Rapport sur le manuscrit de thèse d'Ahmad Assaf

Enabling Self-Service Data Provisioning Trough Semantic Enrichment of Data

This thesis addresses self-service provisioning, i.e. intuitive techniques for datasets discovery, acquisition and integration, in a Business Intelligence context. This thesis was carried out in collaboration between SAP and Eurecom within a CIFRE contract. This research work addresses several challenges: (1) dataset maintenance and discovery, by providing a metadata model and a scalable framework for extracting, generating and validating linked datasets profiles, (2) dataset quality control by identifying quality indicators and integrate most of them in the framework, and (3) dataset integration and enrichment by creating a framework for enriching datasets and an API for aggregating social networks. After a general introduction of the motivation and research challenges, the manuscript is organized in two main parts: part I presents the development of the framework for validating and generating datasets profiles and is divided in four chapters, and part II focuses on the integration of external data and is divided in three chapters. Each part contains a state-of-the-art and the contributions. Then, a general conclusion and an outline of perspectives are given. A big picture of the work is missing in the introduction.

Chapter 2 briefly overviews the semantic web and open data. Then, the data profiling task is described. This chapter is very short and could have been integrated into the introduction.

Chapter 3 surveys models and vocabularies used for describing datasets on the web, and concludes by identifying the need for a harmonized dataset metadata model taking the best of existing models. A classification has been defined from these models and mappings from different models are extracted for designing the Harmonized dataset model covering all the information types. A consequent work has been done for examining all these models and vocabularies specifications and documentation. However, it should be useful to elaborate on the possible automation of the definition of this harmonized model. A question is also to know how the model is updated to cope with the evolution of Linked Open Data.

Chapter 4 addresses the automatic validation and generation of datasets profiles. A framework has bee designed for extracting, validating and enriching datasets profiles. Metadata are associated to these datasets in order to be able to efficiently use them. The motivation is clearly stated. An overview of the Roomba system is depicted in this chapter for the following processes: data identification, metadata extraction, instance and resource extraction, profile validation and finally, profile and report generation. Experiments and evaluation are then presented. Few details are given on the search process from the user side. For example, is it possible to search for datasets on a particular topic, including specific attribute types? Experiments done with Roomba are presented, and the evaluation focuses on profiling correctness and completeness. The overall objective of the evaluation should have been more clearly stated for a better comprehension of the results.

Data quality and trust is a major issue and is addressed in chapter 5. A classification of objective data quality measures is proposed, based on an analysis of the state-of-the-art of data quality for open linked data. Each main category of this resulting classification is then detailed and the way to measure each category is given. Tools for measuring quality aspects are surveyed, and an extension of Roomba was designed to fill the gaps identified in existing tools. This chapter is very useful and well presented. And, nor surprisingly, the evaluation performed in this chapter leads to outline the quality defaults of most datasets.

Part II is dedicated to data integration and semantic enrichment in the SAP ecosystem. This part is composed of three chapters: chapter 6 describes the various tools in the SAP Business Intelligence ecosystem, chapter 7 outlines the need for an enterprise knowledge base, and chapter 8 presents a semantic news aggregation system.

Chapter 6 briefly presents the data integration problem. This part should be more elaborated. Then, Business Intelligence, SAP Hana and social web are also very briefly presented. The reader is not able to understand why social web is depicted there. This chapter should either present the big picture of part II, or be integrated in the following chapters.

Chapter 7 focuses on data integration and proposes a tool for semi-automatically combine data from heterogeneous sources, using the DBPedia knowledge base. A reverse engineering is then performed on the Google Knowledge Graph to find the most relevant properties for an entity. A framework for schema matching has been designed; this tool identifies semantically related data and proposes appropriate mappings to the user; then, data is aggregated and can be visualized or exported to BI tools. New similarity algorithms for string values have been implemented for matching unnamed and untyped columns. Experiments have been conducted on a real scenario. Various combinations of semantic matching algorithms were experimented, resulting in higher number of matches. The performance obtained by combining algorithms is better, but this should be confirmed by more experiments on larger datasets. Important properties of entities are extracted using Google Knowledge Graph, and the result is compared to a user study.

Chapter 8 presents a social semantic news aggregator, retrieving and aggregating news from various platforms. Some examples should be added to this chapter. From a technical point of view, does SNARC process social platforms in real-time? How is it integrated with BI dashboards? How is the aggregated information pushed to the user?

Finally, chapter 9 concludes and outlines future perspectives. A summary of the contributions and a general architecture schema are depicted, showing a global vision of the research work. Coming back to the scenario given in introduction also contributes to clarify the benefits and contributions of this work for an enterprise.

This research work addresses the problem of enriching enterprise information with external data, as data value can be augmented by this external knowledge. A consequent work has been done for implementing, experimenting and evaluating. The quality issue is very well addressed. The idea of using a scenario is a good one, but this scenario should be developed in each chapter to show how the processes are enhanced. Despite the architecture given in conclusion, the reader still misses some glue between the different components, and a clear vision of the industrialization in SAP. This work has been validated by a consequent number of publications, including top conferences and a journal. The whole work demonstrates the complexity of integrating various external data sources, and provides many interesting insights for tackling this challenge.

According to the quality of the research work presented in this manuscript, covering a wide range of topics related to Business Intelligence and Semantic Data Integration, I am in favor of the defense of this thesis to obtain the grade of Doctor of Philosophy, specialty in Computer Science and Multimedia from TELECOM ParisTech.

November 13th, 2015

Marie-Aude Aufaure

Professeur, Université Paris-Saclay, CentraleSupélec