promise

PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

Deliverable 3.6

Semantic Representation and Enrichment of Information Retrieval Experimental Data

Version 1.0, 27 September 2013

SEVENTH FRAMEWORK PROGRAMME

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

Document Information

Deliverable number	3.6
Deliverable title	Semantic Representation and Enrichment of Information Retrieval Experimental Data
Delivery date	27 September 2013
Lead contractor for this deliverable	NUIG
Author(s)	Georgeta Bordea, Paul Buitelaar, Gianmaria Silvello, Nicola Ferro, and Toine Bogers
Participant(s)	NUIG, RSLIS, UNIPD
Workpackage	WP3
Workpackage title	Evaluation Infrastructure
Workpackage leader	UNIPD
Dissemination Level	PU – Public
Version	1.0
Keywords	Experimental Data, Evaluation Campaign, Semantic Enrichment, Expert Search, Expertise Topic

History of Versions

Version	Date	Status	Author	Description
0.10	2012-01-15	Draft	NUIG	Initial draft
0.11	2012-05-23	Draft	NUIG	Updated after UNIPD exchange on semantic representation
1.00	2013-08-31	Final	NUIG, UNIPD, RSLIS	First final version

Abstract

Evaluation campaigns contribute considerably to the advancement of information retrieval systems. Structured data as well as unstructured data, in the form of scientific publications, is produced in this process, in a variety of application areas. In this work we present several steps in the direction of semantically annotating and interlinking these datsets, with the goal of enhancing their interpretation, sharing, and reuse. We discuss the underlying evaluation workflow and we propose a data model for those workflow areas that are directly related to Expert Search. A topic-centric approach for expert search is proposed, addressing the extraction of expertise topics and their semantic grounding.

Several methods for expert profiling and expert finding are analysed and evaluated on a dataset about Information Retrieval publications and workshops. Our results show that it is possible to construct expert profiles starting from automatically extracted expertise topics and that topic-centric approaches outperform language modelling approaches for expert finding.

Contents

Document Information 3

Abstract 3

Executive Summary 7

1 Introduction 9

2 Related Work 10

3 The Evaluation Workflow 11

4 Data Modeling for Expert Profiling 13

5 Semantic Enrichment 18
5.1 Expertise topic extraction . . . 19
5.2 Enriching expertise topics using Linked Open Data . . . 20
5.3 Expert profiling . . . 21
5.4 Expert finding . . . 22

6 Experimental evaluation 24
6.1 Experimental setup . . . 24
6.1.1 Information Retrieval workshop dataset . . . 24
6.1.2 Baseline approaches . . . 26
6.1.3 Evaluation measures . . . 27
6.2 Experiments . . . 27
6.2.1 Semantic grounding of expertise topics . . . 27
6.2.2 Expert profiling . . . 28
6.2.3 Expert finding . . . 29

7 Saffron. An Expert Search system for exploration and discovery of experts and expertise 30

8 Conclusion 35

References 37

Executive Summary

Motivation
Evaluation campaigns aim to create reusable test collections, but without concerted effort to semantically annotate and interlink these datasets, impact is limited to the participants of a shared task. Currently datasets, dataset descriptions, task descriptions and system descriptions have to be accessed separately. A solution to this problem is to enrich existing structured data by automatically extracting topical descriptions from scientific narrative. Because of the inherent complexity and heterogeneity of experimental data, it is difficult to find collaborators with an interest on a given topic or task, or to find all the test collections for a given topic. Identifying, measuring, and representing expertise has the potential to encourage interaction and collaboration, and ultimately knowledge creation, by constructing a web of connections between experts and the knowledge that they create. The combination of experimental data with information extracted from related publications and semantic metadata will enable a more meaningful interaction with this data.

Goals The goal of this work is to develop methods for matching experimental data with underlying publications, extracted topics and people (experts in different research areas, methods, etc.).

Methods We apply the Saffron expert finder system (http://saffron.deri.ie/) to the CLEF setting by bringing together experimental data from CLEF campaigns with information extracted from underlying publications (e.g. methods, tools, experts), wherever possible enriched with semantic metadata available e.g. through Linked Open Data.

Results A Linked Data based data model for two areas of the information retrieval evaluation workflow is proposed, focusing on the resource management area and the scientific production area. Expertise topics are automatically extracted and used to describe documents and to create expert profiles. Several topic-centric measures for expert finding are proposed, allowing users to identify knowledgeable members of the community. Additionally, an evaluation dataset for expert search in Information Retrieval is introduced, relying on scientific publications available online and on implicit expertise information about workshop committee members.

1 Introduction

Information Retrieval (IR) experimental evaluation is a process based on the Cranfield methodology [Cleverdon, 1997] that is carried out in the context of large-scale international evaluation campaigns. The Cranfield methodology makes use of shared experimental collections in order to create comparable experiments and evaluate the performances of different IR systems. Evaluation campaigns aim to guarantee an impartial comparison between different systems, reproducibility of the experiments, and re-use of the data adopted and produced during the campaigns. In this way, evaluation campaigns contribute considerably to the advancement of information retrieval systems by providing an infrastructure and resources for researchers to test, tune, evaluate and compare new approaches.

The scientific data produced by evaluation campaigns forms the basis for subsequent scientific work and system development, constituting an essential reference for the field. Until recently, limited attention had been paid to modeling, management, curation, citation, and access of the produced scientific data, even though the importance of scientific data in general has been highlighted by many different institutional organizations, such as the European Commission [European Union, 2010]. The research group on Information Management Systems (IMS) of the Department of Information Engineering of the University of Padua[1] started a few years ago the challenge of addressing the most common limitations on facing the issue [Agosti et al., 2006] and working on envisaging and defining a necessary infrastructure for dealing with the complexity of the challenge.

Large amounts of data are regularly produced in this process, including structured data about test collections, evaluation activities, evaluation measures, and visual analytics, as well as textual descriptions of shared tasks and reports describing experimental results. This data spans application areas as diverse as cultural heritage, eHealth, intellectual property, image retrieval, XML retrieval, plagiarism detection, question answering, and entity recognition. Additionally, each new campaign brings into focus new application areas.

A main goal of an evaluation campaign is to create reusable test collections, but without concerted effort to semantically annotate and interlink these datasets, impact is limited to the participants of a shared task. The importance of describing and annotating scientific datasets is discussed in [Bowers, 2012], noting that this is an essential step for their interpretation, sharing, and reuse. Currently datasets, dataset descriptions, task descriptions and system descriptions have to be accessed separately. A solution to this problem is to enrich structured data by automatically extracting topical descriptions from existing documents. In this work, we propose a standard representation and schema of IR experimental data as available in the DIRECT infrastructure [Agosti et al., 2012a]. This will enable a seamless integration of datasets produced by different campaigns such as TREC, NTCIR, and CLEF, standardising terms and concepts used to label data across research groups. Wherever possible, experimental data is enriched with semantic metadata available through Linked Open Data (LOD), connecting the dataset with other datasets from the LOD cloud.

Because of the inherent complexity and heterogeneity of experimental data, it is difficult to find collaborators with an interest on a given topic or task, or to find all the test collections for a given topic. Identifying, measuring, and representing expertise has the potential to encourage interaction

[1] http://www.dei.unipd.it/wdyn/?IDsezione=3314&IDgruppo_pass=121

and collaboration, and ultimately knowledge creation, by constructing a web of connections between experts and the knowledge that they create. These connections allow individuals to access knowledge beyond their tightly-knit networks, where members have access to similar information. Additionally, expertise development is accelerated by providing valuable insight to outsiders and novice members of a community. In this way experimental data can be linked with underlying publications and associated people through extracted topics. The combination of experimental data with information extracted from related scientific narrative and semantic metadata will enable a more meaningful interaction with this data.

This report is organised as follows. First, we give an overview of related work in Section 2, then we describe in more detail the overall evaluation workflow used for information access systems in Section 3. Section 4 presents a part of the data model that is relevant to Expert Profiling, using Linked Data principles. This data is further enriched by exploiting publications produced during evaluation campaigns and background knowledge available on the LOD cloud in Section 5. In Section 6, we discuss several experiments related to the semantic grounding of expertise topics, expert finding, and expert profiling. The techniques discussed in this work are integrated in Saffron, a system that allows discovery and exploration of experts and expertise, as we will see in Section 7. We conclude this report by presenting the conclusions and directions for future work in Section 8.

2 Related Work

Expert finding is the task of locating individuals knowledgeable about a specific topic, while expert profiling is the task of constructing a brief overview about the expertise topics of a person. Currently, these tasks received interest mainly for their application in an enterprise setting, but scientific communities can benefit as well from tools that enable collaboration. In an academic setting, existing work on expert finding focused on the task of finding qualified reviewers to assess the quality of research submissions [Mimno and McCallum, 2007; Rodriguez and Bollen, 2008]. In this work, we consider its applications for dissemination and sharing of experimental results in information retrieval.

Initial solutions for expert finding were developed under the area of competency management [Draganidis and Metzas, 2006]. These approaches are based on manual construction and querying of databases about knowledge and skills of an organization's workforce, placing the burden and responsibility of maintaining them on the employees themselves [Maybury, 2006]. A disadvantage of this approach is that because the information about experts and expertise is highly dynamic, considerable efforts are required to keep competency databases up-to-date. This prompted a shift to automated expert finding techniques that support a more natural expertise location process [Campbell et al., 2003].

Expert finding can be modelled as an information retrieval task using queries provided by users to perform a full text search for experts instead of documents. The goal of the search is to create a ranking of people who are experts in a given topic, instead of ranking relevant documents. A lot of ground was covered in terms of evaluating expert search systems by the organisation of three consecutive enterprise tracks by TREC [Bailey et al., 2007], that provided common ground for evaluating different systems and approaches. In this context, the expert finding task is modelled using statistical language modelling [Balog et al., 2006; Petkova and Croft, 2006] or data fusion techniques [Macdonald and Ounis, 2006].

The importance of expert profiling when developing solutions for expert search is discussed in [Balog and de Rijke, 2007], without addressing the problem of discovery and identification of expertise topics. The authors assume that a controlled vocabulary of terms is readily available for the considered domain. Currently such a resource is not available in our application setting, therefore we propose an automatic solution for the extraction of expertise topics by adapting existing term extraction and keyphrase extraction approaches.

