SemLens: Visual Analysis of Semantic Data with Scatter Plots and Semantic Lenses

Philipp Heim
University of Stuttgart, Germany
philipp.heim@vis.uni-stuttgart.de

Steffen Lohmann
Universidad Carlos III de Madrid, Spain
sloehmann@inf.uc3m.es

Davaadorj Tsendorghaa
University of Stuttgart, Germany
dava@visualdataweb.org

Thomas Ertl
University of Stuttgart, Germany
thomas.ertl@vis.uni-stuttgart.de

ABSTRACT
Querying the Semantic Web and analyzing the query results are often complex tasks that can be greatly facilitated by visual interfaces. A major challenge in the design of these interfaces is to provide intuitive and efficient interaction support without limiting too much the analytical degrees of freedom. This paper introduces SemLens, a visual tool that combines scatter plots and semantic lenses to overcome this challenge and to allow for a simple yet powerful analysis of RDF data. The scatter plots provide a global overview on an object collection and support the visual discovery of correlations and patterns in the data. The semantic lenses add dimensions for local analysis of subsets of the objects. A demo accessing DBpedia data is used for illustration.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.5.2 [Information Interfaces and Presentation]: User Interfaces

General Terms
Design, Human Factors, Management.

Keywords
Visual analysis, semantic web, semantic lens, scatter plot, RDF, SPARQL, DBpedia, magic lens, visual interface, visual querying.

1. INTRODUCTION
The Semantic Web has enabled the creation of a large variety of RDF-based datasets that provide extensive information in structured form. A good example of the impressive growth of the Semantic Web is the number of datasets that have been published in the context of W3C’s Linking Open Data (LOD) project [12] in recent years. As of late 2010, the LOD project counted more than 200 interlinked datasets, consisting of roughly 25 billion RDF triples and nearly 400 million RDF links in total [12]. This huge amount of semantic data holds an enormous utilization potential that have only begun to be exploited.

In order to fill the gap between these two categories of approaches, we developed the tool SemLens that enables a simple yet powerful analysis of RDF data. We took two visual concepts that are known for their ease of use and intuitive understanding and applied them to the Semantic Web. The first are scatter plots which provide a simple but highly scalable visualization that greatly supports the visual identification of linear correlations, clusters, patterns, and extreme values. The second are magic lenses which provide an intuitive metaphor for the multi-dimensional exploration of sub-regions of a visualization.

As we will show in this paper, both concepts can be fruitfully combined and applied to the Semantic Web in order to support a highly visual and multi-dimensional information analysis. We start with a summary of the research context and an introduction into the concept of semantic lenses in Section 2. We then describe the tool SemLens and illustrate its implementation with an example from the DBpedia dataset [2] in Section 3.

2. RELATED WORK
Most of the existing approaches to visually access data from the Semantic Web are based on RDF query languages like SPARQL which enable powerful and semantically unique queries against RDF-based datasets. Since the direct formulation of semantic queries requires the query language to be learned in advance,
visual interfaces usually provide graphical representations and abstractions that ease query formulation and interpretation of the results. As mentioned in the introduction, existing approaches can be roughly categorized according to their degree of simplification in complex and simple ones.

2.1 Complex Approaches
While both simple and complex approaches try to facilitate the formulation of semantic queries, the complex approaches also try to preserve the expressive power of the Semantic Web as far as possible. They provide special notations, form-based wizards, and/or graph representations that offer visual support but simultaneously ensure that the full complexity of the underlying query language is available for the users at any time. On the one hand, this allows powerful queries to be formulated without knowing the exact syntax of the query language, but, on the other hand, still requires extensive knowledge of the language’s logical structure and functional behavior.

An example for a complex approach is SPARQLViz [3] which offers a form-based wizard that guides the users through the query construction process. Further examples are NITELIGHT [8] and OpenLink’s isPARQL [7] that provide graphical depictions of SPARQL constructs to reduce the learning curve. All three approaches use graph visualizations to facilitate the query formulation and/or the presentation of the query results. However, though graph visualizations are well suited to display complex relationships, they also tend to produce crossing edges and overlapping nodes in case of large and highly interrelated datasets [13]. This can impede understanding and hinder both formulation of queries and interpretation of results.

2.2 Simple Approaches
To ease interaction and to also enable users without Semantic Web knowledge to create semantic queries and understand the query results, most simple approaches implement the concept of faceted search [4]. Faceted search allows the user to explore a data collection by a number of discrete attributes – the so-called facets. Each facet represents a specific perspective on the data that is usually mutually exclusive.

Applied to the Semantic Web, facets are most naturally created from the properties of an object collection. These objects typically result from some presetting, i.e. initial SPARQL queries and/or interactions that restrict the search space to only those objects that are of interest to the user (e.g. objects of a certain class). Faceted search is then used to iteratively refine the queries and filter the object collection until it meets the user’s information need.

