Exploring Scholarly Data with Rexplore

Francesco Osborne[1, 2], Enrico Motta[1], Paul Mulholland[1]

[1]Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
{francesco.osborne, enrico.motta, paul.mulholland}@open.ac.uk
[2]Dept. of Computer Science, University of Torino, 10149 Torino, Italy
osborne@di.unito.it

Abstract. Despite the large number and variety of tools and services available today for exploring scholarly data, current support is still very limited in the context of sensemaking tasks, which go beyond standard search and ranking of authors and publications, and focus instead on i) understanding the dynamics of research areas, ii) relating authors 'semantically' (e.g., in terms of common interests or shared academic trajectories), or iii) performing fine-grained academic expert search along multiple dimensions. To address this gap we have developed a novel tool, Rexplore, which integrates statistical analysis, semantic technologies, and visual analytics to provide effective support for exploring and making sense of scholarly data. Here, we describe the main innovative elements of the tool and we present the results from a task-centric empirical evaluation, which shows that Rexplore is highly effective at providing support for the aforementioned sensemaking tasks. In addition, these results are robust both with respect to the background of the users (i.e., expert analysts vs. 'ordinary' users) and also with respect to whether the tasks are selected by the evaluators or proposed by the users themselves.

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

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, publication venues, researchers, research groups, events, and others, as well as understanding the relationships which exist between them. Such exploration and sensemaking tasks can take place in a variety of contexts, involving different categories of users. For instance, one of the authors of this paper is Editor-in-Chief of a scientific journal and in such a role he regularly needs to consider competing proposals for special issues, a task which requires (among other things) to analyze the dynamics of one or multiple research areas, in order to formulate a view on whether the proposals in question concern areas that are 'hot' and growing, or are instead to a lesser extent at the cutting edge. In other task contexts, such *scholarly data* are also of great interest to research managers, funding bodies and government agencies, who i) 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; or ii) may need to gather objective evidence about research trends to inform funding policy decisions.

Obviously, there are many tools and services currently available, which already provide a wide variety of functionalities to support the exploration of scholarly data – see Section 2.1 for a review of the state of the art. Nevertheless, as Dunne et al. point out [1], there is still a need for an *integrated solution*, where the different scholarly tasks are provided in a coherent manner, through an environment able to support a seamless navigation between different views and functionalities. In addition, as discussed in detail in the next section, we believe that there are also a number of important functionalities, which are crucial to providing effective support for exploring and making sense of scholarly data, but are currently missing from existing solutions. These include (but are not limited to) 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), and iii) to perform fine-grained academic expert search along multiple dimensions.

To address this gap we have developed a novel tool, Rexplore [2], which integrates statistical analysis, semantic technologies, and visual analytics to provide effective support for exploring and making sense of scholarly data. In this paper, we illustrate the main innovative elements of the tool and we also present the results from a task-centric empirical evaluation, which shows that Rexplore is highly effective at providing support for the aforementioned sensemaking tasks. In addition, these results are robust both with respect to the background of the users (i.e., expert analysts vs. 'ordinary' users) and also with respect to whether the tasks are selected by the evaluators or proposed by the users themselves.

2 Exploring Scholarly Data

2.1 State of the art

A large variety of systems support the exploration of scholarly data, some of them providing an interface to a specific repository of bibliographic data, others integrating multiple data sources to provide access to a richer set of data and/or to provide a richer set of functionalities. The most widely used academic search engine is probably Google Scholar (http://scholar.google.com), which primarily supports search and citation services, providing comprehensive access to the academic literature. DBLP (http://www.informatik.uni-trier.de/~ley/db/) is a well-known computer science bibliography website and can be browsed using FacetedDBLP [3], an interface which exploits the faceted search paradigm to support data exploration. CiteSeer[X] [4] focuses instead on large-scale harvesting and indexing of research papers and includes mechanisms for suggesting relevant papers. These systems mainly focus on providing a good interface for publication search and are not designed to support sensemaking tasks in the academic domain. On the contrary, Microsoft Academic Search (http://academic.research.microsoft.com/) provides a variety of visualizations, including co-authorship graphs, publication trends, and co-authorship paths between authors. In a similar way Arnetminer [5] also offers different visualizations and provides support for expert search and trend analysis. Saffron [6], which builds on the Semantic Web Dog Food Corpus [7], exploits keywords for expert search and estimates the strength of an author/topic relationship by analyzing co-occurrences on the Web. A common aspect of these systems is that they use keywords extracted from publications as proxies for research topics. However these are noisy and lack structure

(see Section 2.2.1 for a detailed discussion on this aspect).

Recently, reference management tools have emerged, such as Zotero (http://www.zotero.org), EndNote (http://endnote.com) and Mendeley (http://www.mendeley.com), as well as specialized social networks sites for researchers –e.g., ResearchGate (http://www.researchgate.net) and Accademia.edu (http://www.academia.edu). However, while these systems support exploration to some degree, again they only provide limited support for sensemaking tasks.

