Semantic User Profiles: Learning Scholars' Competences by Analyzing their Publications

Bahar Sateli[1], Felicitas Löffler[2], Birgitta König-Ries[2], and René Witte[1]

[1] Semantic Software Lab, Department of Computer Science
and Software Engineering, Concordia University, Montréal, Canada
[2] Department of Mathematics and Computer Science
Friedrich Schiller University Jena, Germany

Abstract. Semantic publishing generally targets the enhancement of scientific artifacts, such as articles and datasets, with semantic metadata. However, smarter scholarly applications also require a better model of their *users*, in order to understand their interests, tasks, and competences. These are generally captured in so-called *user profiles*. We investigate a number of existing linked open data (LOD) vocabularies and propose a description of scientists' competences in LOD format. To avoid the *cold start* problem, we suggest to automatically populate these profiles based on the publications (co-)authored by users, which we hypothesize reflect their research competences. Towards this end, we developed the first complete, automated workflow for generating semantic user profiles by analyzing full-text research articles through natural language processing. We evaluated our system with a user study on ten researchers from two different groups, resulting in mean average precision (MAP) of up to 92%. We also analyze the impact of semantic zoning of research articles on the accuracy of the resulting profiles. Finally, we demonstrate how these semantic user profiles can be applied in a number of use cases, including article ranking for personalized search and finding scientists competent in a topic – e.g., to find reviewers for a paper.

1 Introduction

Researchers increasingly leverage intelligent information systems for managing their research objects, like datasets, publications, or projects. An ongoing challenge is the overload scientists face when trying to identify potentially relevant information, e.g., through a web-based search engine: While it is easy to find numerous potentially relevant results, evaluating each of these is still performed manually and thus very time-consuming.

We argue that smarter scholarly applications require not just a semantically rich representation of research objects, but also of their users: By understanding a scientist's interests, competences, projects and tasks, intelligent systems can deliver improved results, e.g., by filtering and ranking results through personalization algorithms [26].

So-called *user profiles* [11,15] have been adopted in domains like e-learning [5], but so far received less attention in scientific applications (we provide a brief background on user profiling in Section 2). We believe that a semantically rich representation of users is important for enabling a number of advanced use cases in scholarly applications. We argue that a new generation of *semantic user profile* models are ideally built on standard

semantic web technologies, as these make them accessible in an open format to multiple applications that require deeper knowledge of a user's competences and interests. In Section 3, we analyze a number of existing Linked Open Data (LOD) [13] vocabularies for describing scholars' preferences and competences. However, they all fall short when it comes to modeling a user's varying degrees of competence in different research topics across different projects. We describe our solution for scholarly user models in Section 4.

Bootstrapping such a user profile is an infamous issue in recommendation approaches, known as the *cold start* problem, as asking users to manually create possibly hundreds of entries for their profile is not realistic in practice. Our goal is to be able to create an accurate profile of a scientist's *competences*, which we hypothesize can be automatically calculated based on the publications of the user. Towards this end, we developed the first text mining pipeline that analyzes full-text research articles for an author's competences and exports the results in linked data format into a user profile. The design and implementation of our approach are detailed in Sections 4 and 5, respectively.

To evaluate our profile generation approach, we performed a user study with ten scientists from two research groups (one in Germany, one in Canada). The participants were provided with two different user profiles each, which were automatically generated based on their publications: One based on the articles' full texts, the second restricted to rhetorical entities (REs) [23]. We asked each participant to rate the relevance of the top-N entries, as well as their competence level. The results, provided in Section 6, show that our approach can automatically generate user profiles with a precision of up to 92%.

Finally, we illustrate in Section 7 how semantic user profiles can be leveraged by scholarly information systems in a number of use cases, including a competence analysis for a user (e.g., for finding reviewers for a new paper) and re-ranking of article search results, based on a user's profile.[3]

2 Background

In this section, we provide background information on user profiling and its applications. We also briefly introduce semantic technologies for user profiling and their connections with natural language processing (NLP) techniques.

2.1 User Profiling and Personalization

A user profile is an instance of a user model that contains either a user's characteristics, such as knowledge, interests and backgrounds, or may focus on the context of a user's work, e.g., location and time [5]. Depending on the application offering personalized content, different features have to be taken into account. For instance, educational learning systems typically model a user's knowledge and background, whereas recommender systems and search applications are more focused on a user's interests. Constructing user profiles requires collecting user information over an extended period of time. This gathering process is called *user profiling* and distinguishes between *explicit* and *implicit* user feedback. Explicit user feedback actively requests interests from a user, whereas

[3] For supplementary material, please visit http://www.semanticsoftware.info/save-sd2016.

implicit user feedback derives preferences from user activities. Commonly used implicit profiling techniques, such as extracting preferences from visited websites and deriving interest weights from the numbers of clicks, are discussed by Gauch et al. [11].

User profiles are the basis for a variety of personalized applications. For instance, recommender systems and personalized news portals utilize user information, specifically purchased articles or search terms, in order to adapt content to user needs. The most dominant representation of user characteristics is a weighted vector of keywords, which is still used in many current adaptive web systems [1,17]. This mathematical description makes it possible to apply classical information filtering algorithms, such as cosine similarity [18], in order to measure item-to-item, user-to-user and item-to-user similarity.

2.2 Semantic Technologies

Semantic technologies have become increasingly important in the management of research objects. They allow automated systems to understand the meaning (semantics) and infer additional knowledge from published documents and data [25,2]. Essential building blocks for the creation of structured, meaningful web content are information extraction and semantic annotations – results that can be obtained from NLP pipelines, for example, to detect rhetorical zones, such as *claims* or *contributions* of a paper [23].

