Understanding Research Dynamics

Francesco Osborne(✉) and Enrico Motta

Knowledge Media Institute, The Open University,
Milton Keynes MK7 6AA, UK
{francesco.osborne,e.motta}@open.ac.uk

Abstract. Rexplore leverages novel solutions in data mining, semantic technologies and visual analytics, and provides an innovative environment for exploring and making sense of scholarly data. Rexplore allows users: (1) to detect and make sense of important trends in research; (2) to identify a variety of interesting relations between researchers, beyond the standard co-authorship relations provided by most other systems; (3) to perform fine-grained expert search with respect to detailed multi-dimensional parameters; (4) to detect and characterize the dynamics of interesting communities of researchers, identified on the basis of shared research interests and scientific trajectories; (5) to analyse research performance at different levels of abstraction, including individual researchers, organizations, countries, and research communities.

Keywords: Scholarly data · Visual analytics · Data exploration · Semantic Web · Semantic technologies · Ontology population · Data mining · Data Integration

1 Introduction

Understanding what goes on in a research area is a complex sensemaking process, which requires exploring information about a variety of entities, such as publications, publication venues, researchers, research communities, events and others, as well as understanding their relationships.

Many currently available tools already provide a variety of functionalities for the exploration of research data. These include bibliographic search engines (e.g., Microsoft Academic Search, Google Scholar), large research databases (e.g., Sciverse Scopus, PubMed), visual analytics tools (e.g., CiteSpace [1]), tools which focus on mining and visualizing relations between researchers (e.g., Arnetminer [2]), and others. These tools however usually miss a number of important functionalities, such as the ability (i) to investigate research trends effectively at different levels of granularity, (ii) to relate authors 'semantically' (e.g., in terms of common interests or shared academic trajectories), (iii) to detect dynamically-characterized research communities (e.g., all researchers working on RDF) and relate them to other entities (e.g., universities, countries, or specific authors), and (iv) to perform fine-grained academic expert search along multiple dimensions. Moreover, while some specific tools may address one or two of the aforementioned functionalities, there is still the need for an integrated solution [3], where the different functionalities and visualizations are provided in a

V. Presutti et al. (Eds.): SemWebEval 2014, CCIS 475, pp. 101–107, 2014.
DOI: 10.1007/978-3-319-12024-9_13

coherent manner, through an environment able to support a seamless navigation between different views, interfaces and entities.

Another important limitation of current tools is their lack of semantic characterization of important entities, such as research areas. Most of the tools use keywords as proxies for research areas [2]; however the keywords associated to academic publications lack structure and are often noisy [4]. Important relations between research areas, such as an area being a sub-area of another one, are neglected: for example, when a user search for papers about "Semantic Web", these systems will ignore the publications tagged only as "Linked Data". Semantic technologies can solve this problem, by allowing for a formal definition of research topics and their relationships.

2 Overview of Rexplore

To address the limitations discussed above, we developed Rexplore [5], a system which leverages novel solutions in data mining, semantic technologies and visual analytics, and provides an innovative environment for exploring and making sense of scholarly data[1]. The back-end of Rexplore is implemented in PHP and Java, while the interface and the visualizations are in HTML5 and JavaScript.

In this short overview we will discuss some of the main features of Rexplore.

Data Integration. Rexplore integrates a variety of data sources in different formats, including: the MAS API[2], DBLP++[3] and DBpedia[4]. The process of generating the populated topic ontology, described in the next subsection, exploits information collected from Google Scholar, EventSeer[5] and Wikipedia.

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 authors. The integration and disambiguation process for the organizations makes use of Linked Open Data and in particular tries to map each organization and location to a DBpedia entity. Rexplore can integrate paper metadata in XML, RDF and SQL, but not yet extract data from PDF. The minimal metadata needed for a paper to be included in Rexplore are the title, the names of the authors and the year. As of June 2014, Rexplore contains 23 million papers and 2.3 million authors.

