remix: A Semantic Mashup Application

Magali Seguran, Aline Senart, and David Trastour

SAP Research, SAP Labs France
805 Avenue du Dr. Maurice Donat
06254 Mougin, France
firstname.lastname@sap.com

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 require helping the user merging data coming from heterogeneous sources.

remix is a Business Intelligence (BI) solution that offers business users a productive environment to easily create highly formatted reports they would ultimately like to see. Via rich visual context-aware interactions, users can quickly combine shared data sets and reuse report parts. This enhanced collaboration combined with the provision of a multi-source semantic layer give users the power to make more effective and informed decisions on virtually any relevant data source or BI resource wherever they are.

Keywords: semantic mashup, semantic web, schema matching, semantic enrichment.

1 Introduction

Whether freely available in the case of "open data", or commercially available in data aggregators or marketplaces, data on the web is growing at fast pace in several domains: government, financial, science, biology. This structured knowledge available in the future e-society will make it feasible for companies to mine huge amount of public data and integrate it in their next-generation enterprise information management systems. Analyzing this new type of data within the context of existing enterprise data should bring them new or more accurate business insights and allow better recognition of sales and market opportunities (*1*).

Unfortunately, current BI applications and solutions cannot easily consume structured data available on the web. These new distributed sources raise tremendous challenges. They have inherently different file formats, access protocols or query languages. They possess their own data model with different ways of representing and storing the data. Data across these sources may be noisy (e.g. duplicate or inconsistent), uncertain or be semantically similar yet different (*2*). Integration and provision of a unified view for these heterogeneous and complex data structures therefore require powerful tools to map and organize the data.

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 312–315, 2012.

2 Proposition

While there are a number of high-quality tools in the BI domain (*3*,*4*), it is very difficult for end users to consume structured data from the web today. By leveraging technology from the linked data community, we propose an innovative self-service BI tool that bridges this gap and allows non-technical business users to easily augment enterprise content with external content. *remix* combines, recommends and presents information from any structured data source from the enterprise and web world. In particular, it offers the following key features:

- **Self-service BI for external data**. *remix* enables mashups of enterprise data and external data. *remix* adds business context to the data and builds semantic links into a unified and consolidated view.
- **Query-free**. Business users do not write queries but simply graphically build what they have in their mind. It's a direct and intuitive way to build reports without requiring the user to understand IT concepts.
- **Guided interaction**. *remix* provides contextualized recommendations to the business users to increase quality when possible, suggest new insights and save them valuable report design time. *remix* helps the business user to find relevant data sources, formulas, visualization, and assists in the data reconciliation process.
- **Collaboration and sharing**. Business users are able to share their data sets and report parts allowing other users to quickly build new reports.

3 Research Challenges

The two key challenges that we encountered in developing *remix* are the provision of the multi-source semantic layer and the recommendation system.

- **Multi-source semantic layer**. In order to interlink data from different sources, we leverage vocabularies and metadata that are readily available on the web. We map cell values with instances, and column headers with types from popular data sets from the Linked Open Data Cloud (e.g. dbpedia). Once raw data has been enriched, schema matching can be performed. Specific algorithms on rich types from vector algebra and statistics have been developed and results show that the use of our multi-source semantic layer greatly improves the matching process (*5*). For example, schemas can be discovered if column headers are not defined and can be improved when they are not named or typed correctly.
- **Recommendation system**. In order to facilitate the reuse of BI artifacts (reports, dashboards, etc.), and easily adapt existing reports to new data, we developed a recommendation system. Based on user profile, historical usage and content similarity (on visualization, query or data elements), the system aggregates data and make recommendations to users at all stages of the data mashup process. The real greatest impediment is to evaluate our results. Until we have a sufficiently large amount of users and BI artifacts in the *remix* repository, the system cannot return interesting advices. To overcome this bootstrapping issue, we focus our efforts on

users' habits and preferences that are well-known and we plan to exploit user traces from other SAP products in the future.

4 Demo and User Benefits

Our demo is based on a use case taken from the healthcare domain. We will run a scenario illustrating how Doctor H., a physician with a speciality of infectious diseases at the Nice Hospital (France) is able to perform easily and quickly data mashup on his medical reports with external data. In the demo, we will show that Doctor H. will discover the source of local suspicious infection with the help of *remix*. Merging different sources of data (external pollution report, hospital medical reports) will lead him to the conclusion that most of his patients suffer from water pollution.