In the area of user modeling, a multitude of semantic approaches have emerged in the last decade that use concepts of domain ontologies in the vector representation, rather than keywords [26,6]. In addition to a common understanding of domain knowledge, using semantic technologies also fosters evolving towards more generic user models. A goal of generic user modeling is facilitating software development and promoting reusability [15]. Semantic web technologies, such as the representation of user characteristics in an RDF or OWL format, can leverage this idea. In the following section, we introduce different proposals for generic user modeling with semantic web models. Furthermore, we discuss scholarly ontologies that describe users, institutions and publications in the scientific domain.

3 Literature Review

We focus our review on two core aspects: Firstly, existing semantic vocabularies that describe scholars in academic institutions with their publications and competences, in order to establish semantic user profiles. And secondly, we examine existing approaches for automatic profile generation through NLP methods.

3.1 Vocabularies for Semantic User Profiles

GUMO [14] was the first generic user model approach, designed as a top-level ontology for universal use. This OWL-based ontology focuses on describing a user in a situational context, offering several classes for modeling a user's personality, characteristics and interests. Background knowledge and competences are considered only to a small degree. In contrast, the IntelLEO[4] ontology framework is strongly focused on personalization

[4] IntelLEO (Intelligent Learning Extended Organizations), http://intelleo.eu/index.php?id=183

and enables describing preferences, tasks and interests. The framework consists of multiple RDFS-based ontologies, including vocabularies for user and team modelling, as well as competences. They are inter-linked and can be connected with other user model ontologies, such as FOAF.[5] Due to its simplicity and linkage to other Linked Open Vocabularies, FOAF has become very popular in recent years and is used in numerous personalized applications [22,20,7]. This RDF-based vocabulary permits describing basic user information with predefined entities, such as name, email, homepage, and interests, as well as modeling persons and groups in social networks. However, FOAF does not provide comprehensive classes for describing preferences and competences. Other ontologies attempting to unify user modeling in semantic web applications are the Scrutable User Modelling Infrastructure (SUMI) [16], the Generic User Model Component (GUC) [27] and the ontology developed by Golemati et al. [12].

For modeling scholars in the scientific domain, VIVO[6] [3] is the most prominent approach and has been used in numerous applications.[7] It is an open-source suite of web applications and ontologies used to model scholarly activities across an academic institution. However, VIVO does not provide for content customization, due to missing classes for user interests, preferences and competences. Further vocabularies modeling scientists and publications in research communities are SWRC,[8] SWPO[9] and LSC.[10]

3.2 Automatic Profile Generation

Generic user models require thinking about new methods for user profiling. Complex user information can be obtained from, e.g., observing a user's browsing behavior, but also from other sources related to the user. Utilizing NLP techniques in user modeling has quite a long history [28]; However, natural language systems are still rarely used for constructing semantic user profiles.

Paik et al. [21] developed <!metaMarker>, an NLP and machine learning pipeline that detects user information in emails. The mined data is used for constructing client profiles in personalized e-commerce applications. The system is able to extract explicit metadata, such as 'name of sender', 'title' or 'affiliation', as well as implicit metadata, like 'mood' or 'intention' of the user. Additionally, they enriched this context-related metadata with new elements, such as 'like', 'dislike', 'interested' and 'not interested', in order to describe a user's preferences. The pipeline consists of seven steps, including Sentence Splitting, Part-Of-Speech Tagging, Stemming and Entity Extraction, generating explicit user information at the end. Through Bayesian probabilistic and k-Nearest Neighbour classifiers, mood and intentions are determined. A training set of 5000 emails was used to build the classifiers for the implicit metadata. The effectiveness of the system was measured with precision and recall, resulting in an average precision of 89%.

LinkedVis [4] is an interactive recommender system that generates career recommendations and supports users in finding potentially interesting companies and specific roles.

[5] FOAF (Friend of a Friend), http://www.foaf-project.org/
[6] VIVO Ontology, http://vivoweb.org/ontology/core#
[7] VIVO Registry, http://duraspace.org/registry/vivo
[8] Semantic Web for Research Communities, http://ontoware.org/swrc/
[9] Semantic Web Portal Ontology, http://sw-portal.deri.org/ontologies/swportal#
[10] Linked Science Core Vocabulary, http://linkedscience.org/lsc/ns#

The authors designed four different user models based on data from *LinkedIn*[11] and extracted interests and preferences from a user's connections, average roles and companies. Two of the four constructed profiles contained meaningful entities instead of plain keywords. A Part-of-Speech Tagger was utilized to find noun phrases that were mapped to Wikipedia articles. The evaluation with a leave-one-out cross-validation revealed that the user models with the semantic enrichment produced more accurate and more diverse recommendations than the profiles based on TF-IDF weights and occurrence matching.

Another approach using NLP methods for online profile resolution is proposed by Cortis et al. [8]. They developed a system for analyzing user profiles from heterogenous online resources in order to aggregate them into one unique profile. For this task, they used GATE's ANNIE[12] plugin [9] and adapted its JAPE grammar rules to disassemble a person's name into five sub-entities such as prefix, suffix, first name, middle name and surname. In addition, a Large Knowledge Base (LKB) Gazetteer was incorporated to extract supplementary city and country values from DBpedia.[13] In their approach, location-related attributes (e.g., Dublin and Ireland) could be linked to each other based on these semantic extensions, where a string-matching approach would have failed. In their user evaluation, the participants were asked to assess their merged profile on a binary rating scale. More than 80% of the produced profile entries were marked as correct. The results reveal that profile matchers can improve the management of one's personal information across different social networks and support recommendations of possibly interesting new contacts based on similar preferences.