A key challenge for a system exploring scholarly data is how to assist users in searching and navigating through a variety of different dimensions –e.g., topic, organization, co-author, etc. A popular paradigm is *faceted browsing* [8], in which a set of objects can be filtered progressively along several dimensions in different orders. The \facet tool [9] exploits this idea to allow for an easier exploration of heterogeneous Semantic Web repositories by using the different resources found in RDF repositories as alternative facets. mSpace [10] tackles the problem of dealing with high-dimensional spaces, by showing a subset of the data at the time, called "a slice", and arranging them in a hierarchy of columns in accordance with user-defined priorities. Other approaches rely on the *pivot* (or multi-pivot) paradigm [11], which allows users to identify key elements in the data space (the pivots), and use these to introduce structure and facilitate the navigation process. For example, PaperCUBE [12] offers advanced data visualization functionalities and it specifically focuses on scholarly data, providing effective visual modalities to browse citation networks and relations between authors and to situate a paper in a research context. However, the focus here is primarily on individual publications and little support is provided for higher-level tasks, such as understanding research dynamics and fine-grained expert search.

2.2 Gap Analysis

As we have seen in the previous section, the space of solutions for exploring scholarly data is large, comprising both powerful systems for crawling and indexing scholarly data, such as Google Scholar, as well as a variety of visualization solutions and data exploration paradigms, some generic in nature, others specifically customized for scholarly data. However, despite the availability of such a variety of systems, exploring scholarly data remains challenging, especially once we move away from basic search (for authors or publications) and we aim to capture the dynamic elements to do with research trends and relationships between authors (which go beyond citation and collaboration), or we aim to perform expert search at a very fine-grained level –e.g., by searching for researchers with expertise in multiple topics, at a certain career stage, within a certain geographical area, who have a track record of publishing in the top conferences associated with one or multiple research areas, etc. In what follows we will discuss these issues in more detail, highlighting the key gaps that Rexplore aims to address.

2.2.1 No semantic characterization of research areas

A key precondition for an effective exploration of scholarly data concerns the mapping of people and publications to the relevant research areas. However, 'research area' is rarely treated as a first class concept and instead systems tend to use keywords as proxies for research areas. This limitation creates a number of problems. For

instance, the Arnetminer page for Enrico Motta includes "International Semantic Web Conference" as a research interest, even though research interests should arguably concern topics[1], rather than conferences. A similar problem can be seen by looking at the Microsoft Academic Search (MAS) page for Enrico Motta, which lists three high level 'fields' for him, "Database", "Web", and "Artificial Intelligence", and then supplements this information with a number of keywords, including "Case Study", which (again) is arguably not a research area.

Another problem stemming from a syntactic, rather than semantic, treatment of research areas is that systems do not take into account important semantic relations between research areas, such as an area being a sub-area of another one, or two labels referring to the same research area. This problem has been traditionally addressed by relying on manually curated taxonomies, such as the ACM classification (http://www.acm.org/about/class/). However these classifications suffer from several problems. First of all, they are very shallow –for example the entry "Intelligent Web Services and Semantic Web" in the ACM classification only contains four sub-topics, thus failing to reflect the variety of topics being tackled by the Semantic Web research community. In addition, because they are manually curated, they evolve very slowly and as a result, they fail to reflect the latest research trends. Finally, they are actually very opaque, as it is not clear what does it mean for a topic to be classified under another topic. For instance, "Ontology Languages" is classified under "Intelligent Web Services and Semantic Web"; however one could argue that it is strange to say that the former is a sub-topic of the latter, given that ontology languages were being designed well before the Semantic Web was recognized as a research area. In addition, these classifications do not cater for situations where there are different ways to refer to the same area. For instance, most people would agree that the labels "Ontology Matching" and "Ontology Alignment" refer to the same area of research.

2.2.2 Lack of granular analysis

Systems such as MAS provide ways to visualize research trends. However, these are considered at a very high-level of abstraction. For example, MAS can visualize publication trends in "World-Wide-Web" and "Databases", but cannot provide this feature for "Semantic Web", let alone more fine-grained topics, such as "Semantic Web Services". However, both researchers and students tend to be interested in rather fine-grained trends – e.g., what's happening with Linked Data, rather than what's happening with the Web. A wider range of topics is provided by Arnetminer, however these still cover only a subset of the research topics (e.g., key topics for the Semantic Web community, such as "Linked Data" and "Ontology Evolution" are not included) and in addition they are provided as a flat list, rather than in a structured, easily navigable form.

2.2.3 Digital library bias

Another limitation of most existing systems in the context of the sensemaking tasks that we wish to support is the emphasis on classic digital library functionalities, such as supporting search for publications and providing citation services. While of course

---

[1] In what follows, we will use the terms 'topic' and 'research area' interchangeably.

these are key functionalities and essential building blocks for more advanced services, they do not necessarily provide the right level of support when the goal is to make sense of what goes on in a research area, rather than to identify a specific paper. For instance, in the example given in Section 1, where a research area needs to be investigated in the context of making a decision about a special issue proposal, what is needed from a system is the ability to support the user in identifying quickly the important trends in the area –such as, whether it is growing (and in this case where are the new researchers coming from) or shrinking (and in this case where are the researchers migrating to), rather than following citation links or locating a specific paper. Another negative side-effect of this 'bias' is the aforementioned problem highlighted by Dunne et al. [1], concerning the lack of an integrated environment, supporting a seamless exploration of the space of scholarly data, as opposed to providing 'atomic' functionalities, to do with static visualizations or search and citation services, which is the situation with most current systems.

