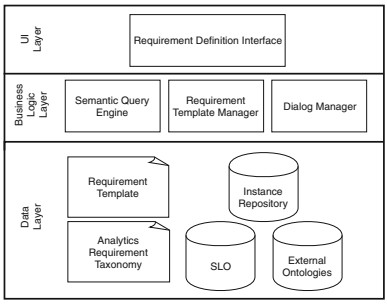

Fig. 1. Proposed system design

– **Data Layer**
At the core of the Data Layer is the Statistical Learning Ontology (SLO) [5] we designed to capture knowledge related to the data analytics space such as variables, prediction models and their relationships. The main components of the SLO are shown in Fig. 2.

Another component of the Data Layer is the Analytics Requirement Taxonomy (ART) which extends the different dimensions of analytics requirements discussed in Sect. 2.2 to suit the organization perspectives. Figure 3 represents the ART we use in the application discussed in Sect. 4. Requirements are classified as strategic and operational. Strategic requirements may be defined by the management of the organization and passed into the analysts, who will define operational requirements in line with them. As illustrated, users can extend the taxonomy to represent a set of requirements related to their domain of analytics.

The next component is the Instance Repository which stores actual data instances of the ontologies related to the analytics process. It contains details about the actual variables, existing prediction models and related publications, links between variables and models, available datasets and accessible data sources. This repository will be kept up to date, so the requirements definitions are generated based on the latest information.

A requirements template provides the structure for a set of requirements definitions. It is defined by selecting relevant requirements from the ART and linking them with the concepts of the ontology repository. Figure 4 provides an example requirements template we defined for predictive analytics with pretrained models. Each requirement in the template is linked to the associated concept in SLO and captures the dependencies and constraints imposed by one requirement on the others. In Sect. 4, Fig. 5 illustrates an instance of this requirements template. What concepts from the instance repository are mapped to define each requirement in the template is shown by associated numbers.

– **Business Logic Layer**
The semantic query engine is responsible for fetching the information from the instance repository to fulfill each requirement defined in the template. It will capture the decisions made by the user at each stage of requirements definition and use them to enrich queries in the following steps. For example, if the user selects a prediction model, independent variables will be selected to match that model.

The requirements template manager provides the ability to update or create new requirement templates.

The dialog manager coordinates the UI layer and communicates with template manager and the semantic query engine to drive requirements definition process. When a user wants to define particular requirements defined in the requirements template, the dialog manager fetches different options available in the instance repository through the semantic query engine.

– **UI Layer**
User interface layer provides an interface for requirements definition. It uses a dialog-based approach to capture strategic and operational requirements of an analytics process. The dialogs are driven by dialog manager, supported by the requirements template manager and the instance repository. This interface guides users to create an instance of the provided requirements template that matches his analytics needs.

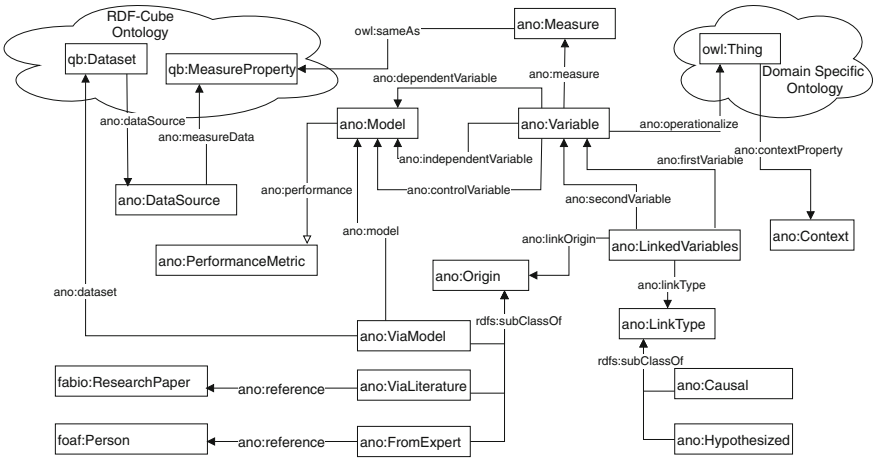

Fig. 2. Main components of the statistical learning ontology [5]

4 Application of the Proposed System

To illustrate how the proposed system can be used to define analytics process requirements specifications, we applied it to a process of developing a house price prediction model.

