Navigating in a heterogeneous data space

Sirko Schindler *
Institute for Computer Science
University of Jena,
Germany
sirko.schindler
@uni-jena.de

Manfred Hauswirth
DERI
National University of Ireland,
Galway, Ireland
manfred.hauswirth
@deri.org

Birgitta Koenig-Ries
Institute for Computer Science
University of Jena,
Germany
birgitta.koenig-ries
@uni-jena.de

ABSTRACT

Empirical work in many professions depends on the analysis and visualization of huge statistical datasets from various sources. For example, researchers have to validate their results against existing empirical data or journalists have to evaluate financial records for their next article. Frequently a single dataset will not suffice for the specific purpose and a combination of several datasets is needed. These datasets, however, will rarely be homogeneous. So, before attending to the actual task of working with the data, the different formats may have to be converted, the data may have to be normalized and data schema have to be unified or adjusted. This process is time consuming and has to be repeated for each change in the sources or the derived results.
We propose a framework to automate this tenuous tasks as much as possible. The user should be able to focus on the actual task instead.

1. MOTIVATION

More and more data becomes publicly available every day: Sensor networks publish measurements, scientists attach collections of data to their publications and open government initiatives publish previously internal statistics.[1] Not only publicly available data grows exponentially, but also datasets within cooperations or research groups. With similar problems arising in all cases, as soon as somebody wants to use more than one source of data. At the moment various standards and opinions exist on how data should be represented. This does not end at the format used for storage, but extends to the way the respective data is organized. Most of the time only the creator of a dataset will have control over these characteristics. Other users have to deal with what is offered to them; rarely he or she has any direct influence on the representation. This can lead to the situation where the exact same data collection in two sources may look entirely different to an outside user.
Now if someone wants to take advantage of all this offered data and, for example start with creating a simple chart using several sources, the first step is transforming the data to a shared and suitable format, assuming the necessary datasets have already been identified. Besides being tedious work, the lack of an overall accepted standard leads to a number of problems along the following dimensions:

- *Heterogeneous storage*: Excel sheets, RDF-based formats, relational databases, ...
- *Heterogeneous measurement systems*: SI system or some other units; meter or kilometer
- *Heterogeneous measurement intervals*: hourly measurements or weekly
- *Conflicting conventions*: for example, is the location of a sensor specified by its latitude and longitude or by the town it is located in?
- *Heterogeneous schemata*: additional columns or enumerated values in a single column

Having resolved these problems, the actual data manipulations may be applied to make the data usable for the problem at hand. So for example, instead of the provided hourly readings of a sensor, one may want to have a daily average or maximum value. Also only a small part of the datasets may be of use, thus the other parts can be discarded. Finally, after having the data transformed successfully, the user can start looking for a suitable visualization, which could have been the actual task at hand.
All of these steps are rather time consuming: The user has to find appropriate conversion tools, do transformations and finally assemble all of the data together into a new dataset. If the new dataset remains static and does not have to adapt to changes in the original sources, this approach may be sufficient. But as soon as the new dataset must adapt to changes in the sources, e.g., an extension of the covered time period in case of sensor data, or the result must be changed in some way to integrate more or other data, the above steps have to be repeated. This repetition of quite simple tasks should not be necessary. Though it may not be possible to determine the characteristics of a dataset completely automatically due to the lack of adopted standards, the process of describing these datasets most of the time can be a one-time task. The results may than be reused in case that dataset is accessed again. Also maintaining the sequence of applied

*The work presented in this paper was supported (in part) by the Líon-2 project supported by Science Foundation Ireland under Grant No. SFI/02/CE1/I131.
[1]See e.g., `http://www.data.gov/` or `http://data.gov.uk/`.

data manipulations would enable the user to easily change the particular parts or add more elements to it from other sources. Such additions would enable the user to focus on the resulting representation instead of having to deal with all the details all over again.
For example, when creating visualizations in this way, one preserves the link between the visualization and the underlying data. This facilitates for new usages that rely on such a connection: For example, there is the possibility to define semantic concepts by example. I.e., suppose some climate data for a town is represented by a chart. Now a user would be able to highlight certain days and tag them with "beautiful weather." The system could create a semantic concept for that user stating that he or she defines "beautiful weather" by a temperature in a certain range and the other factors displayed in the visualization. This would obsolete the need for the user to specify certain values to create a new concept and shift to just highlighting some ranges in the chart representing the underlying data. This idea is not a new concept, but related to query-by-example in databases or visual and/or declarative programming in the programming language domain.