3.3 Discussion

As presented above, there exist only few automatic user profiling approaches using linked named entities and NLP techniques. The most widespread description of a user model in these applications is still a term-based vector representation. Even though keywords are increasingly replaced by linked entities, they still lack an underlying semantic model in RDF or OWL format. With respect to existing application domains, social networks are common sources for gathering personal information. Scholars in particular were not considered in any of the aforementioned systems.

In contrast, we aim at automatically creating semantic user profiles for scholars by means of NLP methods and semantic web technologies. Our goal is to establish user profiles in an RDF format that can be stored in a triplestore. Hosting user information in a structured and meaningful semantic format facilitates data integration across different sources. Furthermore, expressive SPARQL queries and inferences can help to discover related preferences that are not explicitly stated in the profiles.

4 Design

In our approach, we take the publications of an author as input to an automated text mining pipeline, which creates user profiles in LOD format, based on the competences

[11] LinkedIn, https://www.linkedin.com
[12] ANNIE, https://gate.ac.uk/sale/tao/splitch6.html
[13] DBpedia, http://dbpedia.org

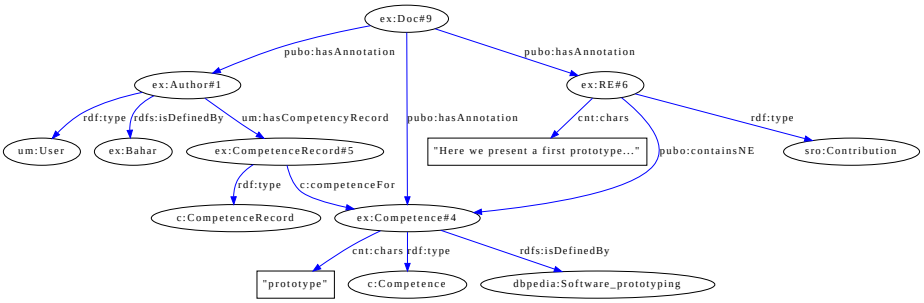


Fig. 1. A semantic scholar profile in form of an RDF graph

detected in the papers. The hypothesis behind our design is that authors of a scholarly publication (e.g., a journal article) are competent in the topics mentioned in the paper to various degrees. Our text mining system performs entity linking from scholarly documents and generates competence relations between a document's authors and its contained LOD named entities using linked open vocabularies. The result is a knowledge base containing the semantic profiles of authors that can be exploited for a variety of use cases by humans and machines alike, as we show in Section 7.

4.1 Semantic Modeling of Users' Competence Records

Modeling semantic scholarly profiles requires the formalization of the relation between authors, their publications, and the topics mentioned in them in a semantically rich and interoperable format. To this end, we decided to use the W3C standard RDF framework to design profiles based on semantic triples. Since RDF documents intrinsically represent labeled, directed graphs, the semantic profiles of scholars extracted from the documents can be merged through common competence URIs, i.e., authors extracted from otherwise disparate documents can be semantically related using their competence topics.

Following the best practices of producing linked open datasets, we tried to reuse existing Linked Open Vocabularies (LOVs) to the extent possible for modeling the extracted knowledge. Table 1 shows the vocabularies used to model our semantic scholarly profiles. We largely reuse IntelLEO ontologies for competence modeling – originally designed for semantic modeling of learning contexts –, in particular the vocabularies for *User and Team Modeling*[14] and *Competence Management*.[15] We also reuse the PUBO ontology [23] for modeling the relation between the documents that we process, the generated annotations and their inter-relationships. Figure 1 shows a minimal example semantic profile in form of an RDF graph.

4.2 Automatic Detection of Competences

Our text mining system accepts a set of publications from an author as input and processes the full-text of the documents to detect competence topics, i.e., grounded Named Entities

[14] IntelLEO User Model Ontology, http://intelleo.eu/ontologies/user-model/spec
[15] IntelLEO Competence Ontology, http://www.intelleo.eu/ontologies/competences/spec

Table 1. Concepts from linked open vocabularies for the semantic modeling of scholar user profiles

LOV Term	Modeled Concept
um:User	Scholar users, who are the documents' authors.
um:hasCompetencyRecord	A property to keep track of a user's competence (level, source, etc.).
c:Competency	Extracted topics (LOD resources) from documents.
c:competenceFor	A relation between a competency record and the competence topic.
sro:RhetoricalElement	A sentence containing a rhetorical entity, e.g., a *contribution*.
cnt:chars	A competence's label (surface form) as appeared in the document.
pubo:hasAnnotation	A property to relate annotations to documents.
pubo:containsNE	A property to relate rhetorical zones and entities in the document.
oa:start & oa:end	A property to show the start/end offsets of competences in text.

um: http://intelleo.eu/ontologies/user-model/ns/ c: http://intelleo.eu/ontologies/competences/ns/
sro: http://salt.semanticauthoring.org/ontologies/sro# cnt: http://www.w3.org/2011/content#
pubo: http://lod.semanticsoftware.info/pubo/pubo# oa: http://www.w3.org/ns/oa/
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# rdfs: http://www.w3.org/2000/01/rdf-schema#

