Making Sense of Research with Rexplore

Enrico Motta[1], Francesco Osborne[2]

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK
e.motta@open.ac.uk
[2] Dept. of Computer Science, University of Turin, Turin, Italy
osborne@di.unito.it

Abstract. While there are many tools and services which support the exploration of research data, by and large these tend to provide a limited set of functionalities, which cover primarily ranking measures and simple mechanisms for relating authors. To try and improve over the current state of affairs, we are developing a novel tool for exploring research data, which is called Rexplore. Rexplore builds on an intelligent algorithm for automatically identifying hierarchical and equivalence relations between research areas, to provide a variety of functionalities and visualizations to help users to make sense of research data. These include visualizations to detect trends in research; ways to cluster authors according to several dynamic similarity measures; and fine-grained mechanisms for ranking authors, taking into account parameters such as ranking criterion, career stage, calendar years, publication venues, etc.

Keywords: Research Data, Bibliographic Data, Data Visualization, Data Exploration, Visual Analytics, Scholarly Semantic Relations.

1 Introduction

Understanding what goes on in a research area is no easy task. Typically, for a given topic, this sensemaking process may require exploring information about a variety of entities, such as publications, researchers, research groups, projects, events, and others, as well as understanding the relationships which exist between them. In addition, different categories of users tend to be interested in exploring different aspects of this space. For instance, a 1st year PhD student in the Semantic Web area would likely be interested in the main approaches, projects, and publications relevant to her topic of choice. She will also be interested in identifying the key people and research groups, but her exploration needs will certainly be very different from those of a company, who may want to improve their expertise about a specific topic by establishing a relationship with an appropriate research group on the basis of their expertise, status in the field, and geographical location. Research data are also of great interest to research managers, funding bodies and government agencies, who may want to find out about the performance of specific individuals and groups, and compare them with their peers both at national and international level.

There are many tools and services currently available, which already provide a wide variety of functionalities to support exploration of research data. These include bibliographic search engines, such as *Microsoft Academic Search* and *Google Scholar*; large research databases, such as *Sciverse Scopus, DBLP* and *PubMed*; reference

management applications, such as *Mendeley*; visual analytics tools, such as *CiteSpace*; tools which focus on mining and visualizing relations between researchers, such as *Arnetminer*; and many others[1]. Nevertheless, as Dunne et al. point out [1], there is still a need for an *integrated solution*, where the different functionalities and visualizations are provided in a coherent manner, through an environment able to support a seamless navigation between the different views and functionalities. In addition, we would also argue that there are a number of important functionalities, relevant to the process of making sense of research data, which are currently not well supported. For instance, as discussed in our companion paper accepted for the ISWC 2012 research track [2], semantic relations exist between research areas, which help to structure the data space and make it possible to go beyond visualizations and searches based on a purely syntactic analysis of the data. Let's consider the Semantic Web again as an example. If our aforementioned PhD student is browsing papers related to this area, she may not be necessarily only interested in papers explicitly labeled "Semantic Web", but, e.g., she may also want to consider papers in Ontology Engineering or Linked Data, even though such papers may not be explicitly tagged as Semantic Web papers. Hence, environments for exploring research data need to make use of algorithms, such as the one described in [2], which can automatically discover relations between research areas and make it possible to go beyond purely syntactic approaches to search, while at the same time also addressing the limitations associated with manually constructed taxonomies [2].

Another weakness of current solutions concerns the limited support for identifying and visualizing relations between researchers. These are arguably crucial to the research sensemaking process, because the different ways groups of researchers co-operate, follow similar research trajectories through different topics, and exhibit other kinds of common patterns in the evolution of their careers and publishing behaviours, arguably provide key indicators of the dynamics of a research area. For instance, it may be very useful for a PhD student to be aware that a significant group of researchers has moved over the past 5 years from topic X to topic Y, exhibiting similar publishing behaviours, while not necessarily collaborating explicitly. While some existing systems already provide different ways of visualizing relations between researchers, these tend to cover simple 'static' ones, such as co-authorship.

In sum, it is our view that there is a need to develop new solutions for exploring research data, addressing the two issues discussed above: i) the need for a seamless integration of views and functionalities in the exploration process and ii) the need for new advanced functionalities, able to go beyond the 'document search' paradigm underlying most existing solutions, to provide new ways to discover patterns and relations between the different classes of entities in the research data space.

2 Making Sense of Research with Rexplore

The semantic relationships among authors and topics are at the heart of many new functionalities of Rexplore. In particular they are used for 1) computing novel kinds of

[1] In this short paper it is not possible to do justice to the huge variety of relevant work, hence we only list a few of the best known solutions. It is also important to note that the above classification is only approximate. In practice many tools integrate different functionalities –e.g., most bibliographic search engines and databases also provide visual analytics functionalities.

similarities and ranking metrics that take in consideration the semantic characterization of research areas; 2) improving the ability of Rexplore to interpret user queries; and 3) enabling a novel graph-based navigation technique, which combines both semantic relationships and automatically computed metrics to generate links between the elements of the domain.