An extensive analysis of expert profiling is presented in [Serdyukov et al., 2011], where the language model used in [Balog and de Rijke, 2007] is considered as one of the features used in a machine learning approach. Other features include a more simple binary model of relevance and the frequency of an expertise topic in expert profiles from the training set. Expertise topics, called tags in this work, are assumed to be known in advance, similar to [Balog and de Rijke, 2007], and are collected through self assessment. An important observation is that the quality of expertise topics is more important than the relevance to a particular person. In their experiments, the most important feature with respect to its performance contribution is the frequency of the expertise topic, a feature that is independent of the particular employee.

We build on this work by using a quality related measure of expertise topics together with relevance based measures for expert profiling. An intermediate conceptual level between documents and experts is introduced, similar to competency management approaches, avoiding their limitations such as manual gathering of data and quickly outdated profiles through automatic extraction of expertise topics.

3 The Evaluation Workflow

An experimental collection is a triple composed by: (i) a set of documents, called also collection of documents, which is representative of the domain of interest in terms of both kinds of documents and number of documents; (ii) a set of topics, which simulate actual user information needs and are often prepared from real system logs; the topics are then used by IR systems to produce the actual queries to be answered; and, (iii) the ground-truth or the set relevance judgements, i.e. a kind of “correct” answers, where for each topic the documents, which are relevant for that topic, are determined.

Experimental collections constitute the basis which allows for comparing different IR systems and a whole breadth of metrics has been developed over the years to assess the quality of produced rankings [Buettcher et al., 2010; Harman, 2011], according to different user models and tasks. Moreover, statistical approaches are adopted to assess significant differences in IR system performances [Hull, 1993; Savoy, 1997] and the quality of the evaluation metrics and experimental collection themselves [Buckley and Voorhees, 2000; Sakai, 2006].

In Figure 1 we can see the main phases of the Cranfield-based IR experimental evaluation workflow [Agosti and Ferro, 2009]; each phase of this workflow takes data as input and produces other data as output. The first phase is the acquisition and preparation of the documents constituting the first component of an experimental collection. The second phase is the creation of topics from

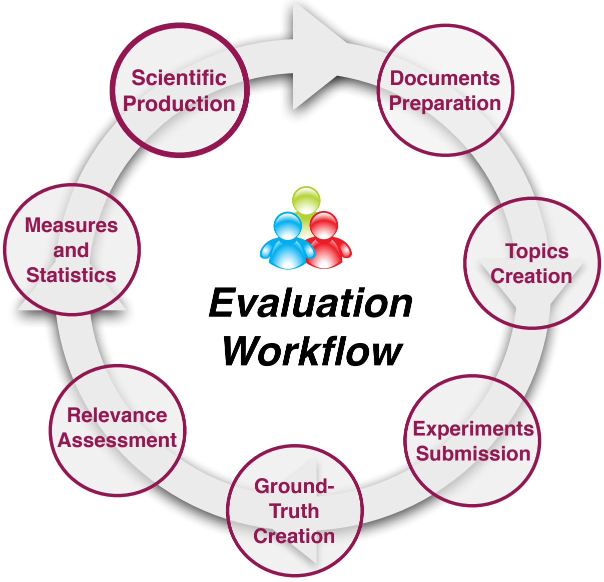


Figure 1: The typical IR experimental evaluation workflow and the data produced

which a set of queries is generated. After these two steps, the participants to the evaluation campaign have everything they need to run experiments and test their IR systems. An experiment is the output of an IR system which usually is composed by a set of ranked lists of documents – one for each topic. When this phase is over, the experiments are gathered by the campaign organizers which exploit them to create the ground-truth; this is done by adopting some appropriate sampling technique to select a subset of documents for each topic, which will be then manually assessed in the "Relevance Assessment" phase. In this phase, indeed, assessors decide whether or not a document is relevant for a given topic. Relevance judgments are raw data composing the experimental collection, but at the same time they represent human-added information connecting documents to topics of an experiment. The documents, topics, and relevant judgments triple is then used to compute performance measures about each experiment. In turn, measurements are used to produce descriptive statistics about the behavior of one or more systems. The last phase of the evaluation workflow regards scientific production where both participants and organizers prepare reports about the campaign and the experiments, the techniques they used and their findings; this phase usually continues also after the conclusion of the campaign as the investigations of the experimental results require a deep understanding and further analyses which may lead to the production of conference and journal papers; this phase involves also external actors who where not originally involved in the evaluation campaign. Indeed, the data employed in the evaluation workflow (i.e. documents, topics, and relevant judgments) as well as the data produced (i.e. experiments, measures and statistics and reports) are usually freely available to the scientific community which exploit them to produce new knowledge in the form of scientific papers. Scientific production is central to the evaluation workflow because it involves all the data used and produced in the process, all the actors who participates to the campaign and external actors who may exploit and elaborate the data.

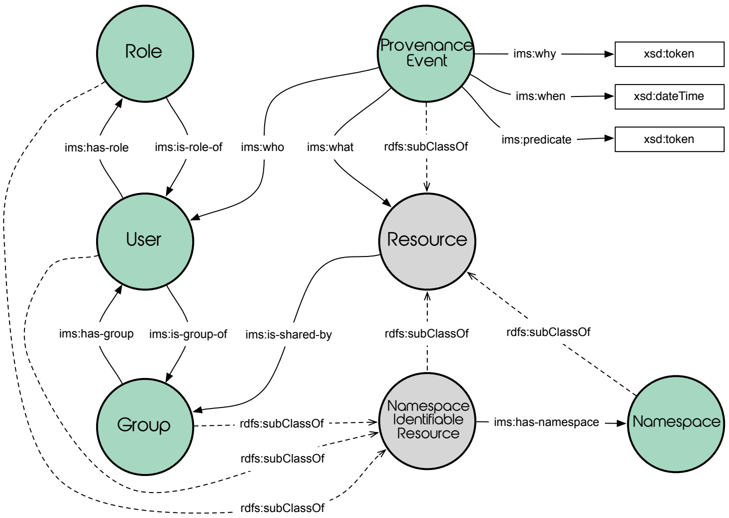


Figure 2: The Resource Management area classes and properties.

Scientific contributions can be considered as the summa of all the work done in the experimental evaluation process as they take into account the data employed and produced as well as they set the basis for further scientific improvements in the field. Scientific contributions represent the knowledge-base from which it is possible to extract information about user expertise, thus defining expert profiles.

4 Data Modeling for Expert Profiling

In order to explicitly take into consideration and model the valuable scientific data produced during an evaluation campaign, we have proposed an extension to the traditional evaluation methodology described above. To this end, we have undertaken the design of an evaluation infrastructure which manages the scientific data produced during a large-scale evaluation campaign [Agosti et al., 2010], as well as supports the archiving, access, citation, dissemination, and sharing of the experimental results [Agosti et al., 2012b; Di Nunzio and Ferro, 2005; Dussin and Ferro, 2009]; the outcome of this effort is *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*.

DIRECT covers all the described aspects of the evaluation workflow which lead to a rather complex system the presentation of which is out of the scope of this report; the full conceptual model and the architecture of DIRECT have been described and discussed in [Agosti et al., 2012b].The creation of expert profiles and the detection of scientific topics related to the data produced by the experimental evaluation, mainly concern two areas covered by DIRECT which we call the “resource management area” and the “scientific production area”. For what it is concerned with these two areas, the conceptual model of DIRECT has been mapped into a *Resource Description Framework (RDF)* model and adopted for enriching and sharing the data produced by the evaluation activities.

Within this model we consider a `Resource` as a generic class sharing the same meaning of resource in RDF[W3C, 2004] where “*all things described by RDF are called resources. [. . .] the*

Table 1: Main datatype properties of the resource management and contribution area classes reported in Figures 2 and 3. Namespace Identifiable Resource, Concept, Group, and Role are not reported because they have no additional datatype properties w.r.t. Resource. "ims" is the prefix for http://ims.dei.unipd.it/data/rdf/ pointing to DIRECT vocabulary terms.

Class	OWL Datatype Properties
Contribution	ims:affiliation, ims:title, ims:pages, ims:additional-information, ims:year, ims:link, ims:copyrighted
Link	ims:score, ims:backward-score, ims:frequency
Namespace	ims:prefix
Provenance-Event	ims:when, ims:why, ims:predicate
Resource	ims:identifier, ims:created, ims:last-modified, ims:description, ims:name, ims:content, ims:content-transfer-encoding, ims:language, ims:country
User	ims:password, ims:first-name, ims:last-name, ims:affiliation, ims:e-mail, ims:birth-date, ims:gender, ims:address, ims:city, ims:state, ims:zip, ims:phone, ims:facsimile, ims:mobile, ims:voip-caller-id, ims:homepage

class of everything." In DIRECT a Resource represents the class of everything that exists in the IR experimental evaluation.

The resource management area models the more general and thick-grained resources involved in the evaluation workflow – i.e. users, groups, roles, namespaces, and concepts – and the relationships among them. Furthermore, it handles the provenance (by means of the so-called Provenance-Event class) of the data. All the classes of this area are defined as subclasses of the general Resource class and they are represented in Figure 2 along the properties connecting them; for sake of readability we omitted from the figure the datatype properties, reported in Table 1, which are non-essential for the comprehension of the model.

The User class represents the actors involved in the evaluation activities such as researchers conducting experiments, organizers of a campaign, assessors, data scientists, and authors of a scientific contributions. The function of a user in the evaluation workflow is defined by the Role class; moreover, the users can be grouped together via the Group class. A user can play none, one or more roles: for instance, a user can be both an organizer of a campaign and a researcher submitting experiments, i.e. a participant to the campaign. On the other hand, there are roles played by more then one user; for instance, a campaign can have one or more participants, e.g. the researchers that are carrying out the experiments for writing a paper. A group is a resource that arrange together users with some common characteristics; for instance, there could be a group formed by all the users belonging to a certain research group.

The Namespace class refers to a logical grouping of identifiers and allows the disambiguation of homonym identifiers belonging to different namespaces. For instance, users are associated with a namespace which in the case of researchers allows us to classify them on an affiliation basis or the terms of an ontology are associated to a namespace allowing us to disambiguate with homonym terms of another ontology. In the RDF model of DIRECT along with the general `Resource` we described above, there is another general class called `Namespace Identifiable Resource` as we can see in Figure 2; this is a subclass of `Resource` always associated to a namespace. Thus, in the RDF model of DIRECT there are two kinds of general resources, the first which has no namespace and the second which has one. Thus, in Figure 2 we can see that `User`, `Group` and `Role` have a namespace, whereas the `Namespace` itself and `Provenance-Event` classes have no namespace.