3 Overview of Rexplore

The goal of Rexplore is to provide an environment capable of overcoming the limitations discussed in the previous section to support users effectively by enabling them i) to detect and make sense of the important trends in one or more research areas, ii) to identify researchers and analyze their academic trajectory and performance in one or multiple areas, according to a variety of fine-grained requirements, iii) to discover and explore a variety of dynamic relations between researchers, between topics, and between researchers and topics, and iv) to support ranking of specific sets of authors, generated through multi-dimensional filters, according to various metrics.

Rexplore addresses the problem of the lack of a semantic characterization of research areas by introducing a fine-grained, automatically populated topic ontology, in which topics are identified and structured according to a number of semantic relationships [13]. The resulting knowledge base is generated using a combination of statistical methods and background knowledge on the basis of a large-scale corpus of publications (Section 3.1) and is then augmented with geographic information (Section 3.2). Research topics can then be browsed and analyzed by means of a variety of visual analytics solutions, which exploit the rich set of relations in the data, and in particular the fine-grained characterization of research areas (Section 3.3). Authors can be investigated by plotting a number of metrics on a timeline, and their associated research areas can be analyzed at different levels of abstraction (Section 3.4). Powerful query/search facilities are also provided, supporting complex multi-dimensional queries that can include logical connectives (Section 3.5). Finally, Rexplore also takes advantage of the fine-grained semantic characterization of authors and topics, to introduce novel relationships between authors, which go beyond co-authorship and focus on their similarity with respect to 'semantic' features, such as research interests and academic trajectories (Section 3.6).

As shown in Figure 1, Rexplore integrates a variety of data sources in different formats, including: DBpedia [14], DBLP++ (http://dblp.l3s.de/dblp.rdf.gz), the MAS API, GeoNames (http://www.geonames.org/), and parsed web pages (e.g., Wikipedia, Google Scholar). The publication and author metadata used in the current version of the system, Rexplore v2.0, come mainly from MAS and DBLP++. The process of

generating the populated topic ontology exploits information collected from Google Scholar, EventSeer (http://eventseer.net/) and Wikipedia. The geographic information and the standardization of the affiliations rely on information from DBpedia and GeoNames (see Section 3.1). As of April 2013, Rexplore integrates metadata on 20 million papers and 2 million authors. The back-end of Rexplore is implemented in PHP, while the interface and the visualizations are in HTML5 and JavaScript. The interface uses the Highcharts (http://www.highcharts.com/) library, and also builds on a heavily modified version of Jit (http://philogb.github.io/jit/).
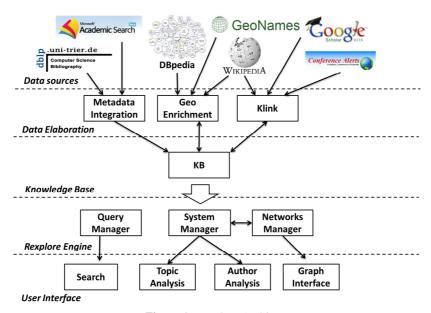

Figure 1. Rexplore Architecture.

3.1 Ontology population with Klink

Rexplore does not consider topics as simple keywords, but relies on an OWL ontology, which characterizes research areas and their relationships. This ontology is automatically populated and periodically updated by the Klink algorithm [13], which takes as input a corpus of publications, annotated with keywords (these can be user-defined or automatically extracted from the text of a publication) and performs three key operations, using a combination of statistical methods and background knowledge:

- It identifies research areas from the given set of keywords, tidying them up by fixing errors and by removing keywords that do not denote research areas – e.g., "Case Study" or "NeOn Project".
- It automatically computes three types of semantic relationships between research areas – see below for more details.
- It returns a knowledge base of semantic relationships expressed in OWL.

In particular, Klink computes the following three relationships between topics:

- *skos:broaderGeneric*. This is used to indicate that a topic, say T_1, is a sub-topic of another topic, say T_2. For instance, "Semantic Web Services" can be

characterized as a sub-topic of both "Semantic Web" and "Web Services".

- *contributesTo.* This is defined as a sub-property of *skos:related* and it is used to characterize relations where there is evidence (gathered through statistical methods and/or background knowledge) that research in topic T_1 is seen as an important contribution to research in topic T_2, but it would be incorrect to say that T_1 is a sub-topic of T_2. An example is the relation between "Ontology Engineering" and "Semantic Web", where there is significant evidence that results from the former are relevant to the latter, but it would be incorrect to say that "Ontology Engineering" is a sub-topic of "Semantic Web", given that it is a much older research area than "Semantic Web" and, even today, there is a lot of work in Ontology Engineering, which is carried out independently of Semantic Web research.
- *relatedEquivalent.* This is also defined as a sub-property of *skos:related* and it is used to indicate that two keywords, e.g., "Ontology Matching" and "Ontology Alignment" are simply different labels for the same research area[2].