(NEs). Each document first goes through a pre-processing phase. In this phase, the full-text of the document is segmented into tokens: smaller, linguistically meaningful parts, like words, numbers and symbols. Subsequent syntactical processing components process the tokenized text into sentences and all sentence constituents are tagged with a Part-of-Speech category. Grammatical processing of sentences helps us to filter out the text tokens that do not represent competences, like adverbs or pronouns. Lastly, we ground (link) nouns and noun phrases in text to their corresponding resource (sense) in the LOD cloud. To this end, we selected the DBpedia Spotlight [19] annotation tool that can link the surface forms of terms in a document to a URI in the DBpedia ontology that serves as the nucleus of the LOD cloud. In this paper, we use the raw frequency of these NEs in documents as a means of ranking the top competence topics for researchers' profiles. Finally, once the documents are processed, we go over the generated annotations and transform them into RDF triples, using the vocabularies described in Section 4.1.

A rather interesting question here is whether all of the detected entities are representative of the authors' interest, or if topics in certain regions of the documents are better candidates? To test this hypothesis, we further process the documents to annotate their so-called *Rhetorical Entities (REs)*, where authors convey their findings in form of claims or arguments, by looking at their linguistic features [24]. In this fashion, we can later evaluate whether the NEs in RE zones of documents better represent the authors' competences.

5 Implementation

In this section, we describe how we realized the semantic user profiling of authors illustrated in the previous section.

5.1 Extraction of User Competences with Text Mining

We developed a text mining pipeline, implemented based on the GATE framework, to analyze a given author's papers to automatically extract the competence records and topics. The NLP pipeline accepts a corpus (set of documents) for each author as input.

We use GATE's ANNIE plugin to pre-process each document's full-text and further process all sentences with a Part-of-Speech (POS) tagger, so that their constituents are labeled with a POS tag, such as *noun*, *verb*, or *adjective* and lemmatized to their canonical (root) form. We use MuNPEx,[16] a GATE plugin to detect noun phrases in text, which helps us to extract competence topics that are noun phrases rather than nouns alone. Subsequently, we use our LODtagger,[17] which is a GATE plugin that acts as a wrapper for the annotation of documents with Named Entity Recognition tools. In our experiments, we use a local installation of DBpedia Spotlight v7.0 with a statistical model[18] for English [10]. Spotlight matches the surface form of the document's tokens against the DBpedia ontology and links them to their corresponding resource URI. LODtagger then transforms the Spotlight response to GATE annotations using the entities' offsets in text and keeps their URI in the annotation's features.

To evaluate whether our hypothesis that the NEs within rhetorical zones of a document are more representative of the author's competences than the NEs that appear anywhere in the document, we decided to annotate the *Claim* and *Contribution* sentences of the documents using our Rhetector[19] GATE plugin [23]. This way, we can create user profiles exclusively from the competence topics that appear within these RE annotations for comparison against profiles populated from full-text.[20] Finally, we create a competence record between the author and each of the detected competences (represented as DBpedia NEs). We use GATE's JAPE language that allows us to execute regular expressions over documents' annotations by internally transforming them into finite-state machines. Thereby, we create a competence record (essentially, a GATE relation) between the author annotation and every competence topic in the document.

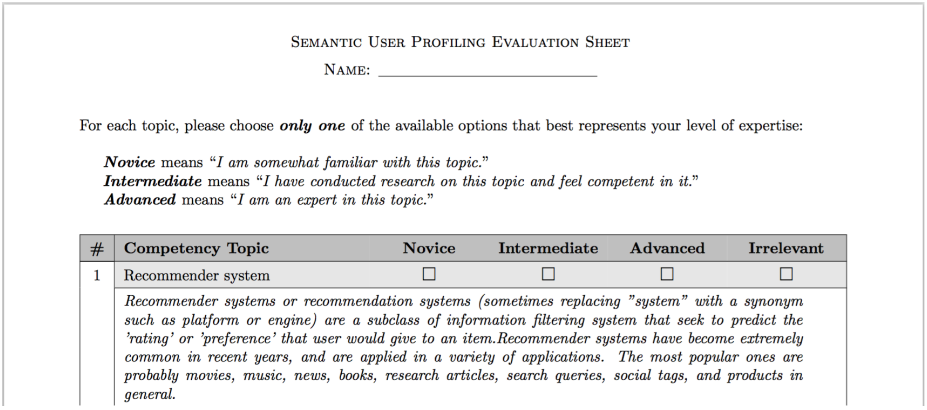
5.2 Automatic Population of Semantic User Profiles

The last step in our automatic generation of semantic user profiles is to export all of the GATE annotations and relations from the syntactic and semantic processing phases into semantic triples using RDF. Our LODeXporter[21] tool provides a flexible mapping of GATE annotations to RDF triples with user-defined transformation rules. For example, the rules:

```
map:GATECompetence   map:GATEtype   "DBpediaNE" .
map:GATECompetence   map:hasMapping   map:GATELODRefFeatureMapping .
map:GATELODRefFeatureMapping   map:GATEfeature   "URI" .
map:GATELODRefFeatureMapping   map:type   rdfs:isDefinedBy .
```

describe that all "*DBpediaNE*" annotations in the document should be exported, and for each annotation the value of its "*URI*" feature can be used as the object of the triple, using "*rdfs:isDefinedBy*" as the predicate. Similarly, we use the LOV terms shown in

