Rexplore: unveiling the dynamics of scholarly data

Conference or Workshop Item

Rexplore: Unveiling the Dynamics of Scholarly Data

Francesco Osborne, Enrico Motta
Knowledge Media Institute
The Open University
Milton Keynes MK7 6AA
United Kingdom
{francesco.osborne, e.motta}@open.ac.uk

ABSTRACT

Rexplore is a novel system that integrates semantic technologies, data mining techniques, and visual analytics to provide an innovative environment for making sense of scholarly data. Its functionalities include: i) a variety of views to make sense of important trends in research; ii) a novel semantic approach for characterising research topics; iii) a very fine-grained expert search with detailed multi-dimensional parameters; iv) an innovative graph view to relate a variety of academic entities; iv) the ability to detect and explore the main communities within a research topic; v) the ability to analyse research performance at different levels of abstraction, including individual researchers, organizations, countries, and research communities.

Categories and Subject Descriptors

I.2 **[Artificial Intelligence],** H.3.7 **[Digital Libraries],** H.2.8 **[Data Mining],** H.5.2 **[User Interface].**

General Terms

Algorithms, Design.

Keywords

Scholarly Data, Visual Analytics, Data Exploration, Empirical Evaluation, Ontology Population, Data Mining, Data Integration.

1. INTRODUCTION

In recent years the quality and number of accessible scholarly data has considerably increased [1]. However, it is still not easy to explore and make sense of these data and often the most interesting pieces of information (e.g., topics trends, research community dynamics) cannot be easily extracted from today's solutions, if at all. Hence, it can be argued that we need better systems for the exploration of the academic world, able to offer novel functionalities for interpreting and analysing data at different levels of granularity, making hidden knowledge explicit, highlighting patterns, and forecasting future tendencies. Some of today tools partially answer these needs by including a variety of research entities (e.g., authors, venues, topics) and allowing the visualization of author trends and academic networks. For example, Microsoft Academic Search displays the trends of a limited number of broad topics (e.g., Artificial Intelligence) and supports the exploration of authors' citation networks. Arnetminer, Scopus and CiteSeerX offer similar services.

However these and others tools usually miss a number of important functionalities, lacking the ability: i) to characterize semantically the research topics (e.g., understanding that topic A is a sub-area of topic B), ii) to investigate research trends and author migrations at different levels of granularity, iii) to relate authors in terms of common interests or shared academic trajectories, iv) to detect diachronic research communities and study their evolution over time, and v) to perform fine-grained academic expert search taking into account also relations between authors.

Moreover, while some specific tools allow for advanced sense-making tasks (e.g., community detection), there is still the need for an integrated solution, in which the different functionalities and visualizations are provided in a coherent manner through an environment able to support a seamless navigation between different views, interfaces and entities.

2. OVERVIEW OF REXPLORE

Rexplore [2] is a modern environment for the exploration of scholarly data, which addresses the aforementioned limitations. It allows exploring a rich set of entities (e.g., authors, research areas, venues, organizations, countries), which can be analysed at different levels of abstraction, using a variety of visualizations and metrics. The system integrates a number of data sources, including: DBLP++ (dblp.l3s.de), Microsoft Academic Search API (academic.research.microsoft.com), DBpedia (dbpedia.org) and GeoNames (www.geonames.com). Rexplore implements also a disambiguation module, which uses a number of features (e.g., co-authorships, topic similarity) to assign each publication to the correct author. As of March 2014, the Rexplore database contains 23 million papers and 2.2 million authors, mostly about Computer Science. In this short overview we will discuss some of the main features of Rexplore.

The Graph View. The graph view is a highly interactive tool for exploring the space of research entities and their relationships using faceted filters. It takes as input authors, organizations, countries or research communities and generates their relationship graph, allowing the user to choose among a variety of connections, ranking criteria, views and filters. Entities, represented by nodes, and relationships, represented by links, can be clicked on to obtain additional information. The graph view offers four types of relations: co-publication, co-citation, topic similarity and temporal topic similarity. The topic similarity reflects how similar two authors are with respect to their research areas. The temporal topic similarity builds on the topic similarity and makes possible the identification of researchers who appear to have worked on similar topics at the same time, i.e., who share the same *research trajectory*. These connections can be further characterized by a number of filters. For example, it is possible to plot the collaboration network of authors who work in the same field or who have published in the same venue.

Research Area Analysis. While most systems use keywords as proxy for research topics, Rexplore relies on an OWL ontology, which characterizes research areas and their relationships. This