The `Provenance-Event` class is not related to a namespace because it does not need to be disambiguated given that it exists only in the context of DIRECT. Indeed, a provenance event keeps track of the full lineage of each resource managed by DIRECT since its first creation, allowing granted users to reconstruct its full history and modifications over time. As shown in Figure 2, `Provenance-Event` is a subclass of `Resource` and it is composed by two object properties and three datatype properties, where:

- **who**, is the property associating the provenance event with the user who caused the event;
- **what**, is the property associating the provenance event with the specific resource originated by the event – please note that every resource in the model can be related to a provenance event;
- **when**, is the datatype property associating the provenance event with the timestamp at which the event occurred;
- **why**, is the datatype property associating the provenance event with the motivation that originated the event, i.e. the operation performed by the system that led to a modification of the resource;
- **predicate**, is the datatype property associating the provenance event with the action carried out in the event, i.e. CREATED, READ, or DELETED.

Modeling provenance is central for the definition of expert profiles and topic extraction because it allows for guaranteeing the quality and integrity of the data produced by the evaluation workflow [Buneman, 2013]. As we discussed above, the data produced by experimental evaluation are not raw data, but they are the product of a series of transformations which involve inputs from scientists and experts of the field. Keeping what was done with the data is crucial if we want to verify the quality or if we want to reproduce the experiments [Buneman, 2013]; moreover, these data are used for scientific production which in turn are exploited for expert profiling, two activities that must rely on high quality data. The `Provenance-Event` class allows us to record the five aspects (i.e. who, what, when, why and predicate) required for keeping the lineage of data [Cheney et al., 2009] and, consequently, the reliability of the information we extract and infer from these data.

In Figure 3 we can see the classes and the properties of the scientific production area. This part of the RDF model is central for the expert profiling activity because it handles scientific contributions, their relations with scientists and authors, and the scientific topics that can be extracted or inferred from them. In Figure 3 there are three main classes which are `Concept`, `Contribution` and `Link`.

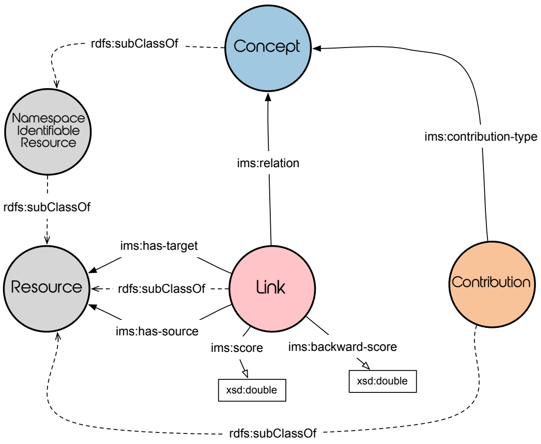


Figure 3: The Scientific Production area classes and properties.

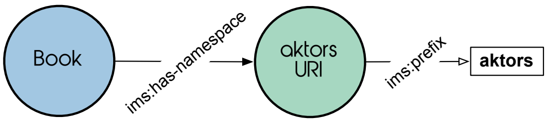


Figure 4: The RDF graph of the term “Book” of the Advanced Knowledge Technology reference ontology (i.e. aktors).

Concept is defined as an idea or notion, a unit of thought; it is used to define the type of relationships in a semantic environment or to create a vocabulary (e.g. contribution types) and, in some sense, resembles the idea of concept introduced by *Simple Knowledge Organization System (SKOS)* [W3C, 2009a,b]. Concept is a subclass of Namespace Identifiable Resource and thus every instance of it has a namespace. In DIRECT every vocabulary we create or import is handled via the Concept class. Let us consider a the term “Book” taken from the “Advanced Knowledge Technology reference ontology” which has http://www.aktors.org /ontology/portal# as *Uniform Resource Identifier (URI)* and prefix “aktors” as reported in Table 2. We can represent this term by instantiating the model shown in Figures 2 and 3 as shown in Figure 4 where we can see that the URI of the ontology is retained by the URI of the instance of the Namespace class (which in the figure is renamed as “aktors URI” for convenience), whereas the prefix is represented by the datatype property ims:prefix; the term “Book” is an instance of the Concept class associated to its namespace via the ims:has-namespace property. In Table 2 we can see all the vocabularies adopted in DIRECT for the Resource Management and the Scientific Production areas.

The Contribution class represents every publication concerning the scientific production phase of the evaluation workflow. We can see that it is related to Concept via the ims:contribution-type property which can be instantiated as shown in Figure 4.

The Link class connects two resources via the ims: has-source and ims:has-target properties with a typed relationship realized throughout a concept connected to the link via the ims:relation property. This allows us for creating a typed relationship between two generic resources involved

Table 2: Namespaces and Prefixes of the vocabularies adopted in DIRECT for the Resource Management and the Scientific Production areas.

Prefix	Namespace	Description
aktors	http://www.aktors.org/ontology/portal#	Advanced Knowledge Technology reference ontology
bibo	http://purl.org/ontology/bibo/	Bibliographic ontology
dcterms	http://purl.org/dc/terms/	Dublin Core terms
foaf	http://xmlns.com/foaf/0.1/	Friend of a friend
gn	http://www.geonames.org/ontology#	GeoNames Ontology
ims	http://ims.dei.unipd.it/data/rdf/	DIRECT vocabulary terms
owl	http://www.w3.org/2002/07/owl#	OWL vocabulary terms
prov	http://www.w3.org/ns/prov#	The ontology supporting the interchange of provenance on the web
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF vocabulary terms
rdfs	http://www.w3.org/2000/01/rdf-schema#	RDF Schema
swrc	http://swrc.ontoware.org/ontology#	Semantic Web for Research Communities ontology
vann	http://purl.org/vocab/vann/	Vocabulary for annotating descriptions of vocabularies
vcard	http://www.w3.org/2006/vcard/ns#	vCard electronic business card profile defined by RFC 2426
xsd	http://www.w3.org/2001/XMLSchema#	XML Schema

in the evaluation workflow. We can instantiate the graph in Figure 3 in several ways; a first very simple example was shown in Figure 4, where we represented a term belonging to a vocabulary. This very example can be extended by representing a taxonomy of terms belonging to one or more vocabularies. In the upper part of Figure 5 we can see how the “Book” term presented above can be related throughout an “is-a” relation to the more general term “Publication”; please note that in this graph we omit literals and namespaces in order to focus on the Link class. So, in this case Link is instantiated by a generic “LinkA” resource, which relates two concepts, i.e. “Book” and “Publication”, via the ims:has-source and ims:has-target datatype properties. The datatype property ims:relation allows us to define the type of the relationship – “is-a” in this case – between the two associated concepts.

The concept “Book” is associated to the instance “contributionX” of Contribution by means of the ims: contribution-type property. Moreover, in the lower part of Figure 5 we can see another possible instantiation of the Link class; indeed, in this case it is used to say that a user (i.e. “userY”) is “author” of “contributionX”.

Link has two datatype properties: ims:score and ims:backward-score, which allow us to add weights on any typed relationship; both score and backward score are xsd:double in the interval $[0, 1]$. Indeed, we can establish a relation between user and concept with two scores on it in order to say that a user is expert in a given scientific topic. This lets us define expert profiles; for instance, we can say that “userY is an expert in Information Retrieval” where “userY” is an instance of the User class and “information retrieval” is a term defined as an instance of Concept; the score represents the strength of the relation between a user and a concept, and the backward score represents the strength of the relation between a concept and a user. This means that the relationship between User and Concept is not symmetric; for instance, we can say that “UserY” is an expert in “Information

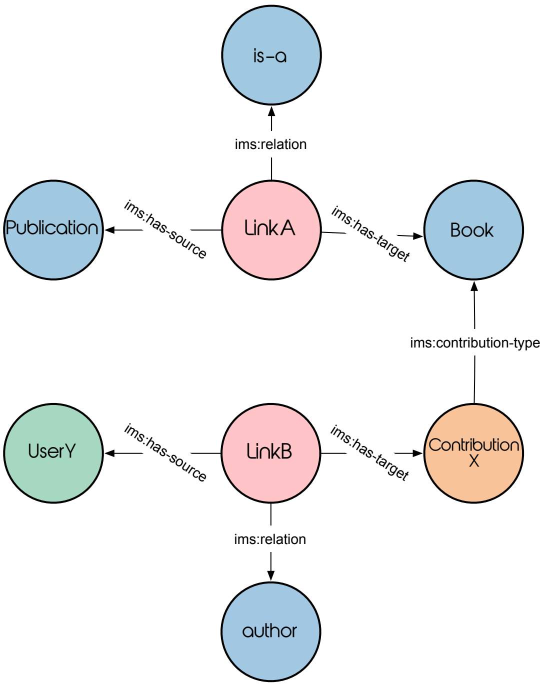


Figure 5: The RDF graph of an instantiation of the model shown in Figure 3.

Retrieval" with score 0.9 and this means that information retrieval is the main area of expertise for the user. On the other hand, there are people more expert in information retrieval than "UserY", so the backward score can be set to be only 0.1, and this would mean that "UserY" is just one of the experts in "Information Retrieval" and that we expect to find out other users with a higher expertise level (backward score) in the considered topic. The RDF graph of the user profile just described is shown in Figure 6.

In Figure 7 we can see another possible use of `Link`, in this case for representing the relationship between a contribution and a scientific topic. Indeed, semantic enrichment techniques are employed for extracting scientific topics from the data produced by the evaluation workflow and then relating them with pertinent contributions. We can see that "contributionX" is related to the scientific topic "Information Retrieval" via an `ims:relation` called "feature"; also in this case the typed relation between `contribution` and `concept` is weighted; the score is set to 0.7 meaning that "contributionX" mainly talks about "Information Retrieval", whereas the backward score is set to "0.3" meaning that among contributions about "Information Retrieval", "contributionX" is not one of the top relevant contributions.