4.1 Application Description

This application aims to develop a framework which builds on existing predictive house price models and advance them through a range of approaches: studying and applying other econometric models, testing a range of additional economic variables, re-focusing predictive model on a 20-year horizon for Australian real estate prices in metropolitan areas. One key aspect of this study is voluminous heterogeneous datasets used in the study as house price prediction problem is addressed by the different experts with diverse perspectives. In addition, a plethora of analytics techniques (statistical learning techniques in this case) is used by users. As a result, knowledge acquisition plays a pivotal role in defining analytics requirements when implementing the project. The analytics

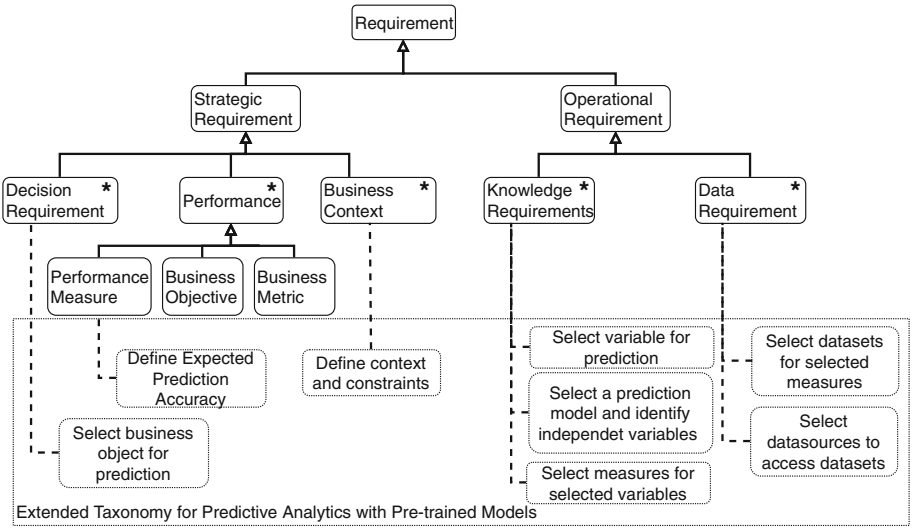

Fig. 3. Analytics requirement taxonomy - extension of the dimensions (represented by *) proposed by Taylor [13]

team comprised a group of academics from the two fields: computer science and economics. They had to go through literature and survey different prediction models, variables etc. used for house price prediction in different time spans and various countries. They use spreadsheets to accumulate findings from the literature. They focused on 30 previous studies and it was difficult for them to link those studies together, and to identify what studies are using similar models or variables. There was no naming convention, so the same variable was named differently or different names were used to refer to the same variable in different studies. Navigating through such spreadsheets and understanding was difficult and time-consuming. They needed a better approach to accumulate all that knowledge in 30 studies and pick the insights that are useful for study-at-hand.

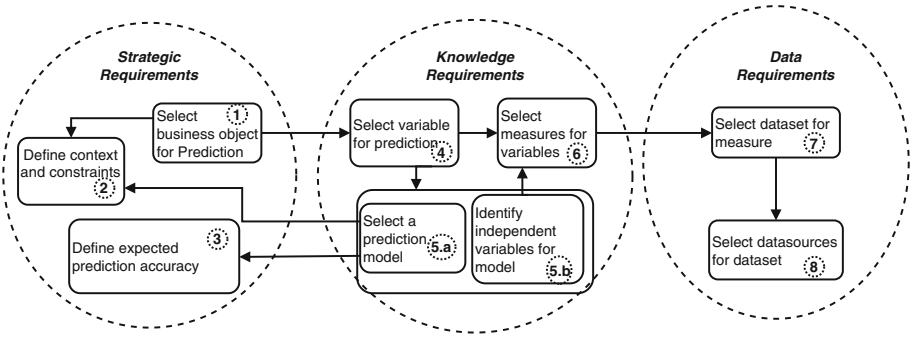

Fig. 4. Requirement template for predictive analytics with Pre-trained models

4.2 Template Instantiation

We used the proposed system to generate requirements definition for this application. We used the proposed SLO and created an instance repository of concepts identified in the literature related to house price prediction. Figure 5 represents one template instance generated based on the SLO, driven by different requirements user specified in the requirements template shown in the Fig. 4. The numbers are used to match the requirements in the template with the concepts defined in instance repository. Furthermore, each concept type according to the SLO is indicated in italic.

