Visualization of Large Datasets using Semantic Web Technologies

Suvodeep Mazumdar,

Department of Information Studies,
University of Sheffield
Regent Court - 211 Portobello Street, S1 4DP, Sheffield, UK
s.mazumdar@sheffield.ac.uk

Abstract. Visualization technologies provide means to comprehend, understand and explore data. Observing patterns and anomalies via visualization tools help users to understand issues and take informed decisions. Semantic web technologies used to represent different data types, conforming to particular standards can be exploited to provide meaningful and intuitive visualizations. In this paper, we propose how we intend to provide intuitive and interactive visualizations for large datasets, formalized by multiple ontologies.

Keywords: Information Visualization, semantic web, dynamic queries

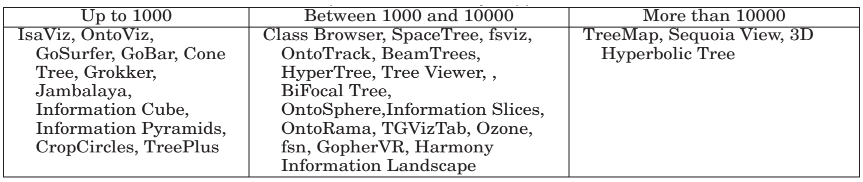
1 The Research Problem

This research looks at highly complex domains such as aerospace engineering. A jet engine's life cycle can last up to 50 years, requiring regular maintenance, overhauls, tests and services. Each of these activities involves documentation in the form of text reports, numeric data, images, in-flight data, CAD drawings etc. The volume of this information can easily exceed several terabytes and some structuring is needed for this extra large heterogeneous information set to be usable. Information extraction and semantic web technologies can provide a standardized and structured representation of the multimedia information. An overarching domain ontology is essential to provide an overall view of the entire domain. In order to gain homogeneity, the overarching ontology will be effective, but doing so would be at the cost of losing details embedded in the document. Hence, each document type can be formalized by its own representative ontology, thereby providing more detailed information respect to the global (overarching) representation. Therefore different ontologies provide different lenses to look at the same document type. It is therefore possible to explore the data at different levels of granularity: a coarse view provided by the domain ontology, and a fine-grain view that makes use of the document ontology. How these two different levels are combined in an effective user interface and how can the users effectively manipulate and explore them is our main research question.

2 Related work and motivation

Several tools for data visualization and exploration have been proposed. For example, Semaplorer [4] visualize people, tags, photos etc on a geographical map; GapMinder[1] provides an exploratory tool for visualizing statistical trends in data over time; ManyEyes[2] allows users to upload their own data and create visualizations. However, most of these visualization tools do not address the main questions of this research, generality and scalability for an effective user interaction. For example current visualization techniques cannot handle very large volumes of data. A.Katifori et al. [2] looked at tools for visualizing ontologies, the available visualization methods and the number of nodes they intend to support (Table1). They found very few visualization methods capable of handling more than 10,000 nodes. GRIDL[3] provides an approach that is scalable by hierarchically presenting each axes and for each axis element, a statistical display (bar chart) is presented; GreenMax [6] provides tree visualization for a million nodes on a representative smaller network of much fewer clusters.

Table 1. Visualization methods according to the number of nodes they intend to support [2].



Up to 1000	Between 1000 and 10000	More than 10000
IsaViz, OntoViz, GoSurfer, GoBar, Cone Tree, Grokker, Jambalaya, Information Cube, Information Pyramids, CropCircles, TreePlus	Class Browser, SpaceTree, fsviz, OntoTrack, BeamTrees, HyperTree, Tree Viewer, , BiFocal Tree, OntoSphere,Information Slices, OntoRama, TGVizTab, Ozone, fsn, GopherVR, Harmony Information Landscape	TreeMap, Sequoia View, 3D Hyperbolic Tree