Topic Ontology and Klink. While most systems use keywords as proxies for research topics, Rexplore relies on an OWL ontology, which characterizes research areas and their relationships. This ontology is automatically populated and periodically updated by Klink [4], an algorithm that uses statistical and machine learning techniques (1) to identify research areas from the given set of keywords, filtering out keywords that do not denote research areas (e.g., "Case Study" or "Large Scale"), (2) to compute three

[1] http://technologies.kmi.open.ac.uk/rexplore
[2] http://academic.research.microsoft.com
[3] http://dblp.l3s.de/dblp++.php
[4] http://dbpedia.org
[5] http://eventseer.net

types of semantic relationships between topics and (3) to return a fully populated OWL ontology describing the topic domain.

The three semantic relationships detected by Klink are (1) *skos:broaderGeneric* (topics T_1 is a sub-area of topic T_2; e.g., "Linked Data" is a sub-area of "Semantic Web"), (2) *contributesTo* (research in topic T_1 is an important contribution to research in topic T_2, however T_1 is not a sub-topic of T_2; e.g., "Ontology Engineering" contributes to "Semantic Web"), and (3) *relatedEquivalent* (T_1 is equivalent to topic T_2; e.g., "Ontology Matching" is equivalent to "Ontology Alignment"). Klink has been tested mainly on Computer Science, but we plan to evaluate it soon on other fields. The returned topic ontology is used in a variety of ways, e.g., for rewriting queries by taking in consideration topic relationships, for analysing authors' trends at different levels of granularity, and for enhancing the community detection algorithm.

Semantic Topic Analysis. A simple but effective method to take advantage of the OWL knowledge base is to consider every publication tagged with topic T_1 to be also about topic T_2, if T_2 is *broaderGeneric* than T_1, or *relatedEquivalent* to T_1 (it should be noted that *broaderGeneric* is transitive). This has a dramatic effect on the quality and size of data available for each topic: for example, our knowledge base includes 11,998 publications tagged with the *string* "Semantic Web", while the publications regarding the *topic* "Semantic Web" (including sub-topics, such as "Linked Data") are almost twice as many (22,143).

For analysing a topic, Rexplore provides an interface that includes: (i) general information about the topic, e.g., the relevant authors and publications, (ii) the *topic navigator,* an interface to browse topics via their semantic connections, (iii) visual analytics on *broaderGeneric* and *contributesTo* topics, (iv) visual analytics on authors' migration patterns, (v) a graph view to explore the research communities active in the topic and their relationships with authors, countries and organizations. For a given topic, Rexplore allows users to visualize on a timeline three kinds of trends: publication trends, author trends and migration trends. The first two provide a concise view of the number of publications and the number of researchers working on the topic over time. The latter illustrate the number of estimated migrations between two topics and it is computed by analysing the degree of shifting in authors' interest. More information on how Rexplore handles topic trends can be found in [6].

Multi-criteria Search. Rexplore offers a fine-grained search functionality for authors, publications and organizations with respect to detailed multi-dimensional parameters. For example, authors can be filtered by (i) name or part of it, (ii) career age (i.e., the time from the first published work), (iii) topics of interest, (iv) venues in which they have published and (v) country/organization. Both venue and topic fields accept multiple values, which can be combined using logical connectives. Moreover, the search interface is enhanced by the *graph view*, which shows the connections of query results with other entities. Hence, the search results can be further refined, explored or filtered by considering their connections. This solution allows building with a few clicks complex queries such as "the career young co-authors (with expertise in Machine Learning) of the prominent researchers in Semantic Web and Data Mining who work for a UK institution". Rexplore also supports the subsequent data exploration by remembering the initial queries and highlighting the related concepts in the following

pages. For example, if the user searches for "authors with expertise in Semantic Web who published in ESWC", the system will highlight the research area and the venue in the following views.

The Graph View. The *graph view* is a highly interactive tool to explore 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 connections, represented by links, can be clicked on to obtain additional information. The dimensions of the nodes are proportional to the metric chosen by the user, e.g. if the user chooses "citations in Artificial Intelligence" the entities with more citations in this topic will be the biggest.

Users can choose from 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 and takes advantage of the topic ontology generated by Klink. The temporal topic similarity (TTS) (see [4, 7]) builds on the topic similarity and makes possible the identification of researchers who worked on similar semantic topics at the same time. Both the nodes and the relationships can be filtered by a variety of parameters. For example, the user can visualize only the collaboration in the field of "Ontology Matching" with career young researchers who published in ESWC.