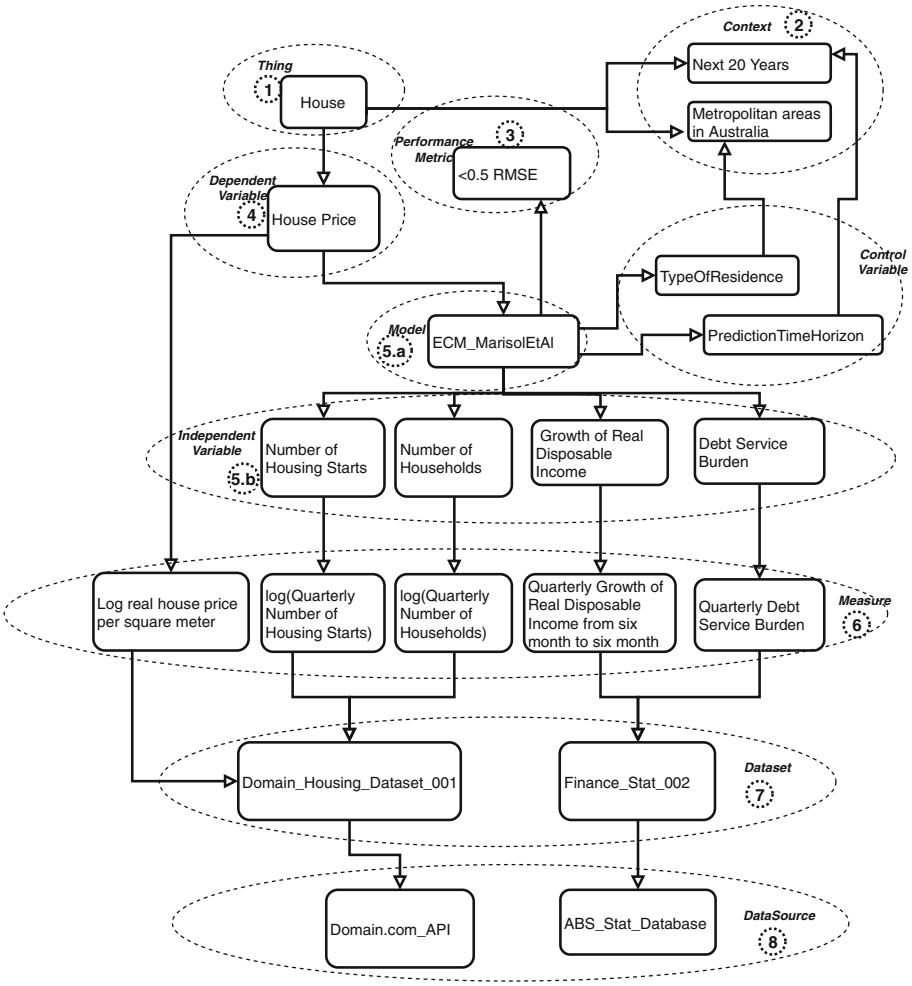

Fig. 5. Requirement template instance - predictive analytics with Pre-trained models

4.3 Prototype Implementation

We are developing a complete prototype of the proposed system. As the first step, the ontologies and instance repository are created on a semantic triple store. The ART and requirements templates are also defined in OWL XML syntax. A semantic query engine is being implemented as a REST API where results are generated when the requirements and parameters are passed.

A web-based front end will be designed to drive the interaction for requirements definition based on the requirements templates. Requirement template manager and dialog manager will be developed on the server side of the web-based application.

5 Conclusion and Future Work

In this paper, we propose a system for defining requirements related to data analytics process via ontologies and a requirements taxonomy. The main advantage of leveraging ontologies is that it provides traceability between different strategic and operational requirements as well as related domain or contexts. Once an organization develops a rich knowledge repository following the proposed system design, it becomes easy and efficient to define new analytics requirements. We express the capabilities of the system by applying it to developing an analytics solution for house price prediction.

