Realizing the Hidden – Interactive Visualization and Analysis of Large Volumes of Structured Data

Olaf Noppens
Institute of Artificial Intelligence
Ulm University, Germany
olaf.noppens@uni-ulm.de

Thorsten Liebig
Institute of Artificial Intelligence
Ulm University, Germany
thorsten.liebig@uni-ulm.de

ABSTRACT

An emerging trend in Web computing aims at collecting and integrating distributed data. For instance, various communities recently have build large repositories of structured and interlinked data sets from different Web sources. However, up to date there is virtually no support in navigating, visualising or even analysing structured date sets of this size appropriately. This paper describes novel rendering techniques enabling a new level of visual analytics combined with interactive exploration principles. The underlying visualisation rationale is driven by the principle of providing detail information with respect to qualitative as well as quantitative aspects on user demand while offering an overview at any time. By means of our prototypical implementation and two real-world data sets we show how to answer several data specific tasks by interactive visual exploration.

1. MOTIVATION

The trend of collecting and integrating distributed data into one large repository is gaining more and more momentum. As an example, community efforts such as DBpedia [2] or ReSIST [1] have recently extracted large volumes of structured data from the Web (Wikipedia, US Census Data, DBLP, Citeseer, ACM, etc.). Those repositories are extreme in the sense that they are extraordinary in size and dominated by data sets incorporating only a small and typically lightweight schema.To the best of our knowledge there is no support in navigating, visualising or even analysing large volumes of data in an interactive way appropriately.

As an example, consider a social-network describing members of research communities defined by concepts such as `Project`, `Person`, `Publication`, `Institution` as well as relationships such as `has-Author`, `has-Project-Member`, `has-Research-Interest` etc. A researcher, for instance, can be a working person, an affiliated person, a student or a PhD student (or even an arbitrary combination of these characteristics). An employee can be characterised by his title, his affiliation(s), his degree(s), his project membership(s) in the past and present, his research interests etc. Keeping this in mind, we have to deal with a broad network of people and different kinds of relationships. In this paper we present our approach of combining techniques from visual analytics and interactive exploration of large volumes of heavily interrelated data sets in order to answer data specific tasks. The following describes various selection, exploration and analysis techniques which have been implemented and integrated into our OntoTrack [7] framework. This is done on the basis of two data sets introduced in the next section.

AVI '08, 28-30 May , 2008, Napoli, Italy.

2. DATA SETS

For the rest of the paper, we have chosen two real-world data sets from different domains in order to show how data can easily understood with help of our approach. The first one has been extracted from the MONDIAL Database and the second from the ReSIST Network of Excellence. Both data sets consist of more than hundred thousands entities.

The MONDIAL Database

The Mondial Database[1] (MONDIAL) is a collection of geographic information compiled from different Web data sources such as the World Factbook, Global Statistics and the Terra database [8]. The core of a MONDIAL record consists of data about countries, cities as well as deserts, rivers, or ethnic groups mainly collected from the World Factbook. In addition the collection includes statistical data about populations, area, or length. Entities are typed in a lightweight manner with respect to common geographical concepts such as countries, rivers etc. In addition, various relationships relate entities among each other: for instance, the relationship `has-City` relates countries to cities, `flows-through-country` tells us which countries a river flows through.

ReSIST Network – Resilience Knowledge Base

The Resilience Knowledge Base (RKB) has been created during the first year of the European Network of Excellence in Resilience Computing[2]. The RKB aims at supporting researchers in accessing knowledge on resilience concepts, methods, tools, and the community itself. For that purpose, resilience data has been captured from each partner's information resources such as research interest, details and courseware. This data has been complemented by external sources captured from research information services CORDIS and NSF. Moreover, meta-data about publications

[1] http://www.dbis.informatik.uni-goettingen.de/Mondial/
[2] http://www.resist-noe.org/

and the RISKS index of "Computer-related Risks to the Public" has been gathered from the Citeseer and ACM repositories forming a social-network of researchers and publications. The data is held in a RDFS triple store and accessible via a SPARQL interface.[3] The system incorporates a consistent reference service which maps different URIs from various sources into one reference [5].