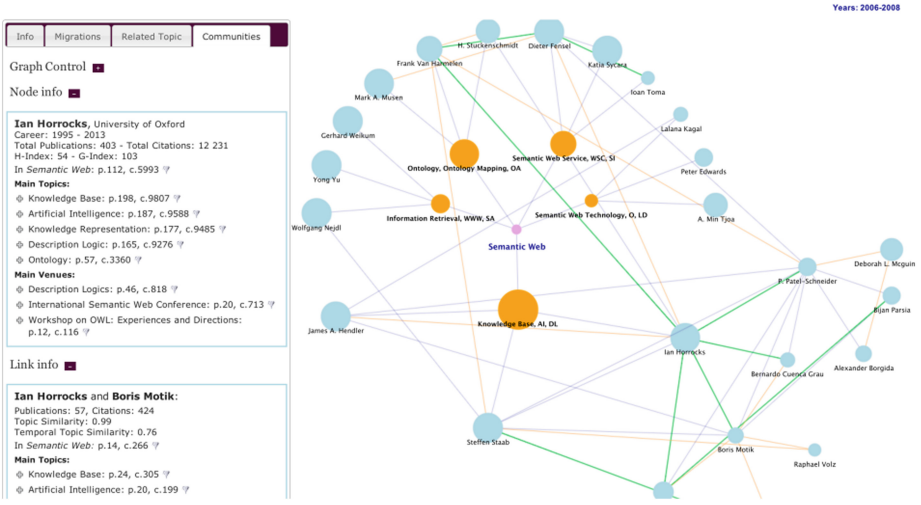

Fig. 1. The main Semantic Web communities in 2006–2008 and some of their most significant authors.

Community Detection. Rexplore integrates a novel algorithm called *TST* [7] (*Temporal Semantic Topic-Based Clustering)*, which exploits the TTS to identify communities of researchers who appear to follow a similar *research trajectory.* Technically this is achieved by running a Fuzzy C-Mean algorithm, which uses as

norm a variation of the *temporal topic similarity* metric, applied to distributions of semantic topics over time, associated with each researcher.

The communities produced by the TST algorithm have some very interesting features. First, they are not snapshots of static collaborations, but rather they are diachronic entities, with topic distributions and interests evolving over time, mirroring trends, technological breakthroughs and new visions. Hence, they allow users to make sense of the dynamics of the research world – e.g., migrations of researchers from one topic to another, new communities being spawn by older ones, shifts of interests, communities splitting, merging, ceasing to exist, etc. Secondly, in contrast with methods that rely on co-authorship or citation networks, their computation does not require a complete graph of relations between community members. Finally, since they are fuzzy clusters, they can address the common situation in which a researcher is active in more than one community. For a full description of TST see [7].

Rexplore relies on TST to detect the communities within a certain *broad topic* (e.g., Semantic Web) and offers a graph view in the topic page to explore their most significant authors and organizations. Hence, it makes it easy to gain an immediate knowledge about the main dynamics of a research area. For example, Fig. 1 shows the top 5 sub-communities in the Semantic Web area in the interval 2006-2008 (shown as first level nodes in the graph view): Knowledge Base/AI/Description Logic, Ontology/Ontology Matching, Information Retrieval/WWW, Semantic Web Service/Semantic Interoperability and Semantic Web Technology/Linked Data. Here the user has chosen to visualize some of the most significant researchers of each community and is exploring the co-authors of Ian Horrocks, which is one of protagonists of the "Knowledge Base/AI/Description Logic" community. Links in the graph view can also be inspected and Fig. 1 also shows additional details about the academic connections between Ian Horrocks and Boris Motik.

