A Model for Assisting Business Users along Analytical Processes

Corentin Follenfant[1,2], David Trastour[1], and Olivier Corby[2]

[1] SAP Research, SAP Labs France SAS
805 avenue du Dr. Maurice Donat, BP 1216, 06254 Mougins Cedex, France
firstname.lastname@sap.com
[2] INRIA Sophia Antipolis Méditerranée,
2004 route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France
firstname.lastname@inria.fr

Abstract. User-centric business intelligence aims at empowering analysts who interact with complex tools, by allowing them to perform accurate data manipulations and analysis without necessarily requiring IT expertise and knowledge of underlying data specifications. Recommender systems contribute to easing their tasks but most of them operate inside walled gardens and cannot assist properly the user throughout his BI workflow. In this paper we introduce a lightweight vocabulary intended to capture fragments of analytical workflows as multidimensional data transformations, within a Semantic Web framework. We utilize this model for calculating content-based recommendations.

Keywords: business intelligence, content-based recommendation, analytical layers, usage semantics

1 Introduction

Traditional Business Intelligence (BI) platforms provide tools that are designed to cover a wide range of operations in a data-driven decision making workflow. The prerequisites steps concern data extraction, cleansing and integration. On top of them come what we call analytical processes: it includes querying, analysis and visual data consumption. These operations often require various technical competencies, for instance SQL expertise and a good understanding of underlying relational models. Since the current businesses landscapes rarely allow users to maintain both technical and analytical profiles, this hinders the decision makers' capacity to leverage the tools at their full potential without requiring extensive assistance from IT departments.

A common approach to tackle this problem is by providing contextual assistance through recommender systems in order to suggest items such as datasets, business entities, queries or visualizations, depending on the current step of the user into his analytical process. Although those systems start to work beyond the legacy $User \times Item$ space and integrate broader contextual information [1],

they can hardly be applied on the whole analytical process as items become heterogeneous and implicit rating functions complex.

In this paper we propose an information model based on Semantic Web technologies, designed to capture semantics of sequential transformations applied on multidimensional data structures. We describe a content-based recommendation use case of this model, where items' granularity vary from business entities to analytical processes aspects, and their utility is computed by arbitrary functions over their usage statistics.

2 Context

BI systems architecture can be split in three layers: first, raw data mainly comes from operational systems where it is stored into heterogeneous databases. Secondly, Extract, Transform, Load (ETL) and integration processes federate those sources into data warehouses. Combined with metadata management components, they expose data through a (hyper)cube model of business entities named after users' familiar terminology: *measures* (factual data, e.g. `Revenue`) that can be driven by *dimensions* (dimensional data, e.g. `Country`, `Year`, `Store`). Thirdly, analytical processes of end user applications such as reporting tools begin by querying the data warehouse layer to retrieve multidimensional data, before applying transformations and visualizations as the user authors his report.

Number of efforts are devoted to making these tools more usable and accessible by involving recommender systems for specific steps of analytical processes. This goes from querying the data warehouse [2, 4], to higher-level workflows such as exploration [6, 3]. Assisting users throughout the analytical process requires a common metamodel to capture multidimensional operations.

3 An Information Model to Capture Usage Semantics

The RDF Data Cube vocabulary introduced by Cyganiak et al.[3] is mainly intended to enable the publication of statistics, and thus provides a metamodel for multidimensional datasets. In order to enable high level description of analytical processes that can be performed within reporting tools, we extend the vocabulary[4] as presented in figure 1. Processes are split into sequences of multidimensional data transformations that

are derived from users' interactions with the report design tool. Each transformation corresponds to a `mda:AnalysisLayer` subclass instance. Like a web service in the OWL-S[5] fashion, it consumes and exposes interfaces of multidimensional data structures described by `qb:DataStructureDefinition` individuals

[3] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

[4] The RDFS classes and properties of our extension use the `mda:` prefix, for MultiDimensional Analysis. The `qb:` prefix refers to Data Cube vocabulary.