3. VISUAL ANALYSIS THROUGH INTERACTIVE EXPLORATION

One lesson learnt from the visual analysis of large data sets in general is that it is not advisable to arbitrarily visualise both all dependencies and all particulars at any time [6]. Therefore, our approach follows the *Visual Information-Seeking Mantra* of "Overview first, zoom and filter, then details-on-demand" [9] by providing detail information only on user demand while offering an overview at the same time.

3.1 Abstraction and Clustering

Following the Information-Seeking Mantra and similar studies, all the connections and relationships between entities can not be visualized and understood at once. We believe that, from the user's point of view, entities with similar characteristics should build obvious or "natural" clusters. For instance, in the MONDIAL domain all European capitals and all countries to which these capitals belong to should automatically be pooled within a cluster as shown in Figure 1. Here, entities are visualised as small filled circles within clusters. Relationships are represented by clubs originating from the set of entities which are considered as the relationship's subject to its objects. However, not only the union of all entities in a cluster can form the origin of a club but also single entities as one can see in Figure 2.

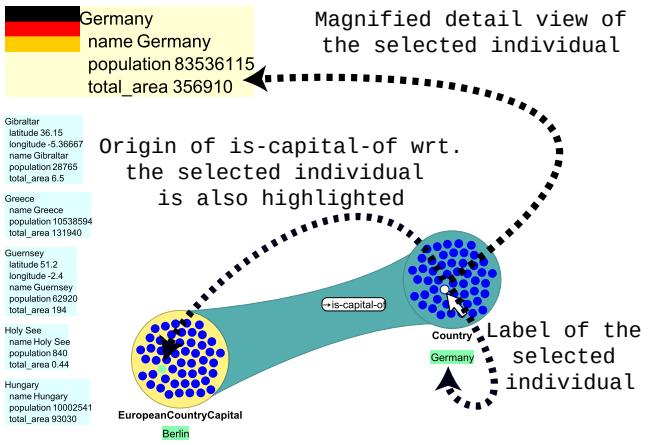

Figure 1: Clustering and club visualisation.

Abstraction also means that entities in a cluster are only drawn if their number is below a user-definable limit. Moreover, the diameter of clusters showing no entities explicitly approximates the number of entities and allows to easily compare the number of entities by the cluster's rendering size. In addition, the number of entities are drawn within a cluster. Additional detail information for each entity such as an image is provided in an optional list as shown on the left hand side of Figure 1. When hovering over an entity with the mouse pointer the list of detail view entries will be scrolled to the corresponding entry and it will be magnified.

To easily grasp all related entities to a focused source entity they are highlighted in all visible clubs when hovering over the source circle. In addition, the labels of these entities are rendered at the bottom of the cluster. For instance, in Figure 1 the mouse pointer is hovering over "Germany". As a result its label is rendered and because "Berlin" is the only origin of the `is-capital-of` relationship the corresponding graphical representation is also highlighted and its label also rendered at the bottom of the `EuropeanCountryCapital` cluster (left hand side of the club).

3.2 Interactive Exploration

When exploring heavily interconnected data sets it is not advisable to show all entities and all their relationships at once. In order to prevent the user in being overwhelmed with currently non-relevant information pieces our user-directed interactive exploration strategy allows for focusing on relevant parts of a data set, or fractions thereof which promise to unveil deeper insights. Initially, one can either start with an user selected entity (e. g. as the result of a query) or with entities showing the same characteristics such as all European capitals in case of the MONDIAL domain. This will result either in showing the graphical representation of that entity or in the case of a set of entities, a slice containing all these entities. As the visualisation and analysis component is integrated into our OntoTrack framework the latter task is carried out by dragging a concept from the schema representation pane on the data analysis pane.