[7] and [8] present a faceted searching and visualization interface for heterogeneous data by mapping them to known vocabularies extracted from the web and visualizing the data using pre-defined interface widgets. The presence of interactive multiple visualizations is desirable since it helps in effectively exploring the underlying data. One such example is Exhibit, part of the SIMILE Project[4], that allows swapping between different perspectives such as timeline or maps. [3] proposes a more advanced approach as the multiple visualizations are available and updated simultaneously. These visualizations, however, do not fuse different document sets formalized by different ontologies, the first goal of the research. Similarly, the work done by the information visualization community has been mainly limited to homogeneous data. To overcome these limitations, this research combines Semantic Web technology, used to aggregate and structure dispersed and heterogeneous data, with findings from the information visualization community, to provide large-scale, intuitive visualizations and manipulation of heterogeneous data. Indeed, despite evidence that highly dynamic interaction tools effectively support

[1] GapMinder, http://www.gapminder.org/

[2] ManyEyes, http://manyeyes.alphaworks.ibm.com/manyeyes/

[3] Graphical Interface for Digital Libraries, www.cs.umd.edu/hcil/west-legal/gridl/

[4] The SIMILE Project: http://simile.mit.edu; Exhibit: http://simile.mit.edu/wiki/Exhibit

users in data exploration, very little has been done in the area of Semantic Data visualization. This research builds upon our previous work [3] that implements the concept of dynamic queries [1] to provide highly interactive manipulation of multiple visualization, namely tables, timeline, geographical and topological plots.

3 Proposed Approach

We aim to engage the user communities at Rolls Royce actively during the research period. Our approach is to follow the process of iterative user-centered design. Since there are different types of users from different areas of Rolls Royce aerospace engineering domain (design, manufacturing and service), the target system must be able to generate visualizations that are equally interactive, intuitive and informative for all. We intend to conduct personal interviews of users to understand their daily jobs and the kinds of visualizations they are used to. We will then present the users with use case scenarios supported by low-fidelity mockups and sketches of the system that we perceive will benefit the users. We will be following this process iteratively to gain a sound understanding of the user's requirements and expectations.

We will also be studying the different ontologies and their inter-dependencies that have been developed for different sets of data currently in use at Rolls Royce. This will help generate a taxonomy that will relate the data types of the concepts to their corresponding visualizations and interactions they can support. We will be using the results from the participatory design sessions to decide the best interactions for different visualizations, so that the user can seamlessly explore the data in different hierarchical layers.

The usability of the semantic data visualization tools would be core. Applying filters to millions of documents generates very large retrieved sets with thousands of results, too much information for the user to process. Past proposals to mitigate this problem include: increase display area by using 3D plots instead of 2D, cluster or hide nodes or utilizing every pixel in the visualization space [5]. Our approach is radically different and uses classification, clustering and overlapping of data to provide contextual layered visualizations, where each layer contains information only relevant to that layer. For example consider a pie chart generated on the basis of the domain ontology and intended to provide a generic overview of the distribution of the data respect to a specific concept; if the user clicks on a pie chart section which has more detailed information formalized by another supported ontology, then further details corresponding to the specific ontology will be displayed providing a semantic zoom.

4 Methodology

Adopting the User-Centered approach discussed above, a core part of the research is understanding how Rolls Royce engineers conduct their daily work and what tools will be useful during data analysis. The starting point will be observations conducted at Rolls-Royce premises aiming at identifying current practices of data display and analysis. By collecting examples of artifacts currently in use we aim at finding inspiration for a design that will be naturally usable because already familiar. We have already started a series of participatory design sessions with several potential users from different areas of Rolls Royce aerospace engineering domain (design, manufacturing and service). In these sessions we are discussing mock-ups of the visualizations and related interactions so as to actively involve the user community in selecting the - potentially optimal - solution(s). This requirements gathering is paired with the system architecture design to be completed in first year of research.