5 Semantic Enrichment

In this section we describe several methods for semantically enriching IR experimental data by analysing unstructured data available in scientific publications. First, we propose a method to au-

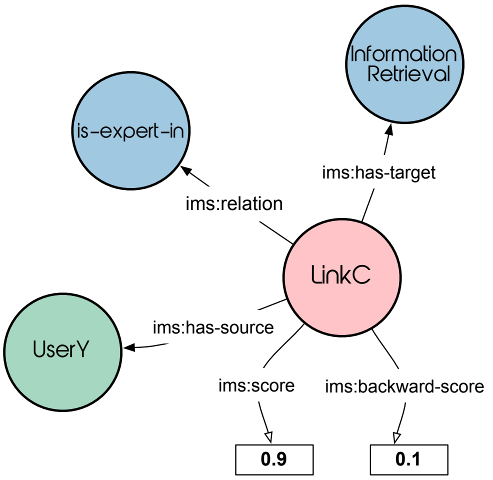


Figure 6: Instantiation for representing an expert profile.

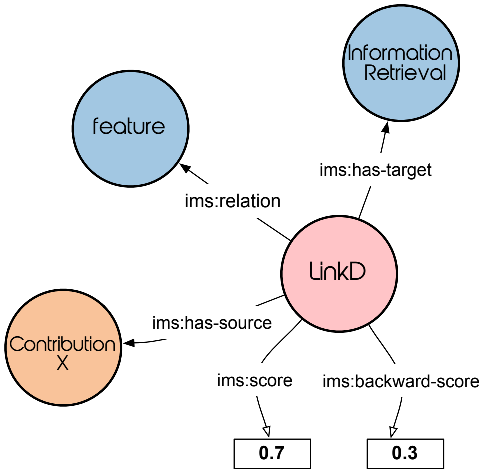


Figure 7: Instantiation for associating a contribution with a scientific topic.

tomatically extract expertise topics in Section 5.1. Then, these topics are enriched by grounding them on the Linked Open Data cloud in Section 5.2. An approach for expert profiling based on automatically extracted expertise topics is discussed in Section 5.3. In Section 5.4 we present several measures that can be used to rank experts for a given topic.

5.1 Expertise topic extraction

Topic-centric approaches for expert search put emphasis on the extraction of keyphrases that can succinctly describe expertise areas, also called expertise topics, using term extraction techniques [Bordea et al., 2012]. An advantage of a topic-centric approach is that topical profiles can be constructed directly from text, without the need for controlled vocabularies or manual identification of terms. Expertise topics are extracted from a domain specific corpus using the following approach. First, candidate expertise topics are discovered from text using a syntactic description for terms (i.e., nouns or noun phrases) and contextual patterns that insure that the candidates are coherent within the domain. A domain model is constructed using the method proposed in [Bordea et al., 2013b] and then noun phrases that include words from the domain model or that appear in their immediate context are selected as candidates. Candidate terms are further ranked using the scoring function s, defined as:

$$s(\tau) = |\tau| \log f(\tau) + \alpha e_\tau \tag{1}$$

where τ is the candidate string, $|\tau|$ is the number of words of candidate τ, f is its frequency in the corpus, and e_τ is the number of terms that embed the candidate string τ. The parameter α is used to linearly combine the embeddedness score e_τ and is empirically set to 3.5. The top ranked expertise topics extracted from Information Retrieval publications are presented in Table 3. These topics describe core concepts of the domain such as *search engine*, *IR system*, and *retrieval task*, as well as prominent subfields of the domain including *image retrieval*, *machine translation*, and *question answering*.

Only the best 20 expertise topics are stored for each document, ranking expertise topics based on their overall score $s(\tau)$ multiplied with their *tf-idf* score. In this way, each document is enriched with keyphrases, taking in consideration the quality of a term for the whole corpus in combination with its relevance for a particular document.

Rank	Expertise Topic
1	information retrieval
2	image retrieval
3	retrieval systems
4	search engine
5	information retrieval system
6	retrieval task
7	QA system
8	query expansion
9	language model
10	text retrieval
11	target language
12	training data
13	retrieval model
14	visual features
15	question answering system
16	Natural Language
17	machine translation
18	relevance feedback
19	IR system
20	annotation task

Table 3: Top 20 expertise topics extracted from IR scientific publications

5.2 Enriching expertise topics using Linked Open Data

Expertise topics can be used to provide links between information retrieval experimental data and other data sources. These links play an important role for cross-ontology question answering, large-scale inference and data integration [Ngonga Ngomo, 2012]. Also, existing work on using knowledge bases in combination with information retrieval techniques for semantic query expansion shows that background knowledge is a valuable resource for expert search [Demartini, 2007; Thiagarajan et al., 2008]. Additional background knowledge, as found on the Linked Open Data (LOD) cloud [2], can inform expert search at different stages. Manually curated concepts can be leveraged from a large number of domain-specific ontologies and thesauri. The LOD cloud contains a large number of

[2] Linked Data: http://linkeddata.org

datasets about scientific publications and patent descriptions that can be used as additional evidence of expertise.

A first step in the direction of exploiting this potential is to provide an entry point in the LOD cloud through DBpedia [3], one of the most widely connected datasources, that is often used as an entry point in the LOD cloud. Two naive but promising approaches for semantic term grounding on DBpedia are described and evaluated in section 6.2.1. Our goal is to associate as many terms as possible with a concept from the LOD cloud through DBpedia URIs. Where available, concept descriptions are collected as well and used in our system. Initially we find all candidate URIs using the following DBpedia URI pattern.

http://dbpedia.org/resource/*{DBpedia_label}*

Where *DBpedia_concept_label* is the expertise topic as extracted from our corpus. A large number of candidates are generated starting from a multi-word term as each word from the concept label can start with a letter in lower case or upper case in the DBpedia URI. Take for instance the expertise topic *"Natural Language Processing"*, all possible case variations are generated to obtain the following URI:

http://dbpedia.org/page/Natural_language_processing

To ensure that only DBpedia articles that describe an entity are associated with an expertise topic, we discard category articles and we consider only articles that match the *dbpedia-owl:title* or the final part of the candidate URI with the topic. Multiple morphological variations are extracted and stored from our corpus for each expertise topic. Each of these variations is used to search for a URI, increasing in this way the number of matches.

5.3 Expert profiling

Expertise profiles are brief descriptions of a person's expertise and interests, that can inform the selection of experts in different scenarios. Whenever we refer to an expertise profile throughout this work, we mean a topical profile. Although a person frequently writes about a subject area, the way they combine this area with other topics is more interesting, because a person is rarely an expert on every aspect of a topic [Mimno and McCallum, 2007]. A recent study [Berendsen et al., 2013] identified several requirements for an expertise profile including coherence, completeness, conciseness and diversity. The same study states that an important requirement for expertise topics is that they have to be at the right level of specificity.

Following [Balog and de Rijke, 2007], we define a topical profile of a candidate as a vector of expertise topics along with scores that measure the expertise of a candidate. Therefore, the expertise profile p of a researcher r is defined as:

$$p(r) = \{s(r,t_1), s(r,t_2), ..., s(r,t_n)\} \tag{2}$$

where t_1, t_2,...,t_n are the expertise topics extracted from a domain specific corpus.

[3] DBpedia:http://dbpedia.org/

A first step in constructing expertise profiles is to identify terms that are appropriate descriptors of expertise. A large number of expertise topics can be extracted for each document, but only the top ranked keyphrases are considered for expert profiling. Keyphrases are assigned to documents by combining the overall termhood rank of a candidate term with a measure of relevance for each document, as described in the previous section. Once a list of expertise topics is identified, we proceed to the second step of expert profiling, the assignment of scores to each expertise topic for a given expert. We rely on the notion of relevance, effectively used for document retrieval, to associated expertise topics with researchers. A researcher's interests and expertise are inferred based on their publications. Each expertise topic mentioned in one of these publications is assigned to their expertise profile using an adaptation of the standard information retrieval measure tf-idf [Baeza-Yates et al., 1999]. The set of documents authored by a researcher is aggregated in a virtual document, allowing us to compute the relevance of an expertise topic over this virtual document. An expertise topic is added in the expertise profile of a researcher using the following scoring function:

$$s_{ep}(r,t) = termhood(t) \cdot tfirf(t,r) \tag{3}$$

Where $s_{ep}(r,t)$ represents the score for an expertise topic t and a researcher r, $termhood(t)$ represents the score computed in Equation 1 for the topic t and $tfirf(t,r)$ stands for the tf-idf measure for the topic t on the aggregated document of researcher r. In this way, we construct profiles with terms that are representative for the domain as well as highly relevant for a given researcher.

5.4 Expert finding

Expert finding is the task of identifying the most knowledgeable person for a given expertise topic. In this task, several competent people have to be ranked based on their relative expertise on a given expertise topic. Documents written by a person can be used as an indirect evidence of expertise, assuming that an expert often mentions his areas of interest. We rely on the tf-irf measure described in the previous section to measure the relevance of a given expertise topic for a researcher. Each researcher is represented by an aggregated document that is constructed by concatenating all the documents authored by that person. Therefore, the relevance score $R(r,t)$ that measures the interest of a researcher r for a given topic t is defined as:

$$R(r,t) = tfirf(t,r) \tag{4}$$

Expertise is closely related to the notion of experience. The assumption is that the more a person works on a topic, the more knowledgeable they are. We estimate the experience of a researcher on a given topic based on the number of their publications that have the query as a keyphrase. Let $D_{r,t}$ be the set of documents authored by researcher r, that have the expertise topic t as a keyphrase. Then, the experience score $E(r,t)$ is defined as:

$$E(r,t) = |D_{r,t}| \tag{5}$$

where $|D_{r,t}|$ is the cardinality, or the total number of documents, in the set of documents $D_{r,t}$. It can be argued that it is not only the number of publications that indicates expertise, but the quality