Currently, the following functionalities and visualizations are provided[2]:

- **Author Ranking and Activity.** Author ranking is a standard functionality, which is provided by most systems and, likewise, Rexplore provides a wide variety of ranking mechanisms, including h-index, citations, publications, etc. These rankings can be parameterized with respect to career stage, calendar years, and publication venues, thus providing the user with fine-grained control over the visualizations. For example, not only Rexplore makes it possible to rank Semantic Web authors by number of publications – a functionality already provided by many existing tools, but it also makes it possible, for example, to focus on the ranking of the best early-career researchers over the past n years, taking into account only data related to the top publication venues in the Semantic Web. This is particularly useful in scenarios, such as recruitment, where the focus tend to be on people who are at a specific career stage. Rexplore also makes it possible to plot the impact of an author over time, both in absolute terms and relatively to the default standard for a particular area. Multiple integrated visualizations of an author's activity are also provided, including the ability to visualize her citations or publications over time, and to parameterize these with respect to the relevant topics.
- **Relations between Authors.** Rexplore makes it possible to visualize a variety of relations between authors, most of which are dynamically constructed on the basis of the patterns emerging from their publishing behaviour and impact over time. For example, Rexplore makes it possible to cluster together researchers who exhibit similar publishing and impact trajectories, whether in the same or different fields. In addition, it is also possible to visualize similarity relations between authors who follow the same research path, by looking at the similarities between their research interests over time. Here we make use of the *Klink* algorithm [2], which ensures that the matching between research areas is 'semantic', rather than simply based on keyword matching. This solution is actually very generic and can also be used by applications in other domains, which wish to consider semantic relations when calculating similarity metrics.
- **Topic Evolution.** Rexplore provides a variety of ways to support a user's understanding of the dynamics of a research area. For example, it makes it possible to visualize *migration patterns* across areas, thus allowing users to understand where people working in a new area are coming from, and whether an area is growing or reducing –i.e., whether there is a gain or loss of researchers between two areas. Another view shows the evolution of a topic over time, highlighting, for example, the main sub-topics, identified automatically using the Klink algorithm, which are emerging, as well as those which are decreasing in importance.

[2] While the ultimate aim of this work is to provide a comprehensive set of functionalities, covering a wide range of entities relevant to the research space, the current version of Rexplore (*alpha v0.9*) only covers authors, groups (of authors), topics, and publications.

Rexplore is implemented mostly in PHP and the visual part of the application uses JavaScript to ensure we do not depend on any external plugin. In particular we use the *Highcharts* library for the charts and a modified version of *JavaScript InfoVis Toolkit* for the graphs. The metadata we use come mainly from Microsoft Academic Search (http://academic.research.microsoft.com/) and DBLP (http://www.informatik.uni-trier.de/~ley/db/). The first comprises over 30 million papers, while the latter is a database for computer science that covers more than two million articles. As of August 2012, Rexplore contains the metadata regarding 15 million papers, focusing in particular on the Computer Science area. These data are enriched by means of a number of algorithms, which are able to infer new information –e.g., by discovering similarities and patterns in the data, by creating links between research topics, etc.

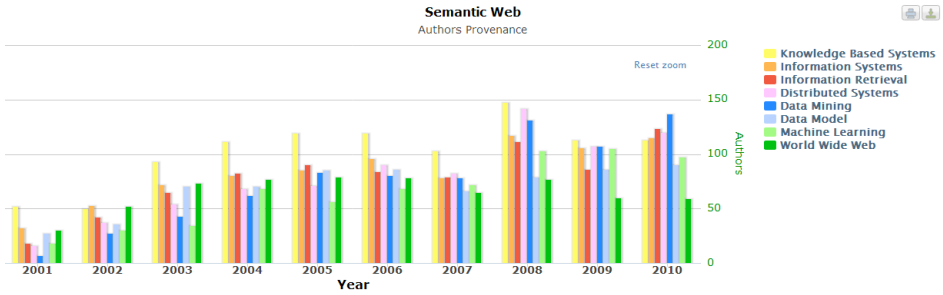

Fig. 1. One of the many visualizations provided by the current version of Rexplore. The snapshot shows the main research areas from which Semantic Web authors originated in the past decade. The figure shows that over the years most newcomers have come from the Knowledge Based Systems area, up until 2010, when for the first time most new authors came from Data Mining.

3 Conclusions

The current version of Rexplore already provides an array of interesting functionalities, many of which go well beyond what is available in other current tools. Nevertheless, we are still at a relatively early stage and many more functionalities are planned. In particular we plan to improve substantially the look and feel of the system, which currently is very much 'browser-like'. We will also extend the range of inputs to the system, by adding information gathered from social networks and other web sources and we also plan to add geographic visualizations to create maps of research groups and topic tendencies. Finally we are also working on improving interactivity and customization, with the aim of allowing users to customize the topic structures generated by Klink, as well as other aspects of the system.

References

1. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., and Dorr, B. (2012). Rapid Understanding of Scientific Paper Collections: Integrating Statistics, Text Analytics and Visualization. To appear in *JASIST*, 2012.
2. Osborne, F. and Motta, E., (2012). Mining Semantic Relations between Research Areas. *11th International Semantic Web Conference (ISWC 2012)*. Boston, MA.