2. FRAMEWORK

In order to accomplish this, we propose a basic architectural framework. The framework consists of three components: a directory of know datasets, mediators, and a user interface.
(1) Directory of datasets: Most of the time it will not be possible to infer the characteristics of a dataset automatically in the required detail, so we perform a semi-automatic discovery step: The system will try to gather as much metainformation about a dataset as possible, before having a user revise and augment these findings. These metainformation include structural information (e.g., amount of columns and their respective ranges) as well as semantic information about the content of the dataset (e.g., semantic annotations to the columns or information about the provenance of the dataset). Finally the resulting metainformation is stored in the directory, which makes the dataset available throughout the system along with the generated metainformation. For the semantic descriptions, public ontologies are used to provide a common ground for data manipulations and make the resulting datasets compliant to the principles of (open) linked data [1].
(2) Mediator: The function of this component is twofold. The formats of the sources are transformed into a common format and the data will be manipulated in some form, for example, aggregation of time series or filtering. As inputs the mapper takes a list of datasets from the directory and a set of rules for data manipulation. It then gathers the required data, performs the transformations and sends the resulting dataset back to the user interface.
(3) User interface: The user interface enables the user to search for the required sources by concepts and keywords. For example, a search for climate data for Ireland would not only return nationwide datasets, but also datasets of Irish cities or European datasets where the Irish data is only a part of. Furthermore, the user will be provided with the means to filter and alter the dataset, add other datasets and export the result into a visualization or a new dataset. At the time of creation there is no way to know in advance which formats may be used in future. This results in the necessity to design all data access in a highly modular fashion, enabling easy integration of new modules for the kinds of data sources that were previously not accounted for.

3. CURRENT STATUS

So far we have been focusing on describing data sources in such a way, that enough semantic information and characteristics of the data are provided to enable an automated processing afterwards. Furthermore, the generation of this description should involve as few user interaction as necessary. To achieve this, the prototypical directory has been implemented using a module to import data from Eurostat.[2] The Eurostat datasets are well structured and publicly available. In their current form, however, they lack semantic grounding. This has been targeted by several projects like [3]. Despite this work, it may in reality proof impossible to force a common standard upon all data publishers and thus there will always be data sources without semantic descriptions or with conflicting / heterogeneous descriptions. So we use the bare Eurostat data as an example for those datasets to build our detection algorithms upon.
The challenge here lies in finding a generally applicable approach to deduce the meaning (semantics) of the data and then describe its purpose using public ontologies like DbPedia.[3] In this process we can only start from column headers and maybe some entity labels within the columns. This problem is related to the field of tag disambiguation in folksonomies (e.g., [2]), where tags attached to objects shall be augmented with semantic concepts. In our case however, the context, which a tag or description appears in, is even sparser. It remains to be seen, if we can automatically determine good enough suggestions about a dataset's content to be helpful to the user.

4. CONCLUSION

We outlined a framework to enable users to efficiently make use of different kinds of data sources. The user should be able to use and combine data without having to worry about conversions between different formats and schema. As part of this, the semantic description of datasets is essential: To improve the process dataset discovery and to enable for automation of data conversions and combination of data. Future work will tackle this problem of generating such descriptions out of the little context information available with minimal user support.
Furthermore, we argue that the linkage between visualizations and data can yield new ways of creating additional and possibly semantical information.

5. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst*, 5(3):1–22, 2009.

[2] A. Garcia, M. Szomszor, H. Alani, and O. Corcho. Preliminary results in tag disambiguation using DBpedia, Sept. 2009.

[3] W. Halb, Y. Raimond, and M. Hausenblas. Building Linked Data For Both Humans and Machines. In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008)*, Beijing, China, 2008.

[2] `http://ec.europa.eu/eurostat`
[3] `http://dbpedia.org/`