By using the proposed method in the house price prediction application we observed that it provided a sound approach to define and link the essential properties of the tacit knowledge of the domain experts from which requirements can be easily generated. It also enabled us to link the known ontologies in the domain with the requirements which enabled us to utilize well established knowledge in the filed. In addition, as requirements generation is automated we can update the requirements specifications as the domain knowledge changes.

Next step of our research is to design an effective interface that enables users to communicate with requirements templates, ontologies and instance repository. We are experimenting with a web-based application that drives dialog-based communication with the user to finalize the requirements definition [4]. Further, we are looking at tools that capture and catalog business models through ontologies and how they can be extended to support our system design.

Potential extension of this work is mapping the requirements taxonomy with the traditional requirements definitions expressed in natural languages, enabling requirements definition for users in more intuitive fashion. We are also planning to integrate the system into our business process model based requirements generation model [1]. As the outcome of this system is a well-defined requirements definition connected to low-level artifacts of the analytics process such as data sets, incorporating it with a semantic service model such as SA-REST this system can be enhanced to support requirements driven service orchestration to realize execution level analytics processes.

Acknowledgments. We are grateful to Capsifi, especially Dr. Terry Roach, for sponsoring the research which led to this paper.

References

1. Aysolmaz, B., Leopold, H., Reijers, H.A., Demirörs, O.: A semi-automated approach for generating natural language requirements documents based on business process models. Inf. Softw. Technol. **93**, 14–29 (2018)
2. Baader, F.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, Cambridge (2003)
3. Bandara, M., Rabhi, F.A.: Semantic modeling for engineering data analytic solutions (2018, Under Review)
4. Bandara, M., Rabhi, F.A., Meymandpour, R.: Semantic model based approach for knowledge intensive processes. In: Stamelos, I., O'Connor, R.V., Rout, T., Dorling, A. (eds.) SPICE 2018. CCIS, vol. 918, pp. 215–229. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00623-5_15
5. Behnaz, A., Bandara, M., Rabhi, F.A., Maurice, P.: A statistical learning ontology for managing analytics knowledge. In: Proceedings of Workshop on Enterprise Applications, Markets and Services in the Finance Industry (2018)
6. Bellatreche, L., Khouri, S., Berkani, N.: Semantic data warehouse design: from ETL to deployment à la Carte. In: Meng, W., Feng, L., Bressan, S., Winiwarter, W., Song, W. (eds.) DASFAA 2013. LNCS, vol. 7826, pp. 64–83. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37450-0_5
7. Brijs, B.: Business Analysis for Business Intelligence. Auerbach Publications, Boca Raton (2016)
8. Colazzo, D., Goasdoué, F., Manolescu, I., Roatiş, A.: RDF analytics: lenses over semantic graphs. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 467–478. ACM (2014)
9. Pan, J.Z., Staab, S., Aßmann, U., Ebert, J., Zhao, Y.: Ontology-Driven Software Development. Springer, Berlin (2012). https://doi.org/10.1007/978-3-642-31226-7
10. Pignotti, E., Edwards, P., Gotts, N., Polhill, G.: Enhancing workflow with a semantic description of scientific intent. Web Semant.: Sci. Serv. Agents World Wide Web **9**(2), 222–244 (2011)
11. Rabhi, F., Bandara, M., Namvar, A., Demirors, O.: Big data analytics has little to do with analytics. In: Beheshti, A., Hashmi, M., Dong, H., Zhang, W.E. (eds.) ASSRI 2015/2017. LNBIP, vol. 234, pp. 3–17. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76587-7_1
12. Siegemund, K., Thomas, E.J., Zhao, Y., Pan, J., Assmann, U.: Towards ontology-driven requirements engineering. In: Workshop Semantic Web Enabled Software Engineering at 10th International Semantic Web Conference (ISWC), Bonn (2011)
13. Taylor, J.: Framing requirements for predictive analytic projects with decision modeling (2015)
14. Wiegers, K., Beatty, J.: Business analytic projects. Software Requirements (2013)