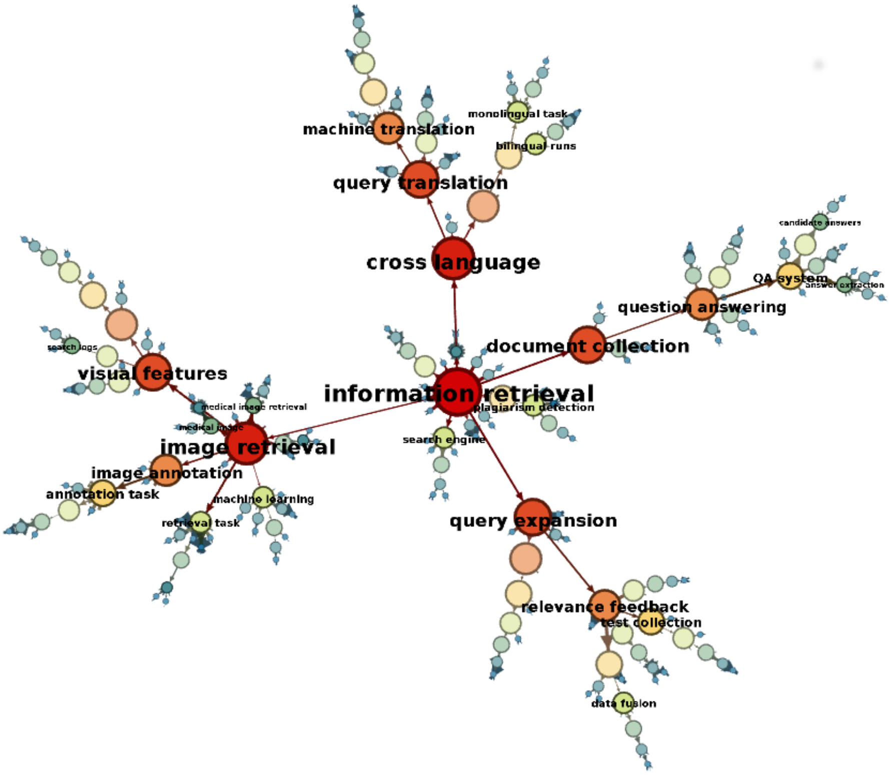


Figure 8: Topical hierarchy automatically constructed for the CLEF evaluation campaign

of those publications as well. We leave for future work the integration of publication impact in this score, measured using citation counts.

Relevance and expertise measure different aspects of expertise and can be combined to take advantage of both features as follows:

$$RE(r,t) = R(r,t) \cdot E(r,t) \tag{6}$$

Both the relevance score and the experience score rely on query occurrences alone. A topical hierarchy, similar to the one constructed in [Hooper et al., 2012], can provide valuable information for improving expert finding results. Take for example the topical hierarchy presented in Figure 8, that is automatically constructed using publications from the CLEF evaluation campaign [4]. When searching for experts in *image retrieval*, we can make use of the information that *image annotation* and *visual features* are closely related expertise topics that are subordinated to the topic of interest. In the same way, when searching experts on the expertise topic *question answering* we can use information about the subordinated terms *QA system* and *answer extraction*.

In the case that the subtopics of an expertise topic are known, we can evaluate the expertise of a person based on their knowledge of specialised fields. A previous study showed that experts

[4] CLEF: http://www.clef-initiative.eu

have increased knowledge at more specific category levels than novices [Tanaka and Taylor, 1991]. We introduce a novel measure for expertise called Area Coverage that measures whether an expert has in depth knowledge of an expertise topic, using an automatically constructed topical hierarchy. Let $Desc(t)$ be the set of descendants of a node t in the topical hierarchy, then the Area Coverage score $C(i,t)$ is defined as:

$$C(i,t) = \frac{|\{t' \in Desc(t) : t \in p(i)\}|}{|Desc(t)|} \tag{7}$$

where $p(i)$ is the profile of an individual i constructed using the method presented in the previous section. In other words, Area Coverage is defined as the proportion of a term's descendants that appear in the profile of a person. Finally, the score $REC(i,t)$ used to rank people for expert finding is defined as follows:

$$REC(i,t) = RE(i,t) \cdot C(i,t) \tag{8}$$

This score combines several performance indicators, measuring the expertise of a person based on the relevance of an expertise topic, the number of documents about the given topic, as well as his depth of knowledge of the field, called Area Coverage.

6 Experimental evaluation

In this section we present an empirical evaluation of the methods proposed in the previous section. We present our experimental setup in Section 6.1 and we discuss several experiments in Section 6.2.

6.1 Experimental setup

A dataset of scientific publications gathered from Information Retrieval conferences is described in Section 6.1.1. The baseline approaches used in our experiments are presented in Section 6.1.2, followed by a discussion of evaluation measures in Section 6.1.3.

6.1.1 Information Retrieval workshop dataset

Evaluating expert search systems remains a challenge, despite a number of data sets that have been made publicly available in recent years [Bailey et al., 2007; Balog et al., 2007; Soboroff et al., 2007]. Traditionally, relevance assessments for expert finding were gathered either through self-assessment or based on opinions of co-workers. On one hand, self-assessed expert profiles are subjective and incomplete, while on the other hand opinions of colleagues are biased towards their social and geographical network. We address these limitations by exploiting expertise data generated in a peer-review setting [Bordea et al., 2013a]. More specifically, we consider conference workshops in the related fields of information retrieval (IR), digital libraries (DL), and recommender systems (RS). About 25 thousand publications were gathered along with data about 60 workshops. Each workshop is associated in average with 15 experts and almost 500 expertise topics were manually extracted to describe these events.

To construct a test collection covering all of these research fields, we used the DBLP Computer Science Bibliography[5], a computer science bibliography website that tracks the most important journals and conference proceedings in computer science. Our initial motivations for constructing a test collection around DBLP were two-fold: (1) the fields of IR, DL, and RS are well-covered in DBLP, and (2) a special version of the DBLP data set, augmented with citation information, is available from the team behind ArnetMiner, which allows for investigations into the use of citation information for expert search.

To make the augmented DBLP collection suited to expert search evaluation, we need realistic topic descriptions as relevance judgments at the expert level. Workshops organized at major conferences covering the fields of IR, DL, and RS are used to collect relevance judgments. To identify relevant workshops, we visited the websites of the CIKM, ECDL, ECIR, IIiX, JCDL, RecSys, SIGIR, TPDL, WSDM, and WWW conferences, which have substantial portions of their program dedicated to IR, DL, and RS. We collect links to workshop websites for all workshops organized at those conferences between 2001 and 2012. This resulted in a list of 60 different workshops with websites that were still online at the time of writing[6].

As a starting point, a test collection covering the aforementioned fields was constructed by using the augmented DBLP data set released by the team behind ArnetMiner. This data set is a October 2010 crawl of of the DBLP data set containing 1,632,442 different papers with 2,327,450 citation relationships between papers in the data set[7]. As this augmented data set contains publications from all fields of computer science, we filtered out all publications not belonging to IR, DL, and RS by restricting the collection to publications in relevant journals, conferences, and workshops.

The list of relevant venues was created in two steps. First, we generated a list of *core venues* by extracting all papers published at conferences used for topic creation: CIKM, ECDL, ECIR, IIiX, JCDL, RecSys, SIGIR, TPDL, WSDM, and WWW. We select these conferences, because as hosts to the topic workshops, they are likely to be relevant venues for PC members to publish in. This resulted in a data set containing 9,046 different publications from these core venues. However, restricting ourselves to these venues alone means we could be missing out on experts that tend to publish more in journals and workshops. We therefore extend the list of core venues with other venues tracked by DBLP that also have substantial portions of their program dedicated to IR, DL, and RS. Venues that only feature incidental overlap with IR, such as the Semantic Web conference, were not included. We also excludes venues that did not have 5 publications or more in the augmented DBLP data set. While this does exclude the occasional on-topic publication in venues that are pre-dominantly about other topics, we believe that this strategy will cover the majority of relevant publications. This additional filtering step resulted in a final list of 78 *curated venues* (core plus additional)[8] covering a total of 24,690 publications.

In addition to citation information, the augmented DBLP data set is also extended with abstracts wherever available. However, the team behind ArnetMiner was only able to add abstracts for 33.7% of the 1.6 million publications (and 43.5% of the 24,690 publications in our test collection). We

[5] Available at http://dblp.uni-trier.de/, last accessed July 9, 2013.

[6] The list of 60 active workshops can be viewed at http://itlab.dbit.dk/~toine/?page_id=631.

[7] Available at http://arnetminer.org/DBLP_Citation, last accessed July 9, 2013.

[8] The list of curated active workshops is available at http://itlab.dbit.dk/~toine/?page_id=631.

therefore attempt to download the full-text versions of al 24,690 publications using Google Scholar. We constructed a search query consisting of the last name of the first author and the full title without surrounding quotes[9]. We then extract the download link from the top result returned by Google Scholar (if available). We were able to find downloads URLs for 14,823 of the 24,690 publications in our filtered DBLP data set for a recall of 60.04%, where recall is defined as the percentage of papers in our filtered DBLP data set that we could find download URLs for. While this is not as high as we would like, it does represent a substantial improvement over the percentage of abstracts present in the augmented DBLP data set. Moreover, a recall rate of 100% is impossible to achieve as tutorials, keynote abstracts, and even entire proceedings are typically not available online in full-text, but they are present in the DBLP data set.

Around 90.15% of these download URLs obtained in this manner were functional, which means we were able to download full-text publication files for 13,363 publications (or 54.12% of our entire curated data set). We performed a check of 100 randomly selected full-text files to see if these were indeed the publications we were looking for and achieved a precision of 97% on this sample. We therefore assume that the false positive rate of our approach is acceptably low.

6.1.2 Baseline approaches

The approaches proposed in this section are evaluated against two information retrieval methods for expert finding and expert profiling. Both methods model documents and expertise topics as bags of words and take a generative probabilistic approach, ranking expertise topics t by the probability $P(t|i)$ that they are generated by the individual i [Balog et al., 2009]. The same probability is used for ranking expertise topics in a person's profile, as well as for finding knowledgeable people for expert finding. The first model constructs a multinomial language model θ_i for each individual, over the vocabulary of documents authored by them. This is similar to our approach that computes the relevance of a topic for an individual on a document that aggregates all the documents authored by that person.

The assumption is that expertise topics are sampled independently from this multinomial distribution. Therefore, the probability $P(t|i)$ can be computed as:

$$P(t|i) = P(t|\theta_i) = \prod_{w \in t} P(w|\theta_i)^{n(w,t)} \qquad (9)$$

where $n(w,t)$ is the number of times the word w appears in the expertise topic t. Smoothing using collection word probabilities is applied to estimate $P(w|\theta_i)$. The smoothing parameters are estimated with an unsupervised method, using Dirichlet smoothing and the average number of words associated with people as the smoothing parameter.