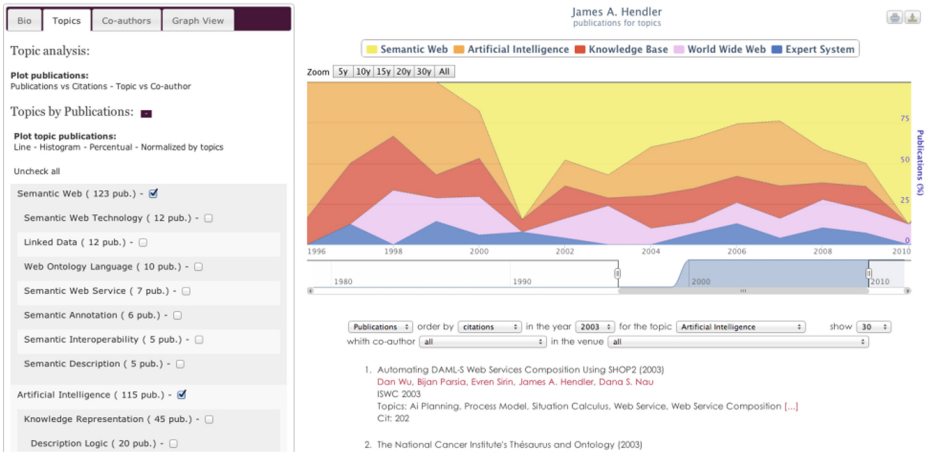

Fig. 2. Topic analysis tool.

Author and Group Analysis. Every author in Rexplore has a personal page which offers a variety of metrics and visualizations to analyse the authors' performance, trends and collaborations. One of the most useful features is the topic analysis tool, which allows users to plot on a timeline the performance of an author in different research areas. Figure 2 shows a view of this tool displaying the main topics of the publications by James Hendler, one of the originators of Semantic Web. On the left the user can select the kind of chart and the topics to be shown. The chart is interactive and the user can click on it to visualize the list of publications relative to a year and a topic. For example, Fig. 2 shows the publications of James Hendler on "Artificial Intelligence" in 2003. The publication list can be further refined by selecting additional filters, such as the co-author and the venue.

By default, Rexplore selects the more general topics (e.g., "Semantic Web" rather than "Linked Data") to show the big picture of the author's interests and how they evolved in time. However the topics and sub-topics are displayed in a multilevel list and the user can choose to adopt different granularity levels. For example a user can conduct a high level analysis by focusing on the main topics (e.g., "Artificial Intelligence" or "Ontology") or otherwise zoom in on one of them (e.g., "Ontology") and further analyse its sub-topics (e.g., "Ontology Engineering", "Ontology Mapping", "Ontology Learning"). Citations and publications can also be normalized according to the average citation numbers of the considered topics, allowing users to easily compare researchers from different disciplines (e.g., Biology and Mathematics). The topic analysis tool can also compare a researcher with those working in the same field or having similar seniority or coming from a specific country/organisation. For example, it can be used to check how a career young researcher from UK working in "Machine Learning" ranks in term of a certain metric (e.g., H-Index) among the researchers with the same characteristics.

Authors' groups, which can be organizations, countries or research communities, have a simpler interface at the moment. It is possible to study the trends of a group in terms of publications and citations and to browse the main researchers and publications by years. Moreover, the user can rely on the graph view to explore the connections of a group with significant authors or with other groups. For example, it is possible to plot the Open University network of collaborations in the Semantic Web and to explore the details of each of them.

3 Conclusion

In this paper we presented Rexplore, an innovative system for exploring scholarly data, which relies on advanced data mining algorithms and semantic technologies. Rexplore implements a variety of innovative functionalities and arguably provides the most advanced solution currently available.

References

1. Chen, C.: CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. **57**(3), 359–377 (2006)
2. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: Proceeding of the 14th International Conference on Knowledge Discovery and Data Mining, pp. 990–998 (2008)
3. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B.: Rapid understanding of scientific paper collections: integrating statistics, text analytics, and visualization. J. Am. Soc. Inf. Sci. Technol. **63**(12), 2351–2369 (2012)
4. Osborne, F., Motta, E.: Mining semantic relations between research areas. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 410–426. Springer, Heidelberg (2012)
5. Osborne, F., Motta, E., Mulholland, P.: Exploring scholarly data with rexplore. In: Alani, H., et al. (eds.) ISWC 2013, Part I. LNCS, vol. 8218, pp. 460–477. Springer, Heidelberg (2013)
6. Osborne, F., Motta, E.: Exploring research trends with Rexplore. D-Lib Mag. **19**(9/10), 4 (2013)
7. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 114–129. Springer, Heidelberg (2014)