Our ontology[3] builds on the BIBO ontology, which in turn builds on SKOS,[4] FOAF,[5] and other standards. Our extensions are very conservative and comprise only the *relatedEquivalent* and *contributesTo* object properties described earlier, and the class *Topic*, which is used to refer to research topics. The resulting OWL knowledge base is exploited to support knowledge-based exploration, pattern extraction and author clustering in Rexplore. Currently it comprises 1500 topics linked by almost 3000 semantic relationships. A detailed description of Klink, including an empirical evaluation of the algorithm can be found in [13].

3.2 Geographic Enrichment

The data sources used by Rexplore offer in most cases only the name of the author's affiliation (e.g., Universities, Research Labs, Hospitals), which is usually derived from parsing research papers and thus it is simply treated as a string. As a result, affiliations may in some cases lack the actual geographical location or may use different ways to refer to the same institution –e.g., "University of Turin" and "University of Torino". Since a correct affiliation linked to the correct geographic location provides valuable information for filtering and exploring authors, we use a simple but effective geographic enrichment procedure which i) defines a standard name for each affiliation, avoiding duplications, and ii) maps the affiliation to

---

[2] Here we could have used *Owl:sameAs*, given that Rexplore functionally treats two *relatedEquivalent* topics as being the same one. However, from an epistemological point of view, it can be argued that this would be too strong a commitment and that in other scenarios one may want to consider topics with different names as different ones. Hence, to avoid overcommitting our ontology, we have introduced the *relatedEquivalent* property.

[3] http://kmi.open.ac.uk/technologies/rexplore/ontologies/BiboExtension.owl.

[4] The most recent specification of the SKOS model, which can be found at http://www.w3.org/TR/2009/REC-skos-reference-20090818/, proposes a new property, *skos:broaderTransitive,* to support the representation of transitive hierarchical relations. However, our ontology currently sticks to the older SKOS specification, primarily because it builds on the BIBO ontology, which in turn builds on the 2004 SKOS model.

[5] http://xmlns.com/foaf/spec/.

GeoNames, a well-known geographic database. The procedure uses initially Wikipedia to retrieve a 'standard' identifier for the affiliation and then searches for the location associated with the affiliation in DBpedia. If the latter search is unsuccessful, then the Wikipedia page is parsed for the tag "location" from which city and country are extracted using a set of heuristic rules. After recovering information about the city or the country, the affiliation is mapped to the correct GeoNames ID. If the search for affiliation and/or location in Wikipedia/DBpedia fails, then the affiliation name is stripped of a set of typical terms, such as "university", "college" or "hospital", and the remaining string is searched for in the GeoNames database. This simple method provided good results, allowing us to correctly map disambiguated affiliations to GeoNames in about 85% of the cases.

3.3 Topic Analysis

Rexplore takes advantage of the Klink-generated OWL knowledge base by considering 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 dimension 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 double (22,143).

Figure 2. Exploring the topic "Semantic Web" in Rexplore.

For analyzing a topic, Rexplore provides an interface that includes: i) general information about the topic, ii) access to the relevant authors and publications, iii) the *topic navigator*, iv) visual analytics on *broaderGeneric* and *contributesTo* sub-topics, and v) visual analytics on authors' migration patterns from other topics to and from the topic in question. As an example, Figure 2 shows the page for the topic "Semantic Web", which (on the left) includes basic statistics, access to basic functionalities, and the topic navigator showing the relevant fragment of the topic hierarchy generated by Klink. On the right hand side of the figure, a histogram is shown, as the user has selected to visualize the publication trends for research in Linked Data, OWL and RDF. In particular, Figure 2 shows that the Linked Data area has exploded in the past

few years, while research in OWL appears to have reached a plateau.

Rexplore is able to visualize different topic trends: 1) publication trends, 2) author trends and 3) migration trends. The first two are the number of publications or authors associated with a semantically enriched topic on a timeline. The latter is defined as the number of estimated migrations between two topics and is computed by analyzing the shifting in authors' interest, as described in [15].

3.4 Author Analysis

Every author in Rexplore has a personal page which includes i) general bio information, ii) author's scores according to different bibliometric measures, iii) *topic analysis*, iv) *co-author analysis*, v) *pattern analysis*, and vi) *graph view*. The page offers the possibility of deploying more than 20 different charts to plot each metric as a function of time. The *topic analysis* makes it possible to browse and plot on a timeline the main research areas in which the author has published or was cited. The topics and sub-topics are displayed in a multilevel list in such a way that it is possible to choose the granularity level. For example it is possible to conduct a high level analysis by focusing on the main topics (e.g., "Semantic Web" or "Artificial Intelligence") or otherwise to zoom in one of them (e.g., "Semantic Web") and further analyze its sub-topics in details, exploiting the semantic structure generated by Klink. The *co-author analysis* section ranks the co-authors according to the number of publications or citations they have in common. It is also possible to select a number of co-authors and visualize their collaboration with the author in question by year and by topic. The *pattern analysis* section groups authors with a similar publications/citations pattern and can be also used to forecast future publication activity and impact for an author (in particular one at a reasonably early career stage). The *graph view* will be discussed in detail in Section 3.6.