978-1-4799-5569-5/14/$31.00

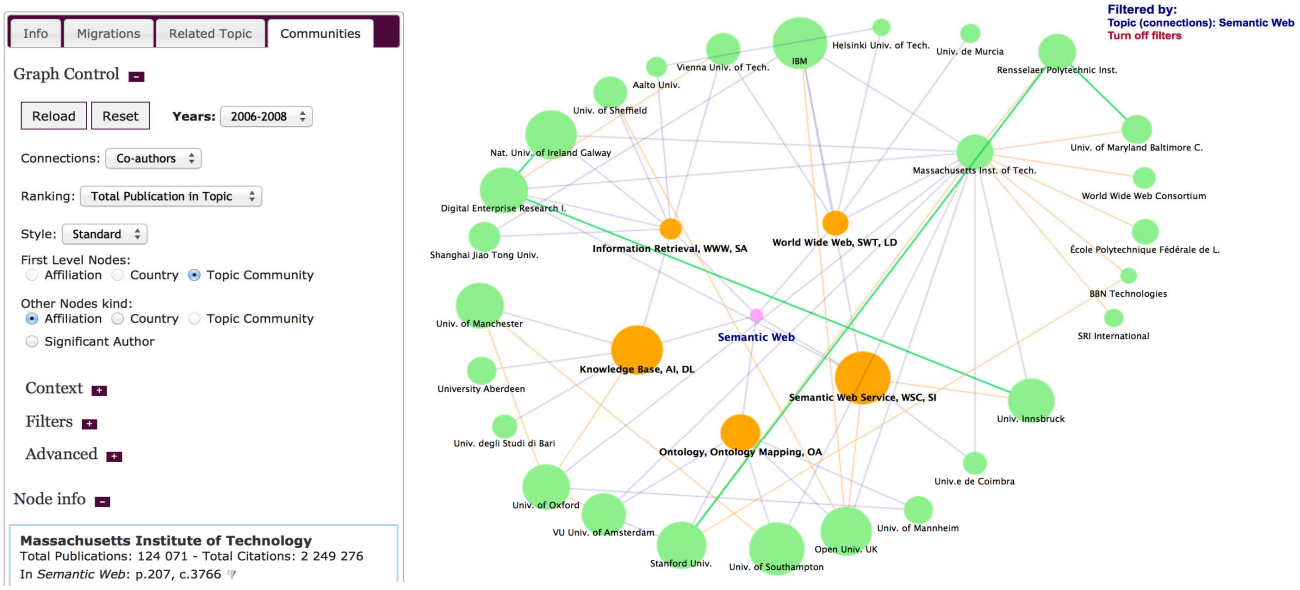

Figure 1. The main Semantic Web communities and some of the associated most significant organizations in 2006-2008.

ontology is automatically populated and periodically updated by Klink [3], an algorithm that uses statistical and machine learning techniques 1) to identify research areas from a given set of keywords, filtering out those that do not denote research areas, 2) to compute three types of semantic relationships between topics and 3) to return a fully populated OWL ontology describing the topic structure. The three semantic relationships detected by Klink are 1) *skos:broaderGeneric* (topics T1 is a sub-area of topic T2), 2) *contributesTo* (research in topic T1 is an important contribution to research in topic T2, however T1 is not a sub-topic of T2) and 3) *relatedEquivalent* (T1 is equivalent to topic T2). The retuned topic ontology is used in a variety of ways for enhancing the search engine, data mining techniques and visualizations provided by Rexplore.

In the topics page it is possible to analyse the dynamics of a topic by visualizing the trends of different sub areas or related topics. For instance, it is possible to see which subtopics are growing the most and thus are most promising for the future. This interface also allows visualizing migrations of researchers between two research topics in subsequent years. Rexplore offers also a graph view to explore the research communities within a topic, their evolution in subsequent years and their relationships with authors, countries and organizations. A user can thus gain an immediate knowledge of the history, the main groups of authors, the collaborations and the organizations active in each research area. Technically, this is achieved by TST [4], an algorithm which identifies communities of researchers who appear to follow a similar research trajectory. For example, Figure 1 shows a graph view of the main Semantic Web research communities in which the user clicked on MIT to obtain additional details.

Author and Group Analysis. Every author in Rexplore has a personal page which offers a variety of metrics and visualizations to analyse his/her performance, trends and collaborations. This interface allows users to plot on a timeline citations and publications in a number of research areas at different levels of granularity. Hence, the user can analyse the author trends in general areas, such as "Machine Learning", but also zoom on sets of more specific topics (e.g., "Named Entity Recognition"). Moreover, this interface allows comparing the performance of an author with the authors returned from any search query. For instance, it is possible to compare an author with all researchers of similar seniority active in the same field. Authors' groups, (i.e., organizations, countries or communities) have also a dedicated page, which allows studying the trends and composition of the group or using the graph view to explore the group connections with significant authors or other groups.

Multi-criteria Search. Rexplore offers a fine-grained search facility for authors, publications and organizations with respect to detailed multi-dimensional parameters. The search results can be further refined, explored or filtered by using the graph view. Users can thus combine search filters and connection filters to formulate complex queries such as "the list of young co-workers (with expertise in Machine Learning) of the rising stars in Digital Library and Data Mining, who published in JCDL and work for a UK institution". Moreover, Rexplore supports the data exploration process by remembering the search queries and highlighting the related concepts in the following pages. For example, if the user searched for "authors with expertise in HCI and publications in CHI", the system will highlight these research area and venue in a number of views.

3. REFERENCES

[1] C. L. Borgman, *Scholarship in the digital age,* Ed. Cambridge: MIT press, 2007.

[2] F. Osborne, E. Motta, and P. Mulholland. "Exploring Scholarly Data with Rexplore." *The Semantic Web–ISWC 2013*, Ed. Berlin, Germany: Springer-Verlag, 2013, pp. 460-477.

[3] F. Osborne and E. Motta. "Mining semantic relations between research areas." *The Semantic Web–ISWC 2012,* Ed. Berlin, Germany: Springer-Verlag, 2012, pp. 410-426.

[4] F. Osborne, G. Scavo, and E. Motta. "Identifying diachronic topic-based research communities by clustering shared research trajectories." The Semantic Web: Trends and Challenges. Ed. Cham, Switzerland: Springer International Publishing, 2014, pp. 114-129.