[16] Multi-lingual Noun Phrase Extractor (MuNPEx), http://www.semanticsoftware.info/munpex
[17] LODtagger, http://www.semanticsoftware.info/lodtagger
[18] DBpedia statistical model for English (en_2+2), http://spotlight.sztaki.hu/downloads/
[19] Rhetector, http://www.semanticsoftware.info/rhetector
[20] Rhetector was evaluated in [23] with an average F-measure of 73%.
[21] LODeXporter, http://www.semanticsoftware.info/lodexporter



SEMANTIC USER PROFILING EVALUATION SHEET

NAME: _____

For each topic, please choose ***only one*** of the available options that best represents your level of expertise:

Novice means "*I am somewhat familiar with this topic.*"
Intermediate means "*I have conducted research on this topic and feel competent in it.*"
Advanced means "*I am an expert in this topic.*"

#	Competency Topic	Novice	Intermediate	Advanced	Irrelevant
1	Recommender system	☐	☐	☐	☐
	Recommender systems or recommendation systems (sometimes replacing "system" with a synonym such as platform or engine) are a subclass of information filtering system that seek to predict the 'rating' or 'preference' that user would give to an item.Recommender systems have become extremely common in recent years, and are applied in a variety of applications. The most popular ones are probably movies, music, news, books, research articles, search queries, social tags, and products in general.				

Fig. 2. Excerpt of a sample generated user profile for evaluation

Table 1 to model authors, competence records and topics as semantic triples and store the results in an Apache TDB-based[22] triplestore.

6 Evaluation

To evaluate the accuracy of the generated profiles, we reached out to ten computer scientists from Concordia University and the University of Jena (including the authors of this paper) and asked them to provide us with a number of their selected publications. We processed the documents and populated a knowledge base with the researchers' profiles. We also developed a Java command-line tool that queries the knowledge base and generates LaTeX documents to provide for a human-readable format of the researchers' profiles (shown in Figure 2) that lists their top-50 competence topics sorted by the number of occurrence in the users' publications. Subsequently, we asked the researchers to review their profiles across two dimensions: *(i)* relevance of the extracted competences, and *(ii)* their level of expertise for each extracted competence.

For each participant, we exported two versions of their profile: *(i)* a version with a list of competences extracted from their papers' full-text, and *(ii)* a second version that only lists the competences extracted from the rhetorical zones of the documents, in order to test our hypothesis described in Section 5.1. To ensure that none of the competence topics are ambiguous to the participants, our command-line tool also retrieves the English label and comment of each topic from the DBpedia ontology using its public SPARQL endpoint.[23] The participants were instructed to choose only one level of expertise for each competence and choose "*irrelevant*" if the competence topic was incorrect or grounded to a wrong sense.

To evaluate the effectiveness of our system, we utilize one of the most popular ranked retrieval evaluation methods, namely the Mean Average Precision (MAP) [18]. MAP

[22] Apache TDB, http://jena.apache.org/documentation/tdb/
[23] DBpedia public SPARQL endpoint, http://dbpedia.org/sparql

Table 2. Evaluation of the generated user profiles

Participant	#Docs	#Distinct Competences		Avg. Precision@10		Avg. Precision@25		Avg. Precision@50	
		Full Doc	REs Only	Full Doc	REs Only	Full Doc	REs Only	Full Doc	REs Only
R1	8	2,718	293	**0.91**	0.80	**0.84**	0.74	**0.80**	0.69
R2	7	2,096	386	**0.95**	0.91	0.90	**0.92**	0.87	**0.91**
R3	6	1,200	76	0.96	**0.99**	0.93	**0.95**	**0.92**	0.88
R4	5	1,240	149	0.92	**0.92**	**0.86**	0.81	**0.77**	0.75
R5	4	1,510	152	0.84	**0.99**	0.87	**0.90**	0.82	**0.82**
R6	6	1,638	166	0.93	**1.0**	0.90	**0.97**	0.88	**0.89**
R7	3	1,006	66	0.70	**0.96**	0.74	**0.89**	0.79	**0.86**
R8	8	2,751	457	0.96	**1.0**	0.92	**1.0**	0.92	**0.99**
R9	9	2,391	227	0.67	**0.73**	0.62	**0.70**	0.56	**0.65**
R10	5	1,908	176	**0.96**	0.91	0.79	**0.80**	0.69	**0.70**
			MAP	0.88	**0.92**	0.83	**0.87**	0.80	**0.81**

indicates how precise an algorithm or system ranks its top-N results, assuming that the entries listed on top are more relevant for the information seeker than the lower ranked results. Table 2 shows the evaluation results of our user study. A competence was considered as relevant when it had been assigned to one of the three levels of expertise (novice, intermediate, advanced). For each participant, we measured the average precision of the generated profiles in both the full-text and RE-only versions. Here, precision is evaluated at a given cut-off rank N, considering only the top-N results returned by the system. Hence, MAP is the mean of the average precisions at each cut-off rank. The results show that for both the top-10 and top-25 competences, 70–80% of the profiles generated from RE-only zones had a higher precision, increasing the system MAP up to 4% in each cut-off. In the top-50 column, we observed a slight decline in some of the profiles' average precision, which we believe to be a consequence of more irrelevant topics appearing in the profiles, although the MAP score stays almost the same for both versions. Analyzing the distribution of answers across the three levels of expertise, the results illustrated in Figure 3 reveal that in both versions, around 60% of the detected competences are related to either the intermediate or advanced level.