3.5 Faceted Search and Data Browsing

Rexplore offers a number of facets to be used both for the formulations of complex search queries and for context-based data navigation and analysis. Indeed, both the topic and author analysis interfaces offer the possibility of focusing on specific combinations of facets, in order to allow the users to navigate/retrieve data according to specific dimensions. For example, authors can be filtered by 1) name or a part of it, 2) career range (that is the time from the first published work), 3) topics of interest and 4) venues in which they published. Both venue and topic fields accept multiple values, which can be combined using logical connectives. Hence it is easy to formulate complex queries, e.g., to retrieve career-young authors, who have worked in both "Semantic Web" and "Social Networks", and have published in ISWC.

The results can be ranked by a variety of metrics that, for author-centric searches, include: 1) number of publications, 2) number of citations, 3) H-Index, 4) G-Index, 5) HT-Index, 6) GT-Index, 7) number of publications/citations in a topic or set of topics, 8) number of publications/citations in a venue or set of venues. Here it is worth to highlight that the fine-grained structure of research topics generated by Klink supports the definition of fine-grained impact metrics, such as "citations in topics", which allow to measure very specific elements of academic impact.

HT-Index and GT-Index are based on the standard G-Index and H-Index, however they are normalized by the number of average citations in each topic. Hence they are

useful for comparing authors who publish in fields with different levels of field-specific impact.

Often users want to start the data exploration process from the query results, for example by analyzing each one of a number of authors. Rexplore assists this seamless navigation by remembering the specified search filters –e.g., when switching from a list of results to a graph view.

3.6 The Graph View

The *graph view* is a novel, highly interactive tool to explore the space of authors and their relationships using faceted filters. It takes as input one or multiple authors and displays their relations allowing the user to choose among a variety of types of links, ranking criteria, and filters. As an example, in Figure 3 we show the graph view displaying the authors most similar to Enrico Motta according to the temporal topic similarity, a novel metric which reflects the similarity of people's research trajectories with respect to the temporal sequencing of their research interests. The radius of the nodes in the graph reflects the number of his/her publications in the Semantic Web area. Other author ranking measures, as discussed in the previous section, can also be used. Users can also choose from six types of relations between authors: co-publication, co-citation, topic similarity, temporal topic similarity, publication pattern similarity and citation pattern similarity.
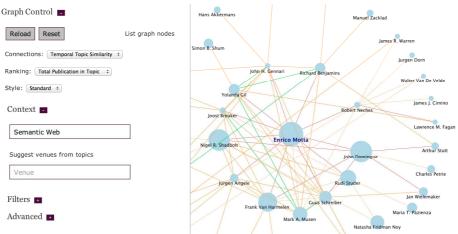

Figure 3. A graph view in Rexplore.

The topic similarity reflects how similar two authors are with respect to their research topics and takes advantage of the fine-grained topic structure generated by Klink and its semantic characterization. A naïve way to compute it would be to directly compare the vectors representing the number of publications associated with a keyword. However, treating topics as strings, as many systems do, would yield poor result. In fact, keywords referring to a related area, to a sub area, or even indicating the same topic with a different name would be considered different. For example, a prominent author in the field of Linked Data would have most of his or her publications associated to a "Linked Data" keyword, and may be considered uncorrelated to authors who have papers tagged as "Semantic Web". Thus we exploit

a variation of the semantic enrichment procedure already mentioned in Section 3.3 on the publication vectors of the authors, assigning each publication on a topic also to its *broaderGeneric* or *relatedEquivalent* topics. However, in this case we want also to include *contributesTo* relationships, which yield important information but cannot be handled in the same way. In fact, it is not automatic that a paper published under a topic (e.g., "Ontology Engineering") is also about its *contributesTo* topics (e.g., "Semantic Web"). It seems however appropriate to use the probability that the contributing topic T_1 refers to a certain topic T_2 to assign an additional bonus to T_2. Thus, in case of a *contributesTo(T_1,T_2)* relationship, we assign to T_2 only a fraction of the publications in T_1 according to the formula:

$$CT(T) = \sum_{i=1}^{n} P\big(T \big| ct(i,T)\big)^{\alpha}$$

where T is a topic, n is the number of publications of an author that are not already associated with T but have at least one topic in a *contributesTo* relationship with T, $ct(i,T)$ is the set of topics associated with the *i-th* publication that are in a *contributesTo* relationship with T, $P\big(T\big|ct(i,T)\big)$ is the probability for a paper with the set of topics $ct(i,T)$ to be also explicitly associated with area T (or with a topic having a *broaderGeneric* or *relatedEquivalent* relationship with T) before the publication date of the *i-th* paper and α is a factor which modulates the *contributesTo* relationship (empirically set to 0.5 in the prototype).

By taking into account the publication date of each paper, the formula considers also the changes in topic relations over time. For example a paper about "Ontology Engineering" in the year 2001 would have a lower probability to be about "Semantic Web" than a paper about the same topic in 2010 and thus should contribute much less to "Semantic Web" in the author publication vector. The topic similarity is finally computed as the cosine similarity between the semantically enriched vectors of publications.