However, one problem of faceted search interfaces is to control how facets are connected with each other. Many different combinations are possible: For instance, selections within a facet, but also selections between facets, can be connected via both AND and OR operators. Finding an intuitive representation and control of these and further logical operators (e.g. NOT operators) for a faceted search interface is difficult. Therefore, most approaches use default operators that are hidden from the users.

Examples of faceted browsers for the Semantic Web are mSpace [9], facet [5], and Parallax [6]. All these tools use the mentioned default configuration for the connection of facets. They furthermore use lists and tables for the presentation of facets and query results, which is good and common practice but limits the ability to show multi-dimensional dependencies and relationships in the results.

Better suited for the clear presentation of linear dependencies, correlations, and patterns are two-dimensional visualizations such as matrices, maps, or scatter plots. However, the analytical dimensions that can be represented in these visualizations are also limited, which complicates the consideration of many factors when analyzing semantic data. For these reasons, we decided to combine scatter plots with magic lenses in our approach to enable the addition of further dimensions to the visual analysis of semantic data and thereby bridge the gap between complex and simple approaches.

2.3 Magic Lenses and Semantic Lenses
The interaction metaphor of magic lenses was first proposed by Bier at al. as one kind of “see-through interface” [1]. In their original definition, magic lenses “combine an arbitrarily-shaped region with an operator that changes the view of objects viewed through that region” [11]. Similar to optical lenses in the real world, magic lenses can be moved over the user interface “to help the user understand various types of information” [11]. Next to their intuitive metaphor and their advantage to show a modified view in the context of the original view, magic lenses can in principle be arbitrarily combined, allowing to add any number of further dimensions to the visual analysis of a dataset.

We developed an application of the magic lens idea for semantic data that we call “semantic lens”. A semantic lens provides a binary filter for the visualization of an object collection that is based on one of the objects’ properties. Depending on the datatype of that property, the filter can be differently configured. If the datatype is, for instance, a number or date, a threshold can be set by the user which filters all objects whose property value is either below or above this threshold, depending on the filter criterion. Likewise, string matching rules can be defined for semantic lenses that are based on properties with a string datatype.

3. SEMLENS
The tool SemLens [10] implements our approach of combining scatter plots and semantic lenses to visually analyze semantic data. It is developed in the open source framework Adobe Flex and works in Web browsers with installed Flash player. As all data access is based on SPARQL, SemLens can in principle be applied to any dataset that provides a SPARQL endpoint. Some parameters can be configured to optimize the visual analysis. For instance, if certain properties are of no use for the analysis, it can be defined that they are ignored by SemLens.

The current SemLens version has been pre-configured for the DBpedia dataset [2], i.e. it accesses DBpedia’s SPARQL endpoint and ignores some properties without value for visual analysis with SemLens (e.g. dhp:thumbnail, dhp:abstract, etc.). Since most DBpedia data has been automatically extracted from Wikipedia and other sources, it cannot be expected that all information is complete or correct. This is, however, not important to illustrate our approach and to give an idea of its potential.

3.1 Global Analysis via Scatter Plots
This subsection explains the presetting of the search space and the creation of the scatter plots, while Section 3.2 details the local analysis via semantic lenses.
3.1 Presetting the Search Space
As in many faceted search approaches, an initial presetting is conducted in SemLens to reduce the search space to only those objects that are of potential interest to the user (cp. Section 2.2). For this purpose, the user enters one or several search terms and gets a list of classes back that have these terms in their labels. The list is ordered by the number of objects the classes contain and thus gives the user already an idea of the size of the search space. The user can either select the most appropriate class or repeat the search with other terms if the retrieved classes do not satisfy his needs.

3.1.2 Creating the Scatter Plot
Once the presetting is completed and the analysis started, a new tab opens showing the object collection resulting from the presetting in a scatter plot. In the screenshot of Figure 1, the user has selected the class “European countries” in the presetting. The scatter plot shows all countries that have the two properties which are set as horizontal and vertical axis (A1, A2). Initially, the two properties that are shared by most objects are used as axes, in order to visualize as large a number of objects of the class as possible at the beginning. The user can change this initial configuration by selecting other properties from the list to the right of the scatter plot (B). In Figure 1, the user has chosen the two properties “longitude” and “HDI rank” (HDI = Human Development Index) as scatter plot axes.

3.1.3 Alternative Table View
As indicated by the numbers in the upper-right corner of the scatter plot (C), some European countries are not shown in Figure 1 due to missing values for the selected properties. Therefore, SemLens provides an alternative table view listing all objects and their values for user-selected properties (D).