Finally, all participants (except R10) informally stated that the RE-only version of their profiles were better representing their competences, corroborating our hypothesis that the topics mentioned in the RE zones of a document are more accurate in representing its authors' competences. This is encouraging because, as shown in Table 2, compared to the number of distinct competences extracted from the full-text of documents, we need an order of a magnitude fewer topics, which not only better represent the users' competences, but also significantly reduces the size of the knowledge base.

7 Application

In this section, we demonstrate a number of use cases in which semantic user profiles can play an effective role.

7.1 Finding all competences of a user

By querying the populated knowledge base with the researchers' profiles, we can find all topics that a user is competent in. Following our knowledge base schema (see Section 4),

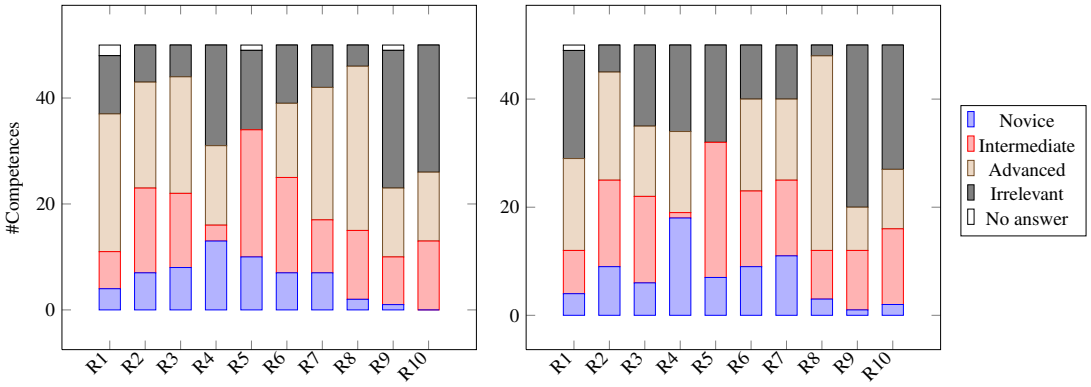


Fig. 3. Distribution of competence levels in full-text (left) and RE-only (right) profiles

we can query all the competence records of a given author URI and find the topics (in form of LOD URIs), from either the papers' full-text or exclusively the RE zones. In fact, the SPARQL query shown below is how we gathered each user's competences (from RE zones) to generate the evaluation profiles described in Section 6:

```
SELECT DISTINCT ?uri (COUNT(?uri) AS ?count) WHERE {
    ?creator rdf:type um:User .
    ?creator rdfs:isDefinedBy <http://semanticsoftware.info/lodexporter/creator/R1> .
    ?creator um:hasCompetencyRecord ?competenceRecord .
    ?competenceRecord c:competenceFor ?competence .
    ?competence rdfs:isDefinedBy ?uri .
    ?rhetoricalEntity rdf:type sro:RhetoricalElement .
    ?rhetoricalEntity pubo:containsNE ?competence .
} GROUP BY ?uri ORDER BY DESC(?count)
```

Table 3 shows a number of competence topics (grounded to their LOD URIs) for some of our evaluation participants, sorted in descending order by their frequency in the documents.

7.2 Ranking papers based on a user's competences

Semantic user profiles can be incredibly effective in the context of information retrieval systems. Here, we demonstrate how they can help to improve the relevance of the results. Our proposition is that papers that mention the competence topics of a user are more *interesting* for her and thus, should be ranked higher in the results. Therefore, the

Table 3. A number of users and their most frequent competence topics

User	Extracted Competence Topics
R1	dbpedia:Tree_(data_structure), dbpedia:Vertex_(graph_theory), dbpedia:Cluster_analysis, ...
R2	dbpedia:Natural_language_processing, dbpedia:Semantic_Web, dbpedia:Entity-relationship_model, ...
R3	dbpedia:Recommender_system, dbpedia:Semantic_web, dbpedia:Web_portal, dbpedia:Biodiversity, ...
R4	dbpedia:Service_(economics), dbpedia:Feedback, dbpedia:User_(computing), dbpedia:System, ...
R5	dbpedia:Result, dbpedia:Service_discovery, dbpedia:Web_search_engine, dbpedia:Internet_protocol, ...

Table 4. Re-ranking of the top-10 search results, originally sorted by a frequency-based method, through integrating semantic user profiles

	Topic Mentions		R8's Profile		R6's Profile	
Paper Title	**Rank**	**Raw Frequency**	**Rank**	**Com. Topics**	**Rank**	**Com. Topics**
A Review of Ontologies for Describing Scholarly and Scientific Documents	1	92	1	312	5	198
BauDenkMalNetz - Creating a Semantically Annotated Web Resource of Historical Buildings	2	50	5	294	4	203
Describing bibliographic references in RDF	3	38	6	269	8	177
Semantic Publishing of Knowledge about Amino Acids	4	25	10	79	10	53
Supporting Information Sharing for Re-Use and Analysis of Scientific Research Publication Data	5	25	4	306	7	185
Linked Data for the Natural Sciences: Two Use Cases in Chemistry and Biology	6	23	2	310	1	220
Ornithology Based on Linking Bird Observations with Weather Data	7	22	8	248	6	189
Systematic Reviews as an Interface to the Web of (Trial) Data: using PICO as an Ontology for Knowledge Synthesis in Evidence-based Healthcare Research	8	19	9	179	9	140
Towards the Automatic Identification of the Nature of Citations	9	19	3	307	2	214
SMART Research using Linked Data - Sharing Research Data for Integrated Water Resources Management in the Lower Jordan Valley	10	19	7	260	3	214

