ANNUAL PROGRESS REPORT – YEAR 1 (2011/2012)

AN INTERACTION FRAMEWORK FOR BUSINESS INTELLIGENCE

Industrial thesis: CIFRE/SAP, EURECOM and EDITE doctoral school
Author: Ahmad ASSAF
Doctoral advisor: Raphaël Troncy
SAP supervisor: Aline SENART

September 19, 2012

1. Introduction

This report aims to give a brief overview of what has been done for our first year of PhD. We joined SAP Research in May 2012 in order to start a PhD thesis with SAP AG and EURECOM while being registered in the EDITE doctoral school in Paris.
Business intelligence (BI) mainly refers to computer-based techniques used in identifying, extracting, and analyzing business data, such as sales revenue. In BI, business users generally perform four types of processes: Data Selection, Data Manipulation, Data Analysis, and Data Visualization. After a full cycle, the user has the possibility to perform these processes again; we call this the BI cycle. The objective of this research is to provide a framework that enables users to have a better interaction with the different BI processes by tackling the limitations that affect user experience. This will allow users to have better insights and analysis of their data, which will lead to more efficient and accurate decisions. This framework will support the full BI cycle, from data selection, manipulation and transformation to data visualization.

2. First year analysis

For the first few months of this PhD, we have performed a literature survey spanning different fields ranging from Business Intelligence, Data Analysis and Integration to the Semantic Web. Doing so helps in enriching our overall knowledge about these fields to narrow down the research topic in the near future.

3. Publications and Deliverables

We have participated in all the deliverables of the SAP internal project remix. Remix is a self-service BI tool that enables non-technical business users to compose existing BI artifacts with new structured internal and external data sources. The contributing teams were a five people team in Mougins, France and a 4 people team in Dresden, Germany. Remix development was followed the Scrum methodology; which is an iterative and incremental agile software development method for managing software projects and product or application development. Our contributions to the project consisted of designing the data model, developing the whole front-end of the product taking into consideration different user experience problems and adapting the design for mobility.

We have published three papers so far:

1. A. Assaf, E. Louw, A. Senart, C. Follenfant, R. Troncy and D. Trastour. “Improving Schema Matching with Linked Data”. The 1st International Workshop on Open Data (WOD 12), Nantes – France.

Abstract:
With today’s public data sets containing billions of data items, more and more companies are looking to integrate external data with their traditional enterprise data to improve business

intelligence analysis. These distributed data sources however exhibit heterogeneous data formats and terminologies and may contain noisy data. In this paper, we present a novel framework that enables business users to semi-automatically perform data integration on potentially noisy tabular data. This framework offers an extension to Google Refine with novel schema matching algorithms leveraging Freebase rich types. First experiments show that using Linked Data to map cell values with instances and column headers with types improves significantly the quality of the matching results and therefore should lead to more informed decisions.

2. A.Assaf, A.Senart. "Data Quality Principles in the Semantic Web". The International Workshop on Data Quality Management and Semantic Technologies 2012, Palermo – Italy.

Abstract:
The increasing size and availability of web data make data quality a core challenge in many applications. Principles of data quality are recognized as essential to ensure that data fit for their intended use in operations, decision-making, and planning. However, with the rise of the Semantic Web, new data quality issues appear and require deeper consideration. In this paper, we propose to extend the data quality principles to the context of Semantic Web. Based on our extensive industrial experience in data integration, we identify five main classes suited for data quality in Semantic Web. For each class, we list the principles that are involved at all stages of the data management process. Following these principles will provide a sound basis for better decision-making within organizations and will maximize long-term data integration and interoperability.

3. A. Assaf, E. Louw, A. Senart, C. Follenfant, R. Troncy and D. Trastour . RUBIX, A Framework for Improving Data Integration with Linked Data. *WOD '12,* May 25 2012, Nantes, France.

Abstract:
With today's public data sets containing billions of data items, more and more companies are looking to integrate external data with their traditional enterprise data to improve business intelligence analysis. These distributed data sources however exhibit heterogeneous data formats and terminologies and may contain noisy data. In this paper, we present RUBIX, a novel framework that enables business users to semi-automatically perform data integration on potentially noisy tabular data. This framework offers an extension to Google Refine with novel schema matching algorithms leveraging Freebase rich types. First experiments show that using Linked Data to map cell values with instances and column headers with types improves significantly the quality of the matching results and therefore should lead to more informed decisions.

4. Conclusion and Future Work

We have started this thesis in May 2012 on a topic that spans different domains. We have identified the following research questions:

- How visualizations will be connected to the underlying data?
- Will the user be able to interact directly with the visualization, hiding completely the data behind it?
- How can the framework recommend certain visualizations that match the user's data?

During the different BI processes, there is a continuous interaction from the user's end. A set of several research questions rises when trying to form a guidance framework:

- What are the requirements for designing an intuitive usable interactive interface that will support data selection, manipulation and transformation?
- How can we control the scope of presented data and what are the levels of details needed to decompose and aggregate it?
- How can we identify different user interactions with the framework and what kind of interactions is significant to store?
    - How can we define and present an interaction?
    - How to store this interaction?
    - How to show the set of saved interactions back to the user?
- Can erroneous interactions be captured with the visualization and can suggestions be provided to correct them? How can we decide if any interaction is erroneous or will lead to one that is?

We plan until the end of this year to continue exploring the several fields covering our topic, find interesting problems, refine our research questions and implement a prototype to build on for the following years.