The second model considered as baseline estimates a language model θ_d for each document from the set D_i of documents authored by the individual i. Words from an expertise topic t are sampled independently, summing the probabilities to generate an expertise topic for each of these

[9] A preliminary test on just the publications from the core venues showed that adding quotes around the publication title decreased recall from 80.3% to 70.86%.

documents. In this case, the probability $P(t|i)$ is calculated using the following equation:

$$P(t|i) = \sum_{d \in D_i} P(t|\theta_d) = \sum_{d \in D_i} \prod_{w \in t} P(w|\theta_d)^{n(w,t)} \tag{10}$$

Again, the probability $P(w|\theta_d)$ is estimated by using the same unsupervised smoothing method. In this case, the smoothing parameter for Dirichlet smoothing is the average document length in the corpus.

6.1.3 Evaluation measures

Given the tasks at hand, several evaluation measures for document retrieval can be used. The expert profiling and the expert finding tasks are evaluated based on the quality of ranked lists of expertise topics and of experts, respectively. From an evaluation point of view, this is not different from evaluating a ranked list of documents. The most basic evaluation measures used in information retrieval are precision and recall. These measure the proportion of retrieved documents that are relevant and the proportion of relevant documents that are retrieved, respectively. Other frequently used effectiveness measures include:

Precision at N (P@N) This is the precision computed when N results are retrieved, which is usually used to report early precision at top 5, 10, or 20 results.

Average Precision (AP) Precision is calculated for every retrieved relevant result and then averaged across all the results.

Reciprocal Rank (RR) This is the reciprocal of the first retrieved relevant document, which is defined as 0 when the output does not contain any relevant documents.

To get a more stable measurement of performance, these measures are commonly averaged over the number of queries. In our experiments, we report the values for the Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). In this setting, recall is less important than achieving a high precision for the top ranked results, because it is more important to recommend true experts than to find all experts in a field.

6.2 Experiments

In this section we discuss the results of several experiments related to semantic grounding of expertise topics (Section 6.2.1), expert profiling (Section 6.2.2), and expert finding (Section 6.2.3).

6.2.1 Semantic grounding of expertise topics

Two approaches for grounding expertise topics on DBpedia are evaluated in this section. The first approach (A1) matches a candidate DBpedia URI with an expertise topic, using the string as it appears in the corpus. The second approach (A2) makes use of the lemmatised form of the expertise

Approach	Precision	Recall	F-score
A1	0.96	0.93	0.94
A2	0.99	0.90	0.94

Table 4: Precision and recall for DBpedia URI extraction

topic. Stemming was also considered but this approach resulted in a decrease in performance, as stems are more ambiguous [10]. In order to evaluate our URI discovery approach, we build a small gold standard dataset by manually annotating 186 expertise topics with DBpedia URIs. First of all, we note that it is only about half of the analysed expertise topics that have a corresponding concept in DBpedia. This is because we are dealing with a general knowledge datasource that has a limited coverage of specialised technical domains.

Although both approaches achieve similar results in terms of F-score, the approach that makes use of lemmatisation (A2) achieves better precision, as can be seen in Table 4. To extract descriptions or definitions of concepts we rely on the *dbpedia-owl:abstract* property, or the *rdfs:comment* property in the absence of the former. For now we are interested in English definitions, therefore we consider triples tagged with the property *lang='en'* alone. Even though English descriptions are available for a larger number of topics, this tag is not always present. Therefore, we can only retrieve descriptions for a smaller number of topics. A manual analysis of matching errors showed that expertise topics that include an acronym (e.g. "NLG system" instead of "Natural Language Generation system") are more difficult to associate with a DBpedia concept, as often acronyms are ambiguous.

Other general purpose data sources, such as Freebase [11], or domain-specific data sources can be linked in a similar manner. A complex problem that we do not address in this work is the disambiguation of an expertise topic when multiple concepts from different domains can be matched. Usually, DBpedia provides a disambiguation page for such cases. In our implementation we did not analyse concepts that redirect to a disambiguation page, grounding only those expertise topics that are specific enough to be used in a single domain.

6.2.2 Expert profiling

The topic-centric approach (TC) for expert profiling proposed in Section 5.3 can be applied for expert profiling without the need for controlled vocabularies, as expertise topics are directly extracted from text. Instead, the language modelling approach used as a baseline in this section, can only be used on datasets where such resources are readily available. The results for the expert profiling task on the IR dataset are presented in Table 5.

Both language modelling approaches achieve better results than the topic-centric approach, with the LM2 approach outperforming the LM1 approach based on precision, but not when considering recall as an evaluation measure. Although the language modelling approach achieves better perfor-

[10] An approach based on a semantic web search engine that uses keyphrase search to find structured data was also considered, restricting the search to the DBPedia domain. The results were disappointing because only a limited number of retrieved results can be analysed. Often, the relevant DBpedia concept does not appear in the top results.

[11] http://www.freebase.com/

Measure	LM1	LM2	TC
MAP	0.1052	**0.1679**	0.0879
MRR	**0.3761**	0.3677	0.3364

Table 5: Expert profiling results for the language modelling approach (LM) and the topic centric approach (TC)

mance on our dataset, this method has the disadvantage that it requires manually identified expertise topics. Instead, the topic-centric approach proposed in this work achieves acceptable results while completely automating the extraction of expertise topics.

6.2.3 Expert finding

We compare several topic-centric methods for expert finding with two language-modelling baselines. The results for the expert finding task are presented in Table 6. The expert finding methods evaluated in this section include Experience (E), Relevance and Experience (RE) and Relevance, Experience and Area Coverage (REC). These methods are described in Equations 5, 6, and 8 respectively, in Section 5.4. The Area Coverage measure makes use of a topical hierarchy, therefore we automatically construct a topical hierarchy for Information Retrieval using the method proposed in [Hooper et al., 2012]. The resulting hierarchy has 4,000 nodes and 3,939 edges, and was constructed by considering all the co-occurrences between two expertise topics in a window of 5 words. An edge is added in the initial graph only if at least three different documents provide evidence for the relation.

Measure	LM1	LM2	E	RE	REC
MAP	0.0599	0.0402	0.1592	**0.1669**	0.1657
MRR	0.1454	0.1231	0.4056	**0.4141**	0.4120
P@5	0.0614	0.0485	0.1771	0.1771	**0.1783**

Table 6: Expert finding results for the language modelling approach (LM), Experience (E), Relevance and Experience (RE), and Relevance, Experience and Area Coverage (REC)

We note that topic-centric approaches achieve the best results for our information retrieval dataset. The experience of an individual measured by the number of documents written on a given topic is the most effective measure of expertise. Only slight improvements can be achieved by considering relevance as well in the RE score. Using a topical hierarchy by computing Area Coverage improves the precision at top 5, but not the overall precision and recall. In a second experiment, we compare topical hierarchies with hierarchical clustering based on the improvements that these structures bring to the task of expert finding. An agglomerative approach with complete linkage clustering is used [Day and Edelsbrunner, 1984], because this approach was shown to outperform other clustering methods when applied to hierarchy construction [Cimiano et al., 2004]. To make sure that the two approaches are comparable, we use the same number of nodes and the same similarity metric in both cases.

Measure	HC	TH
MAP	0.1581	**0.1657**
MRR	0.4052	**0.4120**
P@5	0.1643	**0.1783**

Table 7: Expert finding results using Area Coverage computed based on Hierarchical Clustering (HC) and Topical Hierarchy (TH)

The same list of expertise topics are used for clustering as for the topical hierarchy. The constructed dendrogram is converted in a hierarchy of expertise topics by labelling the resulting clusters. Each intermediate cluster in the hierarchy is labelled with the most frequent expertise topic. The agglomerative clustering algorithm merges two clusters at each step, which results in a large number of self-referring edges. These edges are resulted when the same label is identified for a merged cluster. For the purpose of our experiments, all such edges are ignored. Table 7 presents the results for the REC score described in Equation 8. The Area Coverage measure is computed using a hierarchy constructed through Hierarchical Clustering (method HC in the table) and using an algorithm for constructing Topical Hierarchies (called TH).

Computing Area Coverage using a topical hierarchy achieves better results for expert finding than using a hierarchical clustering algorithm. The improvements are stable for all the evaluation measures. Furthermore, a manual analysis of the constructed hierarchies showed that topical hierarchies are more intuitive, because the pruning algorithm favours closely related terms. Hierarchies constructed through clustering are more difficult to understand, as they rely on similarities in a high-dimensional space, which are more difficult to trace.

7 Saffron. An Expert Search system for exploration and discovery of experts and expertise

The techniques proposed in this work are integrated in Saffron [12], a system that provides insights in a research community or organisation by analysing their main topics of investigation and the individuals associated with them [Monaghan et al., 2010]. Currently, Saffron analyses mainly Computer Science areas, including Natural Language Processing, Information Retrieval, and Semantic Web, but there is an on going effort to extend this to other research domains. We start by giving an overview of the Saffron architecture and then we describe in more detail the main components of the architecture and the connections between them.