The temporal topic similarity builds on the topic similarity measure and makes it possible to identify groups of researchers who appear to be following similar research trajectories, sharing research interests and moving from one topic to another in a similar way. In particular, this is very useful to identify the various sub-communities that populate a particular research area. The temporal topic similarity takes into account the order and the time span in which an author has published on a certain topic and is calculated as the weighted average of the topic similarities computed on different time intervals. Thus, if author A worked on T_1 and then moved to T_2, he or she may be similar to author B who was originally in T_2 and then moved to T_1 in terms of topic similarity, but will be different in terms of temporal topic similarity.
Finally, the publication/citation pattern similarity reflects how similar two authors are with respect to their career progression in terms of number of publications/citations.

The graph view also provides a variety of standard interface operations, such as changing the level of granularity in the view, expanding, closing, or hiding nodes, etc. In addition, both nodes and links can be filtered with respect to specific years, topics, and venues. For example, it is possible to customize a graph and visualize only the co-authors of a particular researcher, who have between 5 and 15 career years, have published in both Linked Data and Social Networks, and have publications in CHI.

4 Empirical Evaluation

4.1 Experimental Setup

For the evaluation we enrolled 17 PhD students and researchers drawn from the members of the Knowledge Media Institute in UK and the University of Turin in Italy. None of these subjects had been involved in the work on Rexplore, or indeed knew the system prior to the evaluation session. At the beginning of the evaluation session, every subject filled a questionnaire about his/her research experience, topics of interest, and familiarity with a list of systems that included Google Scholar (GS), MAS, DBLP, and Citeseerx. This was followed by a 15-minutes tutorial about Rexplore and then the subjects were asked to perform the activities listed in Table 1.

Activity 1. Carry out the tasks shown in Table 2 using Rexplore.
Activity 2. Select one of the three tasks in Table 2 and attempt to achieve it using either Google Scholar (GS) or Microsoft Academic Search (MAS).
Activity 3. Suggest a task you would consider valuable and perform it using Rexplore.

Table 1. Activities in the Evaluation Process.

The rationale for selecting GS and MAS as control systems was that GS is the most widely used bibliographic search engine, while MAS provides a number of features, in terms of time-based visualizations, which go well beyond what is provided by GS.

Warm-up Task. Find the 3 main co-authors (in any field) of the author with most publications in the topic User Modelling.
Task 1. Find the top 3 'rising stars' in the United Kingdom with expertise in both *Semantic Web* and *Social Networks*, in the career range 5-15 years from first publication, ranked in terms of number of citations in these 2 areas.
Task 2. Find the top 5 authors with the highest number of publications in the *Semantic Web* and rank them in terms of number of publications in *Artificial Intelligence*. For each of them find their most cited paper in *Artificial Intelligence*.
Task 3. Which are the 2 sub-topics in *Semantic Web* that have grown the most in 2005-2010 (as measured by the difference between the number of papers in 2010 and in 2005) and who are the top 2 authors (ranked by number of publications in topic) in these 2 topics.

Table 2. Evaluation Tasks.

Each task was recorded with screen capturing software and the time taken for completion was measured; if a task was not solved within 15 minutes, it was recorded as 'failed'. Tasks not completed within the time limit were considered as 15 minutes performance. After completing the various tasks, the 17 participants were requested to fill in a usability SUS questionnaire [16] and a second questionnaire about the strengths/weaknesses of the tested systems. On average the total time required to complete each evaluation session was slightly less than 2 hours.

In contrast with other evaluation studies –e.g., see [17], where participants were divided in different groups and each group would use a different tool to perform the same set of tasks, here we did not carry out a straightforward 'tool shootout', but we instead implemented a more faceted experimental design, comprising usability questionnaires and a task-centric evaluation, and also providing the participants with the opportunity both to suggest their own tasks and also to try out other tools. The

reason for this is that GS and MAS do not directly support the kinds of sensemaking tasks for which Rexplore offers support, hence a 'tool shootout' would have provided little valuable data and most likely caused a high degree of frustration for the subjects. For this reason we decided to focus the bulk of the evaluation on identifying opportunities to evaluate and gather feedback on Rexplore, while still collecting some comparative data.

The tasks given to the subjects cover common scenarios to do with expert search and trend detection. Task 1 is a common expert search task –e.g., for research leaders who wish to identify 'new blood' to fill a certain position. Task 2 is also a common expert search task, where, given a pool of people with expertise in topic A, we want to identify the person in the pool that can be considered as the top expert in topic B. Task 3 is about detecting trends and analyzing research topics. It is a common task for many professionals, such as managers in research funding bodies, who may wish to identify which areas appear to be particularly 'hot' within a broader research field.

4.2 Results

In Activity 1, the 17 subjects were able to complete within the requested 15 minutes 50 of the 51 (17*3) tasks using Rexplore, with a 98% success rate. The only failure was registered in Task 2. Task 1 was the simplest one and was performed on average in about 3 minutes. In fact this task required only the ability to formulate a complex query, followed by the manual identification of a number of authors. Task 2 required a more complex exploration of the system, since the user had to first select five authors and then explore them using the graph view or the author analysis page, to find out their contributions in Artificial Intelligence. Task 3 required the use of visual charts showing the publication trends of the sub-topics and the use of the topic navigator to identify the best authors.