After clicking on the graphical representation of an entity or a cluster a graphical radial preview menu of related entities grouped by their connecting relationships is displayed in an overlay manner. Standard interaction techniques such as mouse-over highlighting and mouse-over zooming as well as intermediate displayed detail information support the user in easily selecting the next club to expand: one or more related clusters are expandable by single mouse-click interaction. For instance, the club shown in Figure 2 represents all project members of "Resilience for Survivability in IST" in the cluster of the right hand side. Here, the preview displays 5 relationships such as `has-author` or `has-affiliation` and one can easily grasp that the entity "Thorsten" for which the menu has been activated is assigned to two affiliations. In order to allow a more flexible exploration of the data set, the exploration direction is not limited to the defined direction of the relationship but also allows to be inversely expanded. The displayed club of Figure 2 represents the relationship `has-project-member` and the preview menu of the selected person also contains the same relationship but in inverse direction allowing a bidirectional exploration of the data set. The direction is denoted by an arrow next to the relationship's label.

Each cluster as well as each (visible) entity can serve as a follow-up point for further expansions. This also allows to branch the expansion by selecting other relationships or de-expand clusters. For instance, in Figure 3 all publications as well as all project memberships (via the inverse expansion of `has-author` resp. `has-project-member` of the members of the Institute of Artificial Intelligence at Ulm University are visible.

To understand how single entities are related to entities in

[3] http://www.rkbexplorer.com/

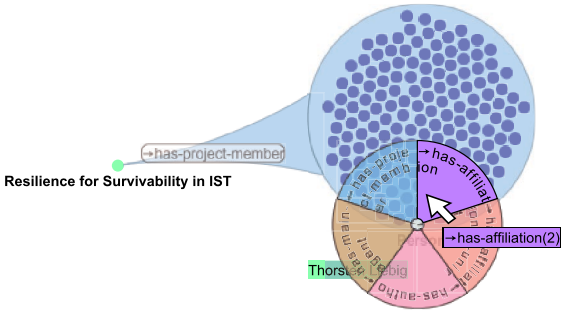

Figure 2: Previews for connected entities grouped by relationships.

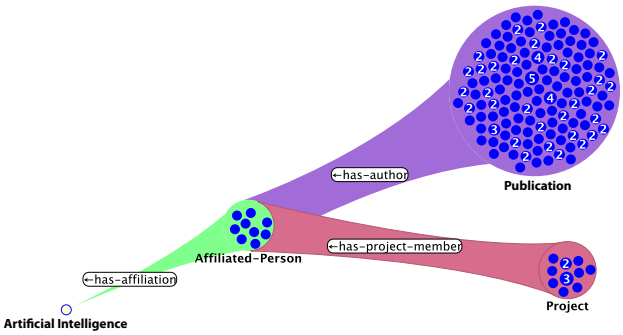

Figure 3: Multiple expansion paths.

preceding or succeeding clusters along the same expansion-path these entities are highlighted as sketched in the following: when hovering over an entity with the mouse pointer all entities in the preceding cluster which are related to that entity will be instantly highlighted. Then, for these entities the procedure repeats recursively. For instance, the leftmost club in Figure 4 shows deserts and the middle one all countries which they are located in. Here, the mouse pointer is hovering over "Algeria" in the second cluster and as a result all deserts which are located there are highlighted in the previous cluster. In addition, their names are displayed on the bottom of the surrounding cluster.