Saffron is developed by DERI's Unit for Natural Language Processing (UNLP)[13], and was applied for several domains and application scenarios, including two organisations as well as several academic conferences and online communities. Technically, Saffron is designed to fulfil three main non-functional requirements:

[12] Saffron:http://saffron.deri.ie/

[13] UNLP:http://www.deri.ie/nlp

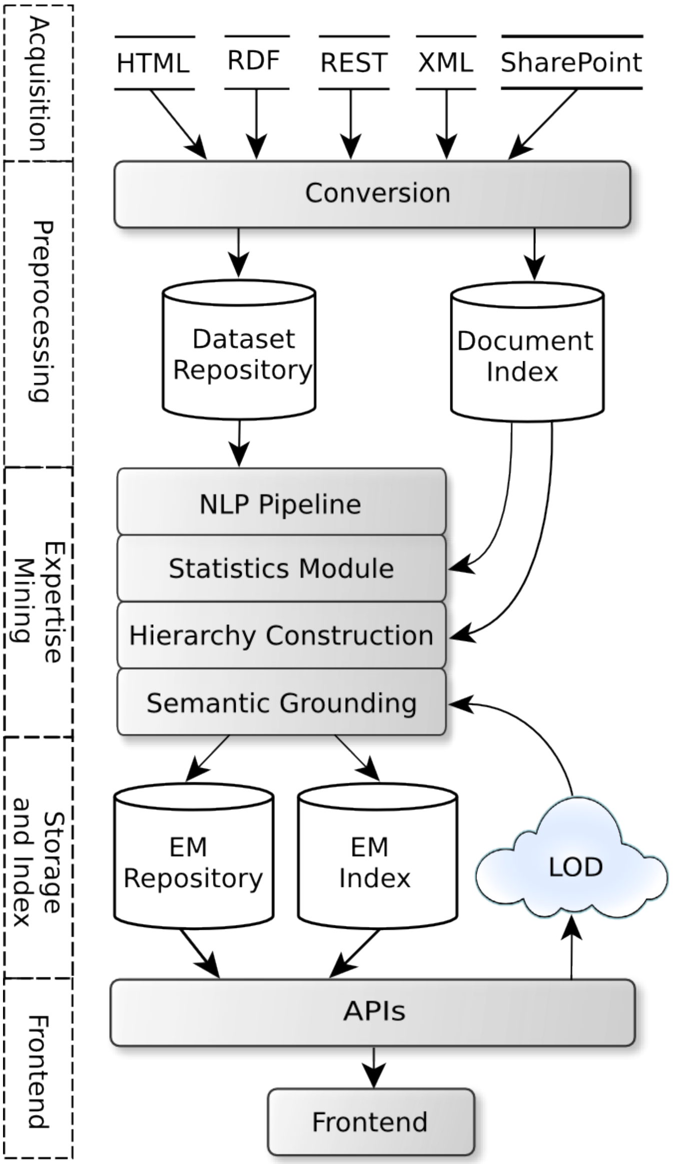


Figure 9: Overview of the Saffron infrastructure stack

1. Ability to cope with datasets acquired from different sources and represented in various formats

2. Minimal need for human interaction

3. Adaptive extraction algorithms that can cover a wide range of domains

These requirements guided the design of the Saffron infrastructure that can be seen in Figure

9. The first level of the infrastructure deals with the acquisition of documents and people associated with them for a given domain.The next layer is the preprocessing layer that converts and stores metadata and indexes documents. The Expertise Mining layer contains the core components of the system. Finally, results are stored, indexed and then made available for further integration with other knowledge management applications through APIs. The final layer in the infrastructure is the Saffron interface.

Data acquisition and data preprocessing

The first layer of the infrastructure is the acquisition of a suitable dataset about individuals and documents authored by them. Relatively clean metadata about documents and associated people is already available for several domains. This metadata is most often represented in XML, but there is an increasing number of datasources available in RDF, through a public SPARQL endpoint. The CL dataset makes use of the XML format to represent data about scientific publications, researchers and academic events, while the SW dataset represents the same types of information by making use of standardised vocabularies in RDF. RESTful Web Services are another way to provide access to expertise datasets, and this is the case of the DIRECT Infrastructure [14].

In the case of academic events such as conferences and workshops, it is often the case that information about publications and authors is not readily available, and has to be collected from dedicated HTML websites through web scrapping. In the enterprise environment, most information about documents and organisation members is not public, and has to be accessed from content management tools, such as SharePoint. Depending on the dataset, people are identified using methods that are more or less ambiguous. Many of these datasets use personal names to identify the author of a document, therefore a name disambiguation and name consolidation component is required. Saffron uses a popular open source relational database, MySQL, as backend, and Lucene, an information retrieval library, is used for indexing full-content documents.

Expertise Mining components

This layer addresses the core tasks of Expertise Mining, including expertise topic extraction, topical hierarchy construction, expert profiling, and expert finding. Candidate terms identified using a NLP pipeline based on the GATE natural language processing framework [Cunningham et al., 2002] and the ANNIE information extraction system, included in the standard GATE distribution. The NLP pipeline is depicted in more detail in Figure 10. We use several off-the-shelf components available in ANNIE for text tokenisation, sentence splitting and part-of-speech tagging. In the figure, components provided by GATE are represented in a lighter shade than the last two components, which are customized components for Expertise Mining.

A gazetteer, called DM Gazetteer[15] in the figure, annotates domain model words extracted from a domain-specific corpus. Saffron identifies candidate terms using extraction patterns constructed starting from a domain model. Finally, candidate terms are annotated using a finite state transducer,

[14] DIRECT: http://direct.dei.unipd.it/

[15] The acronym DM stands for Domain Model

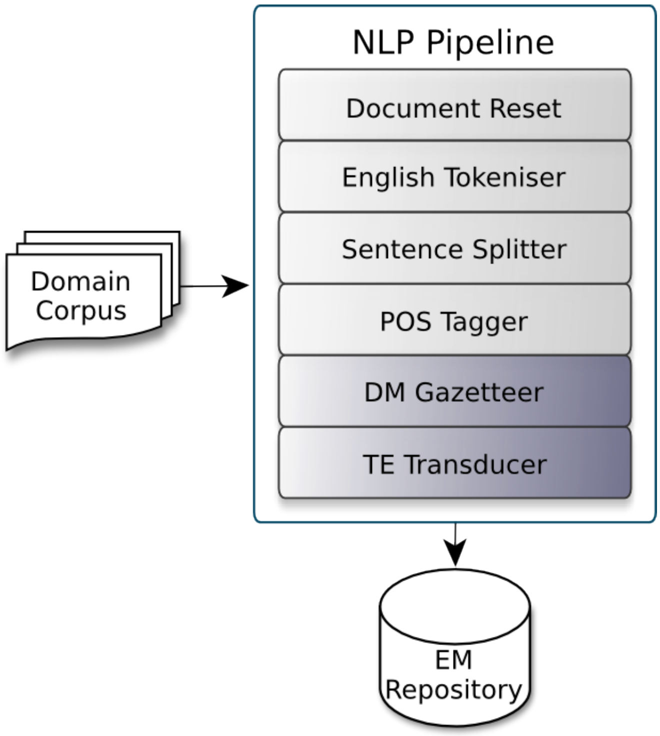


Figure 10: Overview of the NLP Pipeline for expertise topic extraction

called TE transducer[16]. Several rules are used to select terms that contain a word from the domain model, or terms that are introduced by one. The candidates terms are stored in the Expertise Mining (EM) Repository for further analysis.

The Statistics Module is responsible for ranking and filtering candidate terms to identify expertise topics. Word occurrences, as well as relevance measures for expert finding and expert profiling, are computed using a Lucene index. The relations between expertise topics are identified by the Hierarchy Construction component, using a graph-based algorithm for constructing topical hierarchies. Again, Saffron relies on the Lucene index to measure co-occurrences between two expertise topics, making use of the span search functionality available in Lucene. This functionality allows us to perform proximity searches within a predefined window of words. The final core Expertise Mining component is Semantic Grounding, that is responsible for identifying DBpedia URIs and descriptions of expertise topics. In this way Saffron provides an entry point in the Linked Open Data (LOD) cloud, as well as descriptions for expertise topics that can be directly used by the end user.

Expertise storage and index

The next layer of the architecture, the Storage and Index layer prepares the data for high performance access by other applications. This is done either directly through APIs or by making the data

[16] The acronym TE stands for Term Extraction

available on the LOD cloud through a SPARQL endpoint. The EM Repository is a MySQL based solution for storing data. The Expertise Mining results are also indexed by the EM Index component, using Solr, a higly scalable enterprise search engine.

Frontend

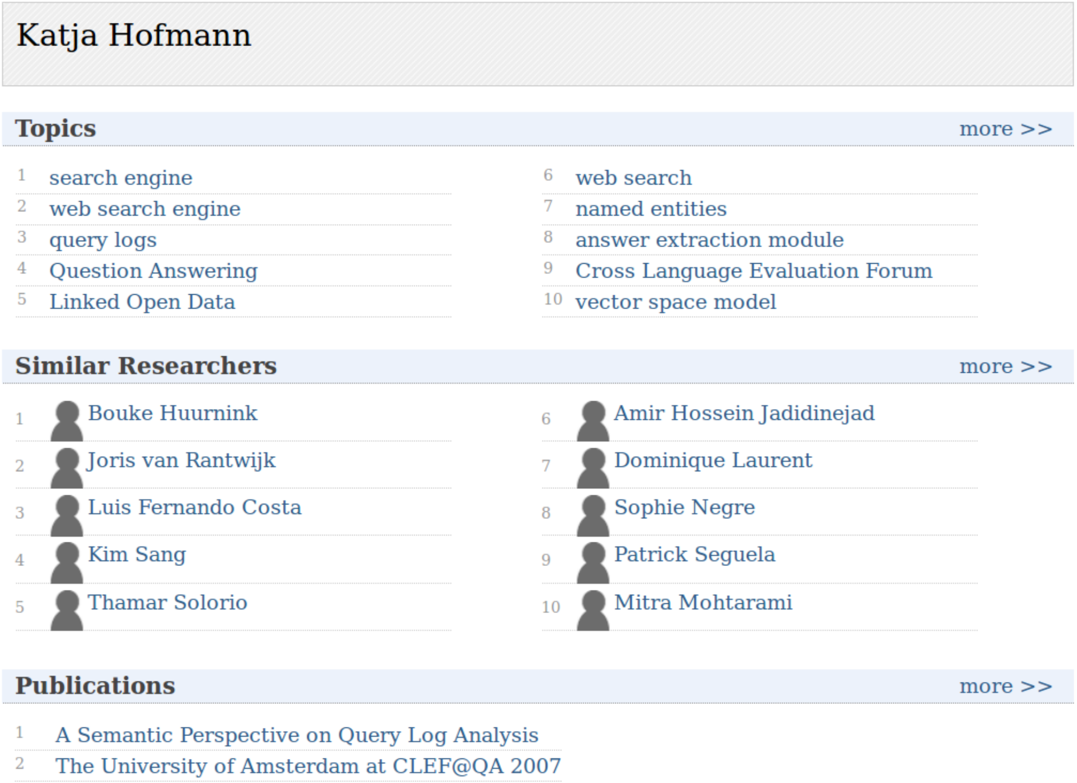
Saffron supports users that have different information needs and varying levels of knowledge of a field to search for experts in a community or organisation. Several scenarios are considered, including novice members trying to establish connections, expert members looking for collaborators, as well as outsiders interested in an overview of the main areas of investigation or activity. The main functionalities of the system allow search and discovery of expertise topics, experts and expert profiles. Saffron provides keyphrase based search, enhanced with an autocomplete feature, for searching experts and expertise topics. Users that are not familiar with the domain are guided by a list of representative expertise topics that are listed on the start page. Table 8 shows the top ranked topics for three instances of Saffron, for Computational Linguistics, Semantic Web, and the CLEF initiative. Users can select any of these expertise topics to find out more information about documents that mention them and associated experts.