diversity and frequency of topics within a paper should be used as ranking features. We showed in [23] that retrieving papers based on their LOD entities is more effective than conventional keyword-based methods. However, the results were not presented in order of their *interestingness* for the end-user. Here, we integrate our semantic user profiles to re-rank the results, based on the common topics in both the papers and a user's profile:

```
SELECT (COUNT(DISTINCT ?uri) as ?rank) WHERE {
    <http://example.com/example_paper.xml> pubo:hasAnnotation ?topic .
    ?topic rdf:type pubo:LinkedNamedEntity .
    ?topic rdfs:isDefinedBy ?uri .
FILTER EXISTS {
    ?creator rdfs:isDefinedBy <http://semanticsoftware.info/lodexporter/creator/R8> .
    ?creator um:hasCompetencyRecord ?competenceRecord .
    ?competenceRecord c:competenceFor ?competence .
    ?competence rdfs:isDefinedBy ?uri .} }
```

The query shown above compares the topic URIs in a given paper to user R8's competences extracted from full-text documents and counts the occurrence of such a hit. Note that the `DISTINCT` keyword will cause the query to only count the unique topics, e.g., if <dbpedia:Semantic_Web> appears two times in the paper, it will be counted as one occurrence.[24] We can then use the numbers returned by the query above as a means to rank the papers. Table 4 shows the result set returned by performing a query against the SePublica dataset of 29 papers from [23] to find papers mentioning <dbpedia:Ontology_(information_science)>. The "*Topic Mentions*" column shows the ranked results based on how many times the query topic was mentioned in a document. In contrast, the R6 and R8 profile-based columns show the ranked results using the number of common topics between the papers (full-text) and the researchers' respective profiles

[24] We decided to the count the unique occurrences, because a ranking algorithm based on the raw frequency of competence topics will favour long (non-normalized) papers over shorter ones.

(populated from full-text documents). Note that in the R6 and R8 profile-based columns, we only count the number of unique topics and not their frequency. An interesting observation here is that the paper ranked fourth in the frequency-based column ranks last in both profile-based result sets. A manual inspection of the paper revealed that this document, although originally ranked high in the results, is in fact an editors' note in the preface of the SePublica 2012 proceedings. On the other hand, the paper which ranked first in the frequency-based column, remained first in R8's result set, since he has a stronger research focus on ontologies and linked open data compared to R6, as we observed from their generated profiles during evaluation.

7.3 Finding users with related competences

Given the semantic user profiles and a topic in form of an LOD URI, we can find all users in the knowledge base that have related competences. By virtue of traversing the LOD cloud, we can find topic URIs that are (semantically) related to a given competence topic and match against users' profiles to find competent authors:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dbpedia: <http://dbpedia.org/resource/>