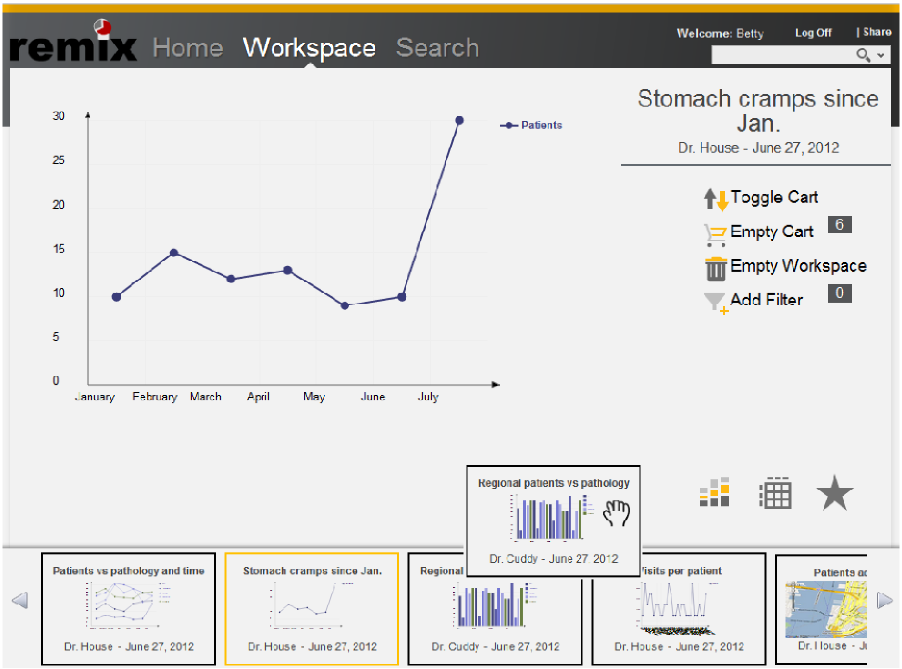
remix
Home
Workspace
Search
Welcome: Betty
Log Off
| Share
Stomach cramps since Jan.
Dr. House - June 27, 2012
Toggle Cart
Empty Cart
6
Empty Workspace
Add Filter
0
Patients
30
25
20
15
10
5
0
January
February
March
April
May
June
July
Regional patients vs pathology
Dr. Cuddy - June 27 2012
Patients vs pathology and time
Dr. House - June 27, 2012
Stomach cramps since Jan.
Dr. House - June 27, 2012
Regional
Dr. Cuddy - June 27, 2012
Visits per patient
Dr. House - June 27, 2012
Patients ac
Dr. House - Ju

Fig. 1. *remix* tool, "your report on my data"

At the beginning of the scenario, Doctor H. builds his report following *remix*'s suggestions and becomes aware of a trend in his patients' visits. He decides to share his report with Doctor C. (his boss) who has access to more data at a regional level. Figure 1 shows how Doctor C. applies her data on Doctor H.'s report, confirming that there is an increasingly large amount of stomach cramps in the last month in the region. Remix suggests automatically how to merge the two reports and pre-selects the columns to match. After some investigation using *remix* (in particular, searching for relevant reports), the physicians perform a second merge between the patients's addresses and a report on the pollution of the river. *remix* shows a clear correlation between the location of patients and the pollution, enabling Doctor H. and Doctor C. to raise a healthcare alert.

This scenario shows that *remix* assists the users in all steps of the process: identification of relevant data sources, recommendation of reusable reports, suggestion in data processing and data alignment. This helps users to produce higher quality business analysis in a shorter time frame, therefore increasing efficiency of their work.

5 Conclusion

We propose to present *remix*, a new self-service BI tool that combines, recommends and presents information from any structured data source from the enterprise and web world. The tool is designed to help business analysts use existing knowledge and generate new knowledge in return. In future work, we plan to improve the recommendation engine and to evaluate mash up on big data.

References

1. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big Data, Analytics and the Path from Insights to Value. MIT Sloan Management Review (2011)
2. Kavitha, C., Sadasivam, Shenoy, N.: Ontology Based Semantic Integration of Heterogeneous Databases. European Journal of Scientific Research 64(1), 115–122 (2011)
3. QlikTech. QlikView: Business Discovery for Everyone, `http://www.qlikview.com/` (accessed 2012)
4. Tableau Software, `http://www.tableausoftware.com/` (accessed 2012)
5. Assaf, A., Louw, E., Senart, A., Follenfant, C., Troncy, R., Trastour, D.: Improving Schema Matching with Linked Data. In: First International Workshop on Open Data, Nantes, France (2012)
6. Peukert, E., Eberius, J.: Rahm Erhard. A Self-Configuring Schema Matching System. In: 28th IEEE International Conference on Data Engineering (2012)