CL	SW	CLEF
training data	Semantic Web	query expansion
target language	Web services	retrieval system
speech recognition	search engine	search engine
spoken language	knowledge base	text retrieval
word sense disambiguation	data set	retrieval task
source language	web pages	target language
web services	information retrieval	relevance feedback
statistical machine translation	social network	retrieval results
user interface	data sources	source language
sign language	user interface	Question Answering

Table 8: Topics from the start page of the Saffron interface for Computational Linguistics (CL), Semantic Web (SW), and the CLEF initiative

The Saffron interface is designed around three types of pages, based on the type of resource they describe: topical page, expert page, and document page. A topical page shows additional information about a topic such as occurrence trends across the time, a description of the topic, and related topics. Additional information includes a list of main experts that work on the topic, and the most relevant documents. An expert page presents the profile of that person, a list of similar experts that can be used if the expert cannot be contacted, and a list of documents authored by the expert. Figure 11 shows an extract from the expert page of a researcher in the Semantic Web community. Finally, a document page shows the authors of the document and several topics that describe the content of the document. Saffron maintains a web of connections between topics, experts and documents, enabling users to navigate from one type of resource to another.

The system has been applied inside organisations as well as at conferences. Further usability



Katja Hofmann

Topics more >>

1 search engine
2 web search engine
3 query logs
4 Question Answering
5 Linked Open Data
6 web search
7 named entities
8 answer extraction module
9 Cross Language Evaluation Forum
10 vector space model

Similar Researchers more >>

1 Bouke Huurnink
2 Joris van Rantwijk
3 Luis Fernando Costa
4 Kim Sang
5 Thamar Solorio
6 Amir Hossein Jadidinejad
7 Dominique Laurent
8 Sophie Negre
9 Patrick Seguela
10 Mitra Mohtarami

Publications more >>

1 A Semantic Perspective on Query Log Analysis
2 The University of Amsterdam at CLEF@QA 2007

Figure 11: Saffron interface for an expert profile at CLEF

studies are required, but this is beyond the scope of this work. Future work will integrate topical hierarchies in the frontend, and use them as a tool for browsing documents and experts.

8 Conclusion

In this report we discussed the data modelling and the semantic enrichment of information retrieval experimental data, as produced by large-scale evaluation campaigns. We described in detail the evaluation workflow used for information access systems and we proposed a Linked Data based data model for two areas of the workflow, namely resource management and scientific production. Unstructured data in the form of scientific publications was used to inform the extraction of various types of semantic enrichment. Expertise topics were automatically extracted and used to describe documents and to create expert profiles. Several topic-centric measures for expert finding were proposed, allowing users to identify knowledgeable members of the community. In this way we created new relationships among existing data, allowing a more meaningful interaction with experimental data.

We introduced an evaluation dataset for expert search in Information Retrieval, relying on scientific publications available online and on implicit expertise information about workshop committee members. Our experiments show that expertise profiles can be constructed using automatically extracted expertise topics and that topic-centric approaches for expert finding outperform state of the art language modelling approaches on the considered dataset.

References

Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., and Silvello, G. (2012a). Directions: design and specification of an ir evaluation infrastructure. *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*, pages 88–99.

Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., and Silvello, G. (2012b). DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In Catarci, T., Forner, P., Hiemstra, D., Peñas, A., and Santucci, G., editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012)*, pages 88–99. Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany.

Agosti, M., Di Nunzio, G. M., Dussin, M., and Ferro, N. (2010). 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In Sakay, T., Sanderson, M., and Webber, W., editors, *Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010)*, pages 16–24. National Institute of Informatics, Tokyo, Japan.

Agosti, M., Di Nunzio, G. M., and Ferro, N. (2006). A data curation approach to support in-depth evaluation studies. In *Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006)*, pages 65–68. `http://ucdata.berkeley.edu:7101/projects/sigir2006/papers/pdf-final/MLIA-2.pdf`.

Agosti, M. and Ferro, N. (2009). Towards an Evaluation Infrastructure for DL Performance Evaluation. In *Evaluation of digital libraries: An insight to useful applications and methods*, pages 93–120.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bailey, P., de Vries, A. P., Craswell, N., and Soboroff, I. (2007). Overview of the TREC 2007 Enterprise Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC)*.

Balog, K., Azzopardi, L., and de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1–19.

Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., and van den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 551–558, New York, NY, USA. ACM.

Balog, K. and de Rijke, M. (2007). Determining expert profiles (with an application to expert finding). In *proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 2007)*.

Balog, K., de Rijke, M., and Azzopardi, L. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research* *and development in information retrieval - SIGIR '06*, pages 43–50, New York, New York, USA. ACM Press.

Berendsen, R., Balog, K., Bogers, T., van den Bosch, A., and de Rijke, M. (2013). On the assessment of expertise profiles. *Journal of the American Society for Information Science and Technology (JASIST)*.

Bordea, G., Bogers, T., and Buitelaar, P. (2013a). Benchmarking domain-specific expert search using workshop program committees. In *Workshop on Computational Scientometrics: Theory and Applications, at CIKM*.

Bordea, G., Kirrane, S., Buitelaar, P., and Pereira, B. O. (2012). Expertise mining for enterprise content management. In *The International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey*, pages 3495–3498.

Bordea, G., Polajnar, T., and Buitelaar, P. (2013b). Domain-independent term extraction through domain modelling. In *10th International Conference on Terminology and Artificial Intelligence*.

Bowers, S. (2012). Scientific workflow, provenance, and data modeling challenges and approaches. *Journal on Data Semantics*, 1(1):19–30.

Buckley, C. and Voorhees, E. M. (2000). Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 33–40.

Buettcher, S., Clarke, C. L. A., and Cormack, G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, Cambridge (MA), USA.

Buneman, P. (2013). The providence of provenance. pages 7–12.

Campbell, C. S., Maglio, P. P., Cozzi, A., and Dom, B. (2003). Expertise identification using email communications. In *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 528–531, New Orleans, LA.

Cheney, J., Chiticariu, L., and Tan, W. C. (2009). Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1(4):379–474.

Cimiano, P., Hotho, A., and Staab, S. (2004). Comparing conceptual, divise and agglomerative clustering for learning taxonomies from text. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 435–439.

Cleverdon, C. W. (1997). The Cranfield Tests on Index Languages Devices. In *Readings in Information Retrieval*, pages 47–60.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Day, W. H. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24.

Demartini, G. (2007). Finding experts using wikipedia. In *Proceedings of the Workshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007*, pages 33–41.

Di Nunzio, G. M. and Ferro, N. (2005). DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns. In Ioannidis, Y., Schek, H.-J., and Weikum, G., editors, *Proc. 8th DELOS Thematic Workshop on Future Digital Library Management Systems: System Architecture and Information Access*, pages 58–63. http://dbis.cs.unibas.ch/delos_website/delos-dagstuhl-handout-all.pdf [last visited 2007, March 23].

Draganidis, F. and Metzas, G. (2006). Competency based management: A review of systems and approaches. *Information Management and Computer Security*, 14(1):51–64.

Dussin, M. and Ferro, N. (2009). Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., and Tsakonas, G., editors, *Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009)*, pages 63–74. Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany.

European Union (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data.* Printed by Osmotica.it, Final report of the High level Expert Group on Scientific Data.

Harman, D. K. (2011). *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA.

Hooper, C. J., Marie, N., and Kalampokis, E. (2012). Dissecting the butterfly: representation of disciplines publishing at the web science conference series. In Contractor, N. S., Uzzi, B., Macy, M. W., and Nejdl, W., editors, *WebSci*, pages 137–140. ACM.

Hull, D. A. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338.

Macdonald, C. and Ounis, I. (2006). Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 387–396, New York, NY, USA. ACM.

Maybury, M. (2006). Expert finding systems. Technical Report MTR 06B000040, MITRE Corporation.

Mimno, D. and McCallum, A. (2007). Expertise Modeling for Matching Papers with Reviewers. In *SIGKDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 500–509.

Monaghan, F., Bordea, G., Samp, K., and Buitelaar, P. (2010). Exploring your research: Sprinkling some saffron on semantic web dog food. In *Semantic Web Challenge at the International Semantic Web Conference*.

Ngonga Ngomo, A.-C. (2012). On link discovery using a hybrid approach. *Journal on Data Semantics*, 1(4):203–217.

Petkova, D. and Croft, W. B. (2006). Hierarchical language models for expert finding in enterprise corpora. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 599–608, Washington, DC, USA. IEEE Computer Society.

Rodriguez, M. A. and Bollen, J. (2008). An Algorithm to Determine Peer-Reviewers. In *'08: Proceedings of the Seventeenth International Conference on Information and Knowledge Management*, pages 319–328. ACM.

Sakai, T. (2006). Evaluating Evaluation Metrics based on the Bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 525–532.

Savoy, J. (1997). Statistical Inference in Retrieval Effectiveness Evaluation. *Information Processing & Management*, 33(44):495–512.

Serdyukov, P., Taylor, M., Vinay, V., Richardson, M., and White, R. (2011). Automatic people tagging for expertise profiling in the enterprise. In Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., and Mudoch, V., editors, *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 399–410. Springer Berlin Heidelberg.

Soboroff, I., de Vries, A. P., and Craswell, N. (2007). Overview of the trec 2006 enterprise track. In *The fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.

Tanaka, J. W. and Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482.

Thiagarajan, R., Manjunath, G., and Stumptner, M. (2008). *Finding experts by semantic matching of user profiles*. PhD thesis, CEUR-WS.

W3C (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax – W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-concepts/ [last visited 2007, March 23].

W3C (2009a). SKOS Simple Knowledge Organization System Primer – W3C Working Group Note 18 August 2009. http://www.w3.org/TR/skos-primer.

W3C (2009b). SKOS Simple Knowledge Organization System Reference – W3C Recommendation 18 August 2009. http://www.w3.org/TR/skos-reference.