3.1.4 Identifying Correlations
The scatter plot provides an overview of the distribution of the object collection and supports the discovery of correlations, clusters, and patterns in the data. The example in Figure 1, for instance, indicates a linear correlation between the HDI rank of a European country and its longitude. In particular two data points catch the user’s attention, as they differ from the correlation: The first represents Portugal (E1) and the second Cyprus (E2). The user continues the visual analysis with the semantic lenses to investigate why these two countries are not in the correlation.

3.2 Local Analysis via Semantic Lenses
The semantic lenses support the local analysis by adding further analytical dimensions to certain regions of the scatter plot. This subsection explains how they are created, used, and combined.

3.2.1 Creating Semantic Lenses
Semantic lenses can be created for all properties of the object collection, i.e. for every entry in the property list (Figure 1, B). Clicking the lens symbol next to a property label opens the “lens details” dialog to create or edit the filter rule of the semantic lens. For numbers and dates, the user can set a threshold and state if values have to be “greater or equal” or “less than” this threshold to get filtered. For strings, a substring search is performed based on one of the four matching criteria “is”, “contains”, “begins with”, and “ends with”.

Fig. 1 shows three lenses: Lens L1 defines a filter for the property “GDP PPP per capita” (GDP = Gross Domestic Product), lens L2 a filter for the property “currency code”, and lens L3 a filter for the property “Gini”. The latter refers to the Gini coefficient which measures the statistical dispersion of countries, commonly in terms of income and wealth.

Figure 1. Screenshot of SemLens accessing DBpedia data about European countries.
The user has defined the filter rule for L1 to be “greater or equal” 25,000, which filters all countries with a GDP PPP per capita below 25,000 U.S. dollar. For L2, he has selected the matching criterion “contains” and has entered “EUR” in the text field. This filters all countries with currencies other than euro. Finally, he defined L3 to filter all countries with a Gini value “less than” 30 (which is the Gini coefficient multiplied by 100 in DBpedia).

In addition, a Boolean operator needs to be set for each lens that specifies how the filter rule of the lens is combined with the filter rules of other lenses. In the example, the user has set the Boolean operators of L1 and L3 to AND and the operator of L2 to OR.

3.2.2 Exploring with Semantic Lenses

The semantic lenses can be placed anywhere on the scatter plot and their rectangular shape can be freely resized. If the semantic lens is moved over a certain region of the scatter plot, the data points in that region change their color: Those that meet the filter criterion of the lens are shown in green and those that do not meet it in red. Data points of objects that do not have the lens property change to grey. All other data points, i.e. those outside the lens and other lenses, remain blue.

The red color of the data point of Portugal (E1) after positioning of L1 in Figure 1 suggests that its relatively low HDI rank has to do with its also relatively low GDP per capita. The green color of Cyprus (E2) after positioning of L2 indicates a positive correlation between the euro currency and HDI rank of a country.

The filters of semantic lenses can also be combined. The green color of Montenegro (F), for instance, is caused by the OR operator of L2 that requires only one filter criterion to be fulfilled to color data points green; so either the one defined by L2 or the one defined by the AND-connection of L1 and L3. This example demonstrates that not only the filter criterion, the region covered by the lens, and the logical operator are important for the combination of lenses, but also the order in which the lenses are placed above one another.

3.2.3 Lens Stack

SemLens provides a list with the lens stack in the lower right corner of the user interface (Figure 1, G). It displays the order of the lenses which can be changed via drag-and-drop. It also allows editing and deleting of lenses and shows the number of objects that are covered by each lens in the scatter plot. The lens stack can also be visually indicated (though not changed) within the scatter plot to avoid the indirection via the list.

The interplay of filter criterion, logical operator, order, and position of semantic lenses is illustrated in Figure 2. Each lens is represented by its filter criterion and Boolean operator (e.g. L1 has filter criterion F1 and operator AND). Furthermore, each lens lies on a distinct user interface layer above the scatter plot and is logically connected with the underlying lenses according to its operator. This connection is performed whenever the regions of lenses overlap each other. The four Boolean expressions at the bottom of Figure 4 indicate how the lenses are connected. The data points under the region of lens L3, for instance, are green if they meet the following filter expression: F3 AND F1 OR F2.

4. SUMMARY

We have presented SemLens, a visual tool that aims to bridge the gap between powerful but complex and simple but restrictive approaches to visually query and analyze semantic data. It combines scatter plots and magic lenses and applies them to the Semantic Web to enable a simple yet powerful data analysis.

Since all data access is based on SPARQL, SemLens can in principle be used with any dataset that provides a SPARQL endpoint. This promises high value against the backdrop of a growing number of datasets available in the Semantic Web.

5. REFERENCES