A series of exhaustive tests on the query response time, loading times, efficiency etc. of the various triple stores will be conducted to select the most efficient system architecture. Once a back-end system is determined, we will be performing tests on loading query results 'on-the-fly'. Tests conducted in X-Media show that there is a significant waiting time for the visualizations to be initialized. This is the base line against which we will work to improve display efficiency, a core issue in user interaction. The software coding phase would be throughout the second year of the research, when we will also be preparing evaluation and trial materials based on the use case scenarios being developed in year one.

The evaluation of the solution will be carried out with the Rolls Royce engineers at their premises during the first two months of the third year. We will follow the methodology we have used previously in [3]: participants will be requested to carry out specific tasks designed in partnership with Rolls Royce experts; the interaction will be logged and the screen activity recorded; participants will then be requested to fill in a questionnaire and answer a few targeted questions in an interview. Results from this user evaluation will be used to re-design and modify the application where needed, following which we would be conducting a long-term user trial. The remainder of the third year would be dedicated to thesis writing and providing bug fixes and minor enhancements.

5 Conclusions and Future Work

The work already done in X-Media shows the importance and effectiveness of multiple visualizations in a large complex organization. The ability of a user to visualize the same data in different dimensions, query them and identify patterns and areas of interest is useful in providing or identifying possible solutions. The findings from the X-Media project has been a good stepping stone for the research we intend to conduct over the next few years.

The research, although organized around the case of aerospace engineering, is expected to be generic and applicable to different domains that share similar characteristics and problems. Specifically, we will test our result with the data from GrassPortal[8], an online resource for accessing data related to grass species, global

[8] GrassPortal, http://www.grassportal.org

environmental data, evolutionary relationships among grasses etc. to test the portability of the approach adopted. This will be a good way to measure how successfully the semantic visualization technology can be ported to other domains represented by their respective domain ontologies.

Acknowledgments. This research is supported by SAMULET, a Rolls Royce and DTI funded project for knowledge management in aerospace manufacturing domain.

References

1. Ahlberg, C., Williamson, C., Shneiderman, B. Dynamic Queries for Information Exploration: An Implementation and Evaluation. CHI'92, 619-626 (1992)
2. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E. 2007. Ontology visualization methods—a survey. ACM Comput. Surv. 39, 4 (Nov. 2007), 10. DOI= http://doi.acm.org/10.1145/1287620.1287621
3. Petrelli, D., Mazumdar, S., Dadzie, A.-S., Ciravegna, F.: Multi Visualization and Dynamic Query for Effective Exploration of Semantic Data. In Proceedings of the 8th International Semantic Web Conference, pp. 505-520. Springer (2009)
4. Schenk, S., Saathoff, C., Staab, S., Scherp, A. 2009. SemaPlorer-Interactive semantic exploration of data and media based on a federated cloud infrastructure. Web Semant. 7, 4 (Dec. 2009), 298-304. DOI= http://dx.doi.org/10.1016/j.websem.2009.09.006
5. Van Ham, F., Van Wijk, J. J. 2002. Beamtrees: Compact visualization of large hierarchies. In Proceedings of the IEEE Conference on Information Visualization. IEEE CS Press, 93–100
6. Wong, P. C., Foote, H., Mackey, P., Chin Jr., G., Sofia, H., and Thomas, J., A Dynamic Multiscale Magnifying Tool for Exploring Large Sparse Graphs, Information Visualization 7, 2
7. M. Hildebrand, J. van Ossenbruggen, L. Hardman, and G. Jacobs. Supporting subject matter annotation using heterogeneous thesauri: A user study in web data reuse. International Journal of Human-Computer Studies, 67(10):887–902, 10 2009.
8. M. Hildebrand and J. van Ossenbruggen. Configuring semantic web interfaces by data mapping. In S. Handschuh, T. Heath, and V. Thai, editors, Visual Interfaces to the Social and the Semantic Web (VISSW 2009), volume 443, February 2009.