	Rexplore (N=17)			MAS (N=9)		
	Task 1	Task 2	Task 3	Task 1 (N=6)	Task 2 (N=2)	Task 3 (N=1)
Average Time	3:06	8:01	7:51	14:46	13:52	15:00
Standard Dev.	0:45	2:50	2:32	0:24	1:35	00:00
Success Rate	100%	94%	100%	33%	50%	0%

Table 3. Experimental results (in min:secs) using Rexplore and MAS. The tasks performed with GS yielded no success, thus their average time is by definition equal to 15:00.

In Activity 2, eight subjects were asked to work with GS and nine with MAS. Task 1 was chosen by 6 users on MAS and 5 on GS, while Task 2 was chosen respectively by 2 and 3 subjects. Task 3 was perceived by the subjects as practically impossible to do with a system without a fine-grained topic analysis functionality, and as a result was tried only by one subject (using MAS). Only three people out of nine completed a task with MAS (overall 33% success rate) and none at all with Google Scholar. Hence, the success rates of the three systems are significantly different: the two by two table comparison between Rexplore and MAS analyzed with a Fisher test (a standard statistical significance test used in the analysis of contingency tables when the numbers involved are small) yields $p=10^{-5}$, whereas the three by two table including also GS yields $p <10^{-7}$. Incidentally, the users who were able to complete the chosen task on MAS (2/8 for Task 1 and 1/2 for Task 2) were among the best

performers in Activity 1, and required for the same task about 5 times longer on MAS than on Rexplore, even after having already successfully completed the task in Rexplore. Table 3 summarizes the time employed for the assigned tasks on Rexplore and MAS and the relative success rate –i.e., the number of jobs completed correctly within 15 minutes.

An important question when using a tool for navigating a research area is how much prior knowledge of the domain affects task performance. The results of the evaluation show that the average time for completing the three tasks by subjects with expertise in Semantic Web (that is the main area of the tasks) is not significantly different from the one obtained by the others ($p=0.63$ according to the t-test). However, the experts in tools for exploring academic data, who are active in fields such as Bibliometric and Learning Analytics, were instead able to get acquainted with the Rexplore system much more quickly and use it more effectively than the other subjects. The average time of the former group on the three tests was $5:01\pm0:02$ min, against $6:52 \pm 0:06$ min of the latter ($p < 0.022$). On the contrary, no correlation was found with the usage of other tools for academic exploration, such as GS, MAS, DBLP, ACM, Citeseerx and Scopus. Hence the data appears to show that no domain-specific expertise is needed to use Rexplore to make sense of a particular research area, while at the same time the tool does not penalize experts in Bibliometrics and Learning Analytics, who are used to carrying out these kinds of analyses.

The tasks proposed by the subjects in Activity 3 were a good mix of routine searches and creative queries, and thus the performances cannot be directly compared. 59% of the subjects chose to investigate a single author, using mainly the topic analysis and the graph view, whereas 23% of them preferred to explore a research area to understand better its migration patterns and trends. The ability to filter by multiple topics or using them for author analysis was widely appreciated: 71% of the proposed tasks involved topic filtering or topic analysis on an author or group of authors. The integration of the different Rexplore functions made it possible to try particularly interesting exploration tasks: for example, a particularly creative subject tried to find a better affiliation for an author by analyzing organizations, topic similarity and prominence of the researchers connected to him through the various links provided by Rexplore (incidentally, he opted for MIT).

15 out of the 17 subjects considered their suggested task satisfactorily concluded. One of them was unable to complete her proposed task because of problems with the original data tagging: the subject was searching for papers of a certain author about Semantic Web, which were actually tagged only as "Knowledge Base" in the original data which Rexplore uses. This suggests that relying exclusively on user-defined keywords may not be sufficient and even when these are available, it may be useful to refine them by analyzing the abstract or the full text of the paper.

Rexplore reached a score of 75/100 on the standard SUS usability test, based on ten multiple-choice questions. A score of 75 can be converted to a percentile rank of 72%, meaning that the usability of Rexplore was considered equal or superior than 72% of the 500 tested systems. In particular 94% of the subjects agreed or strongly agreed on the fact that the functions of the system are well integrated and 82% stated that they would be happy to use Rexplore for their work.

The post-task questionnaire included three sections. In the first and second parts the users were asked for their opinions about the support given by Rexplore and GS/MAS for the assigned tasks. In the third part they were asked to comment about

the support provided by Rexplore for the task suggested by them.

In the first section, 94% of the subjects described Rexplore as "very effective", while 18% described it as "easy/natural/intuitive". Among the most useful features were the faceted filters (59%), the visualization/charts (47%), the graph view (47%) and the semantic characterization of topics (41%). The main weaknesses of Rexplore were found to be its visual complexity (41%) and a not always well-evidenced navigation context (35%). Indeed, according to some users, the high number of functionalities offered by Rexplore may also be overwhelming.

When asked to suggest new features that would facilitate their exploration of academic data, 23% of subjects suggested some "minor interface change", especially in the direction of solving the aforementioned problem of "making the context clearer in any moment" (18%). 23% of them thought Rexplore did not need any additional features and 18% proposed additional filters. Other features that the users suggested include a natural language interface for formulating complex searches and the ability to retrieve and search the full text of a publication from within Rexplore.