3.3 Analysing Quantities

Besides qualities, the representation of quantities is another important dimension of visual analytics: we found out that quite a number of queries inherently require to take quantities into account. Even if one can easily grasp how many entities are related to a given one with respect to a specific source by manually counting them, it is more complex to visually answer how man entities in a cluster are related to a specific one within its successor cluster. For instance, to answer the question within the MONDIAL domain which is the country in which the highest number of deserts can be found, one would start with the cluster representing all deserts. After expanding the club connected via the `located-in-country` relationship one gets a cluster consisting of all corresponding countries as one can see in Figure 4. Derived from well-known methods from visual analytics we have implemented a simple but powerful solution: the diameters of each country circle scales proportionally with the number of related deserts in the predecessor club. In

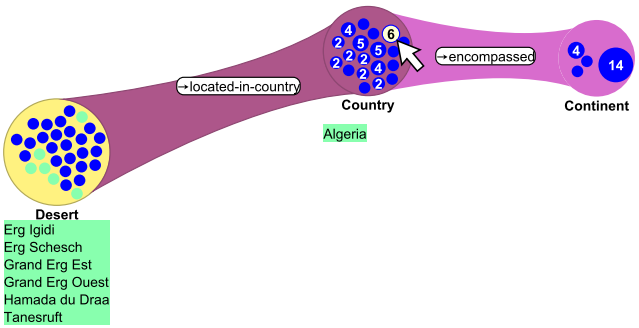

Figure 4: Utilizing quantities to answer questions such as "In which countries can be found the highest number of deserts?".

case that there are more than one single source entity their number is also drawn within the entity circle as shown in Figure 4. Furthermore, one can also see that most deserts can be found on the African continent (when hovering with the mouse pointer over the circle labeled "14" in the `encompassed` cluster the entity's name is shown and the interlinked entities in the preceding cluster are instantly highlighted).

Even if the original question does not refer to quantities, an additional rendering about quantities is a good benefit. Consider a follow-up expansion of Figure 2 (b) to get all co-authors of all publication of "Thorsten Liebig". At a glance one gets the information which is the co-author of most of the publications as shown in Figure 5 (a). The circle labeled "47" is Thorsten himself. Note that the same approach could answer from which affiliation tend to come most of them (Figure 5 (b)).

4. IMPLEMENTATION

Real-world data sets typically consist of thousands of thousand of entities and relationships between them. This makes great demands on the scalability and performance of the implementation of our visualisation approach. We address this with our decision to implement the visualisation and analysing component as a plug-in for our OntoTrack ontology framework: the visualisation is based on the Piccolo framework which can manage huge numbers of graphical objects [3]. The data management is based on a combination of a relational database storage and the wide-spread Java OWL 1.1 API [4] which provides high-level access mechanisms to concepts and relationships.

5. CONCLUSION

In this paper we presented a gainful combination of established methods from visual exploration and visual analytics introducing our new "club visualisation" metaphor. It enables to discover hidden connections between entities while not disturbing the user when exploring large structured data sets. The exploration examples throughout this paper should give an idea how this will help to gain deeper insights into large and heavily interlinked data sets from different domains. The exploration direction as well as the level of detail are determined by the user. In addition to qualities a visual feedback about quantities and the outlining of connected entities enables the user to easily grasp the overall structure as well as on the same time interrelations between