[5] http://www.ai.sri.com/daml/services/owl-s/1.2/overview/

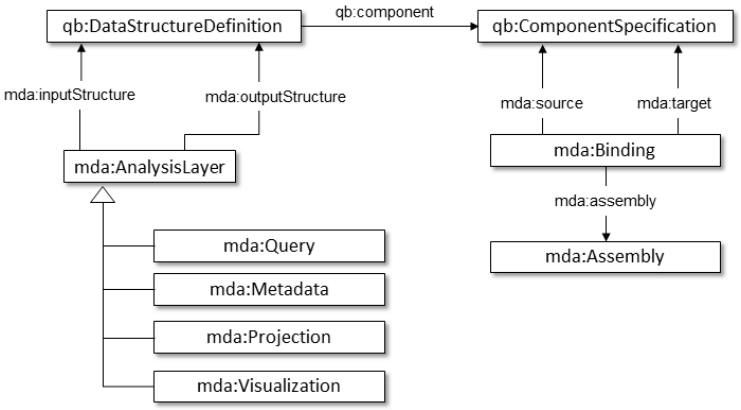

Fig. 1: Multidimensional Analysis extension outline

through `mda:inputStructure` and `mda:outputStructure` properties. Transformations can thus be interchanged and reused for describing different snippets (reports elements) that share layers of the analytical process. These layers are connected together through sets of bindings, assemblies, that plug the multidimensional structures atoms, business entities represented by `qb:ComponentSpecification` individuals.

4 Experiments

Aiming at providing assistance and reuse capabilities to business users who consume reports and have authoring expectations, we leverage our model to compute content-based recommendations. We ran a snippet crawler against a repository of BI reports in order to harvest snippets' underlying analysis sequences and populate an RDF graph with generated triples. The source is an internal repository storing 645 reports used to perform analysis on 101 data warehouse models. A total of 8121 snippets were extracted, all of them being split into up to five layers of transformations over business entities.

Usage statistics measures are extracted from this graph and then used to feed utility functions of a recommender system, for which we identified two granularities of recommendations. First, basic top-k SPARQL queries can suggest business entities that are likely to complete a `ProjectionLayer`, that is adding dimensions or measures to a snippet's axis. As opposed to this horizontal recommendation, the vertical one aims at recommending entire layers in the analytical process in order to assist a user into reusing relevant transformations or visualizations that can be applied on top of a query. To do so, we compute item similiarity measures for `AnalyticalLayer` individuals that are not already connected together through assemblies. The similarity measure strategy is adapted from the Levenshtein distance implemented in the iSPARQL extension [5].

5 Conclusion & Future Work

We introduced an approach to reuse-oriented analytical processes modelling with Semantic Web technologies, which captures the different steps of analysis as multidimensional structures transformations. The first use case concerns BI reporting applications, for which we exemplified our model by triplifying a repository of reports snippets. The graph data resulting from this initial experiment can be queried for basic usage statistics or content similarity measures with simple SPARQL aggregates or iSPARQL statements.

Areas of research that we expect will require further investigation include the formal definition of matching criteria between layers of analytical processes, and its implementation as a specific similarity strategy for analytical layers' RDF resources in iSPARQL; and mechanisms to capture or infer the provenance of the data surfacing into end user visualizations [7]. Finally, we will check the model's genericity by using crawlers for BI applications besides reporting tools, such as dashboarding or exploration ones. This will enable representing analytical processes composed of transformations and data derived from different environments, for instance statistical data published with respect to RDF Data Cube vocabulary and external to the data warehouses.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17, 734–749 (2005)
2. Chatzopoulou, G., Eirinaki, M., Polyzotis, N.: Query Recommendations for Interactive Database Exploration. In: Proceedings of the 21st International Conference on Scientific and Statistical Database Management. pp. 3–18. SSDBM 2009, Springer-Verlag, Berlin, Heidelberg (2009)
3. Jerbi, H., Ravat, F., Teste, O., Zurfluh, G.: Applying Recommendation Technology in OLAP Systems. In: Aalst, W., Mylopoulos, J., Rosemann, M., Shaw, M.J., Szyperski, C., Filipe, J., Cordeiro, J. (eds.) Enterprise Information Systems. Lecture Notes in Business Information Processing, Springer Berlin Heidelberg (2009)
4. Khoussainova, N., Kwon, Y., Balazinska, M., Suciu, D.: SnipSuggest: Context-aware Autocompletion for SQL. Proc. VLDB Endow. 4, 22–33 (October 2010)
5. Kiefer, C., Bernstein, A., Stocker, M.: The Fundamentals of iSPARQL: A Virtual Triple Approach for Similarity-Based Semantic Web Tasks. In: The Semantic Web. Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2007)
6. Marcel, P., Negre, E.: A Survey of Query Recommendation Techniques for Datawarehouse Exploration. In: Proceedings of the 7th Conference on Data Warehousing and On-Line Analysis. EDA '11 (June 2011)
7. Reisser, A., Priebe, T.: Utilizing Semantic Web Technologies for Efficient Data Lineage and Impact Analyses in Data Warehouse Environments. In: Proceedings of the 2009 20th International Workshop on Database and Expert Systems Application. DEXA '09, IEEE Computer Society, Washington, DC, USA (2009)