Trying to perform the kind of task described in Table 2 with MAS or GS frustrated the users: 88% of the subjects using MAS and 89% using GS described the support of those systems as "ineffective". The reasons of their frustration were various: not effective contextual filtering (77% MAS, 65% GS), absence of semantic/structured topics (56% MAS, 63% GS), and poor support for complex/multidimensional queries (33% MAS, 50% GS). Finally, the support provided by Rexplore for user-defined tasks (Activity 3) was also rated positively. 76% of the participants defined such support as "effective/very effective/unique", while 12% of them, though they were able to complete their task, found some "minor problems", usually to do with missing filter options. Indeed it seems that users could do with a variety of filters well beyond what it is normally considered in these systems: one of the subjects suggested a filter able to discriminate genders, while another asked to be able to split publications by the particular author position (e.g., first or second author). Nonetheless the results of the evaluation appear very satisfactory, confirming that Rexplore provide a degree of support that users consider effective and valuable for performing real-world tasks.

5 Conclusions

In this paper we have presented Rexplore, a novel tool for exploring scholarly data, which integrates a semantic foundation with statistical and visual analytics solutions to support users in exploring and making sense of scholarly data. The results from the empirical evaluation confirm the effectiveness of the functionalities provided by the tool and show a high value of user satisfaction. In particular, users rate very highly the semantic underpinning of the tool, which arguably affords a major advantage over other tools in its ability to support i) the visualization of trends at a very fine level of granularity, ii) methods to identify 'semantic' relations between authors, and iii) fine-grained multi-dimensional academic expert search.

For the future we plan to extend the tool by enhancing its functionalities through the integration of other sources of data relevant to academic activities and we also aim to address the minor interface issues identified during the evaluation. We also plan to add to the number of navigation filters, a feature which users appear to value extremely high. Finally, we are actively discussing with a number of commercial providers of scholarly data, with the aim to release a version of the tool with

comprehensive data coverage for use by the scientific community.

References

1. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., Dorr, B.: Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. American Society for Inf. Science and Technology, 63(12), 2351-2369. (2012)
2. Motta, E., Osborne, F.: Making Sense of Research with Rexplore. 11th Int. Semantic Web Conference, Poster&Demo Session, Boston, MA. (2012)
3. Diederich, J., Balke, W. T., Thaden, U.: Demonstrating the Semantic Growbag: Automatically Creating Topic Facets for FacetedDBLP. In: Proceeding of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. (2007)
4. Li, H., Councill, I., Lee, W.C., Giles, C.L.: CiteSeerx: an architecture and web service design for an academic document search engine. In: Proceedings of the 15th Int. Conference on the World Wide Web, pp. 883-884. (2006)
5. 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 Int. Conference on Knowledge Discovery and Data Mining, pp. 990-998. (2008)
6. Monaghan, F., Bordea, G., Samp, K., Buitelaar, P.: Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food. In: Semantic Web Challenge at the International Semantic Web Conference. (2010)
7. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for semantic web dog food—The ESWC and ISWC metadata projects. In: The Semantic Web (pp. 802-815). Springer Berlin Heidelberg. (2007)
8. Yee, P., Swearingen, K., Li, K., Hearst, M..: Faceted Metadata for Image Search and Browsing. In: Proceedings of CHI 2003, pp. 401-408, ACM. (2003)
9. Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A browser for heterogeneous semantic web repositories. In: Proceedings of the 5th Int. Semantic Web Conference (pp. 272-285). Springer Berlin Heidelberg. (2006)
10. Schraefel, M.C., Wilson, M., Russell, A., Smith, D.A.: mSpace: improving information access to multimedia domains with multimodal exploratory search. Communications of the ACM, 49(4), 47-49. (2006)
11. Popov, I.O., Schraefel, M.C., Hall, W., Shadbolt, N.: Connecting the dots: a multi-pivot approach to data exploration. 10th Int. Semantic Web Conference (pp. 553-568). Springer Berlin Heidelberg. (2011)
12. Bergstrom, P., Atkinson, D.C.: Augmenting the exploration of digital libraries with web-based visualizations. ICDIM 2009. IEEE. (2009)
13. Osborne, F., Motta, E.: Mining semantic relations between research areas. 11th Int. Semantic Web Conference, pp. 410-426, Springer-Verlag. (2012)
14. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., et al.: DBpedia-A crystallization point for the Web of Data. Journal of Web Semantics, 7(3), 154-165. (2009)
15. Osborne, F., Motta, E.: Exploring Research Trends with Rexplore. D-Lib Magazine, 19 (9/10). (2013)
16. Brooke, J.: SUS: A "quick and dirty" usability scale. Jordan, P. W. et al. (eds), Usability Evaluation in Industry pp. 189-194. Taylor & Francis, London, UK. (1996)
17. Motta, E., Peroni, S., Gómez-Pérez, J. M., d'Aquin, M., Li, N.: Visualizing and Navigating Ontologies with KC-Viz. In: Proceedings of the 10th Int. Semantic Web Conference, pp. 343-362. Springer-Verlag. (2011)