SELECT ?author_uri WHERE {
    SERVICE <http://dbpedia.org/sparql> {
      dbpedia:Ontology_( information_science ) dcterms:subject ?category .
      ?subject dcterms:subject ?category . }
?author rdf:type um:User .
?creator rdfs:isDefinedBy ?author_uri .
?creator um:hasCompetencyRecord ?competenceRecord.
?competenceRecord c:competenceFor ?competence.
?competence rdfs:isDefinedBy ?subject .
? rhetoricalEntity pubo:containsNE ?competence.
? rhetoricalEntity rdf:type sro:RhetoricalElement . }
```

The query above first performs a federated query against DBpedia's SPARQL endpoint to find topic URIs that are semantically related to the query topic.[25] Then, it matches the retrieved URIs against the topics of the knowledge base users' competence records. This way, for example as shown in Table 5, even if a researcher does not have <dbpedia:Ontology_(information_science)>, but does have <dbpedia:Linked_data> in her profile, she will be returned as a hit, since both of the aforementioned topics are related in the DBpedia ontology. In other words, if we are looking for persons competent in ontologies, a researcher that has previously conducted research on linked data might also be a suitable match.

8 Conclusions

Semantic user profiles are an important extension for semantic publishing applications: With a standardized, shareable, and extendable representation of a user's competences, a number of novel scenarios become possible. Searching for scientists with specific competences can help to find reviewers for a given paper or proposal. Recommendation algorithms can filter and rank the immense amount of research objects, based on the

[25] We assume all topics under the same category in the DBpedia ontology are semantically related.

Table 5. Topics related to the query and their respective competent researchers

Competence Topic	Competent Users
dbpedia:Ontology_(information_science)	R1, R2, R3, R8
dbpedia:Linked_data	R2, R3, R8
dbpedia:Knowledge_representation_and_reasoning	R1, R2, R4, R8
dbpedia:Semantic_Web	R1, R2, R3, R4, R5, R6, R7, R8
dbpedia:Controller_vocabulary	R2, R3, R8
dbpedia:Tree_(data_structure)	R1, R4, R7

profile of individual users. And a wealth of additional applications becomes feasible, such as matching the competences of a research group against project requirements, simply by virtue of analyzing an inter-linked knowledge graph of users, datasets, publications, and other artifacts. The work presented here demonstrates how we can represent scholarly profiles in LOD format. We show how to bootstrap semantic user profiles including scientists' competences through an automated text mining approach with high accuracy. In ongoing work, we are currently integrating the semantic user profiles into a scholarly data portal for biodiversity research, in order to evaluate their impact on concrete research questions in a life sciences scenario.

Acknowledgments. We would like to thank all the participants in our user study.

References

1. Almuhaimeed, A., Fasli, M.: A semantic method for multiple resources exploitation. In: Proceedings of the 11th International Conference on Semantic Systems (SEMANTICS '15). pp. 113–120. ACM, New York, NY, USA (2015)
2. Berners-Lee, T., Hendler, J.: Publishing on the semantic web. Nature 410, 1023–1024 (2001)
3. Börner, K., Conlon, M., Corson-Rikert, J., Ding, Y.: VIVO: A Semantic Approach to Scholarly Networking and Discovery. Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers (2012)
4. Bostandjiev, S., O'Donovan, J., Höllerer, T.: Linkedvis: Exploring social and semantic career recommendations. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13). pp. 107–116. ACM, New York, NY, USA (2013)
5. Brusilovsky, P., Millán, E.: User models for adaptive hypermedia and adaptive educational systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, Lecture Notes in Computer Science, vol. 4321, pp. 3–53. Springer Berlin Heidelberg (2007)
6. Cantador, I., Castells, P.: Extracting multilayered communities of interest from semantic user profiles: Application to group modeling and hybrid recommendations. Computers in Human Behavior 27(4), 1321–1336 (2011)
7. Celma, O.: Foafing the music: Bridging the semantic gap in music recommendation. In: Proceedings of the 5th International Conference on The Semantic Web (ISWC'06). pp. 927–934. Springer-Verlag, Berlin, Heidelberg (2006)
8. Cortis, K., Scerri, S., Rivera, I., Handschuh, S.: An ontology-based technique for online profile resolution. In: Social Informatics, Lecture Notes in Computer Science, vol. 8238, pp. 284–298. Springer International Publishing (2013)
9. Cunningham, H., et al.: Text Processing with GATE (Version 6). University of Sheffield, Department of Computer Science (2011), http://tinyurl.com/gatebook

10. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving Efficiency and Accuracy in Multilingual Entity Extraction. In: Proc. of the 9th International Conference on Semantic Systems (I-Semantics) (2013)
11. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User Profiles for Personalized Information Access. In: The Adaptive Web, chap. User Profiles for Personalized Information Access, pp. 54–89. Springer-Verlag, Berlin, Heidelberg (2007)
12. Golemati, M., Katifori, A., Vassilakis, C., Lepouras, G., Halatsis, C.: Creating an ontology for the user profile: Method and applications. In: Proceedings of the First International Conference on Research Challenges in Information Science (RCIS) (2007)
13. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis lectures on the semantic web: theory and technology, Morgan & Claypool Publishers (2011)
14. Heckmann, D., Schwartz, T., Brandherm, B., Schmitz, M., von Wilamowitz-Moellendorff, M.: Gumo – the general user model ontology. In: User Modeling 2005, Lecture Notes in Computer Science, vol. 3538, pp. 428–432. Springer Berlin Heidelberg (2005)
15. Kobsa, A.: Generic user modeling systems. User Modeling and User-Adapted Interaction 11(1-2), 49–63 (2001)
16. Kyriacou, D., Davis, H.C., Tiropanis, T.: A (multidomainsional) scrutable user modelling infrastructure for enriching lifelong user modelling. In: Lifelong User Modelling Workshop (in conjunction with conference UMAP 2009), Trento, Italy (2009)
17. Malhotra, A., Totti, L.C., Jr., W.M., Kumaraguru, P., Almeida, V.: Studying user footprints in different online social networks. CoRR abs/1301.6870 (2013)
18. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press (2008)
19. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In: Proceedings of the 7th International Conference on Semantic Systems. pp. 1–8. ACM (2011)
20. Orlandi, F., Breslin, J., Passant, A.: Aggregated, interoperable and multi-domain user profiles for the social web. In: Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12). pp. 41–48. ACM, New York, NY, USA (2012)
21. Paik, W., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., Amice, C.: Applying Natural Language Processing (NLP) Based Metadata Extraction to Automatically Acquire User Preferences. In: Proceedings of the 1st International Conference on Knowledge Capture (K-CAP '01). pp. 116–122. ACM, New York, NY, USA (2001)
22. Raad, E., Chbeir, R., Dipanda, A.: User profile matching in social networks. In: The 13th International Conference on Network-Based Information System (2010)
23. Sateli, B., Witte, R.: Semantic representation of scientific literature: bringing claims, contributions and named entities onto the Linked Open Data cloud. PeerJ Computer Science 1(e37) (2015), https://peerj.com/articles/cs-37/
24. Sateli, B., Witte, R.: What's in this paper? Combining Rhetorical Entities with Linked Open Data for Semantic Literature Querying. In: Semantics, Analytics, Visualisation: Enhancing Scholarly Data (SAVE-SD 2015). p. 1023–1028. ACM, Florence, Italy (2015), http://www.www2015.it/documents/proceedings/companion/p1023.pdf
25. Shadbolt, N., Hall, W., Berners-Lee, T.: The semantic web revisited. Intelligent Systems, IEEE 21(3), 96–101 (2006)
26. Sieg, A., Mobasher, B., Burke, R.: Web Search Personalization with Ontological User Profiles. In: Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM '07). pp. 525–534. ACM, New York, NY, USA (2007)
27. van der Sluijs, K., Houben, G.J.: Towards a generic user model component. In: Workshop on Personalization on the Semantic Web (PerSWeb05), Edinburgh, Scotland. pp. 47–57 (2005)
28. Zukerman, I., Litman, D.: Natural language processing and user modeling: Synergies and limitations. User Modeling and User-Adapted Interaction 11(1-2), 129–158 (2001)