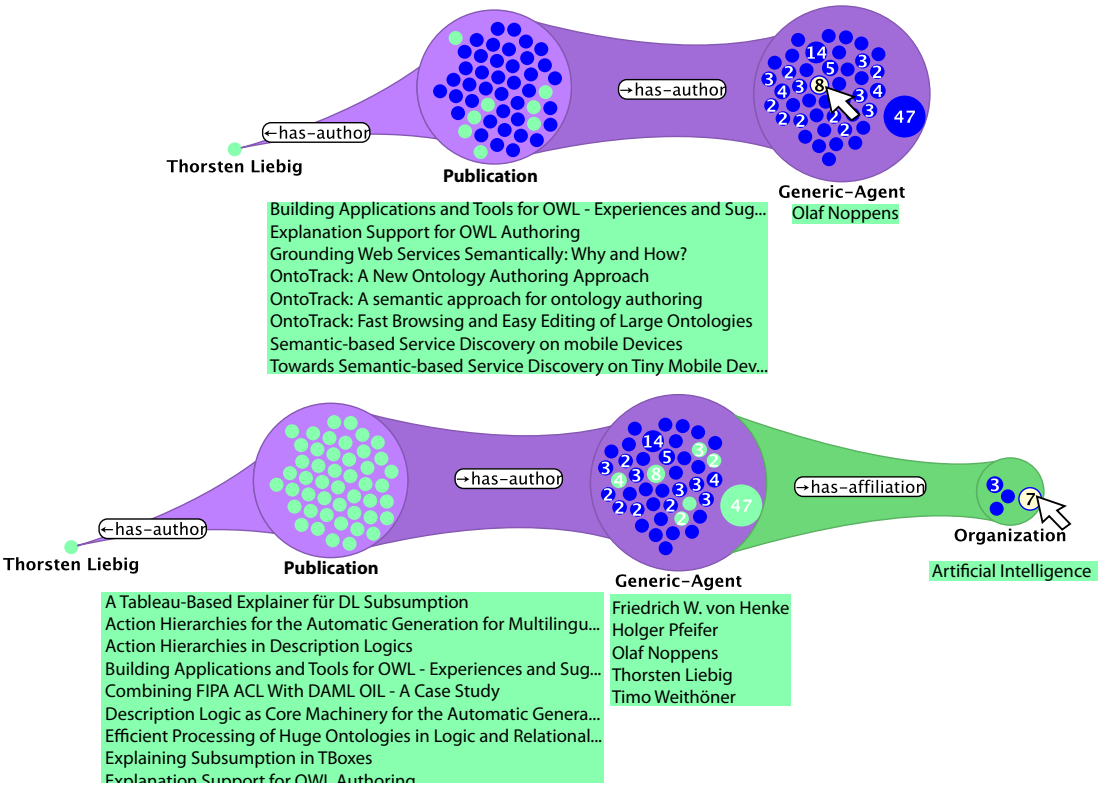

Figure 5: (a) Who is the most co-author of papers involving Thorsten. (b) From which affiliation the most co-authors come from?

specific entities. The interlocking of these techniques adds new exploration and understanding possibilities not found in current tools. All these features have been implemented and integrated into our OntoTrack framework.

Acknowledgments.

This work has been partly supported under the ReSIST Network of Excellence, which is sponsored by the Information Society Technology (IST) priority in the EU Sixth Framework Programme (FP6) under contract number IST 4 026764 NOE.

6. REFERENCES

[1] T. Anderson, Z. Andrews, J. Fitzgerald, B. Randell, H. Glaser, and I. Millard. The ReSIST Resilience Knowledge Base. In *Proc. of the 37th IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2007)*, June 2007.

[2] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of the 6th International Semantic Web Conference (ISWC 2007)*, volume 4805 of *LNCS*, pages 722–735. Springer, 2007.

[3] Ben Bederson, Jesse Grosjean, and Jon Meyer. Toolkit Design for Interactive Structured Graphics. Technical Report CS-TR-4432, University of Maryland, January 2002.

[4] Matthew Horridge, Sean Bechhofer, and Olaf Noppens. Igniting the OWL 1.1 Touch Paper: The OWL API. In *Proc. of the 3rd OWL Experiences and Directions Workshop (OWLED'07) at the ESWC'07*, Innsbruck, Austria, 2007.

[5] Afraz Jaffri, Hugh Glaser, and Ian Millard. URI Identity Management for Semantic Web Data Integration and Linkage. In *Proc. of the 3rd International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2007)*, Vilamoura, Algarve, Portugal, 2007. Springer.

[6] Daniel A. Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, 2001.

[7] Thorsten Liebig and Olaf Noppens. OntoTrack: A semantic approach for ontology authoring. *Journal of Web Semantics*, 3(2):116 – 131, 2005.

[8] Wolfgang May. Information extraction and integration with Florid: The Mondial case study. Technical Report 131, Universität Freiburg, Institut für Informatik, 1999.

[9] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. of the IEEE Symposium on Visual Languages*, pages 336–343, Washington, USA, 1996.