Expertise Modelling in Community-driven Knowledge Curation Platforms

Hasti Ziaimatin, Tudor Groza and Jane Hunter
e-Research Group, School of ITEE,
The University of Queensland, Australia
{h.ziaimatin, tudor.groza, j.hunter}@uq.edu.au

Abstract

Expertise modelling has been the subject of extensive research in two main disciplines - Information Retrieval (IR) and Social Network Analysis (SNA). Both IR and SNA techniques build the expertise model through a document-centric approach providing a macro-perspective on the knowledge emerging from large corpus of static documents. With the emergence of the Web of Data, there has been a significant shift from static to evolving documents, characterized by micro-contributions. Thus, the existing macro-perspective is no longer sufficient to track the evolution of both knowledge and expertise. The aim of this research is to provide a comprehensive, domain-agnostic model for expertise profiling in the context of dynamic, living documents and evolving knowledge bases. Our approach combines: (i) a fine-grained provenance model, (ii) weighted mappings of Linked Data concepts to expertise profiles, via the application of IR-inspired techniques on micro-contributions, and (iii) collaboration network analysis - to create, refine and enrich expertise profiles in community-centred environments, based on the relationships between networks of collaborators.

Keywords: Micro-contributions, expertise profile, annotation, ontology

1 Introduction

Acquiring and managing expertise profiles represents a major challenge in any organization, as often, the successful completion of a task depends on finding the most appropriate individual to perform it. Furthermore, the use of expertise profiles to identify, acknowledge and recommend experts from within an online community, motivates additional participants to contribute to the community knowledge base. This collaborative input is vital to the capture and integration of diverse viewpoints and the efficient assembly of an extensive body of knowledge[1].

In particular, many scientific research environments are increasingly dynamic and subject to rapid evolution of knowledge. Major scientific challenges such as global pandemics require teams of collaborators with expertise from a wide range of domains and disciplines. Better "expertise finders" would help identify the optimum set of researchers for a critical scientific challenge at any given time.

The topic of expertise modelling has been the subject of extensive research in two main disciplines: information retrieval (IR) and social network analysis (SNA). From the IR perspective, static documents authored by individuals (e.g., publications, reports) can be represented as bags-of-words (BOW) or as bags-of-concepts (BOC). The actual expertise identification is done by associating individual profiles to weighted BOWs or BOCs either by ranking candidates based on their similarities to a given topic or by searching for co-occurrences of both the individual and the given topic, in the set of supporting documents. Such associations can then be used to compute semantic similarities between expertise profiles (Thiagarajan, Manjunath and Stumptner October 2008). From the SNA perspective, expertise profiling is done by considering the graphs connecting individuals in different contexts, and inferring their expertise from the shared domain-specific topics (Zhang, Tang and Li 2007). Both IR and SNA techniques build the expertise model through a document-centric approach that provides only a macro-perspective on the knowledge emerging from the documents (due to their static, final nature, i.e., once written, the documents remain forever in the same form). However, the content of living documents changes via *micro-contributions* made by individuals, thus making this macro perspective no longer adequate for monitoring changes in either the knowledge or the expertise.

With the emergence of the Semantic Web (Lee, Hendler and Lassila 2001) and Web 2.0 (O'Reilly and Musser 2006), there has been a significant shift from static documents to evolving documents. This trend has followed in the scientific publishing process in some scientific communities. There has been a shift from the traditional document-centric approach towards a finer-grained contribution-oriented approach in which hypotheses or domain-related innovations (in form of short statements) replace the publications. Examples of this new trend can be seen via nano-publications (Mons and Velterop 2009) or liquid publications (Casati, Giunchiglia and Marchese 2007). In this new setting, mapping such micro-contributions to expertise will be essential in order to support the development of reputation metrics. Similarly, Wikis or knowledge bases (e.g., AlzSWAN[2] or SKELETOME[3] from the biomedical domain) also support this shift by enabling authors to incrementally refine the content of the embedded documents to reflect the latest advances in knowledge in the field. For example, AlzSWAN captures and manages

[1] http://www.nature.com/ng/journal/v40/n9/full/ng.f.217.html

[2] http://www.alzforum.org/res/adh/swan/default.asp

[3] http://skeletome.metadata.net/skeletome

hypotheses, arguments and counter-arguments in the Alzheimer's disease domain. SKELETOME supports discussion statements on skeletal dysplasias (see Fig. 1). While Wikipedia allows authors to state opinions and raise issues in the Discussion pages. These content increments, or *micro-contributions*, give the knowledge captured within the environment a dynamic character. Hence, generating expertise profiles from this constantly evolving knowledge-base raises a new and different set of challenges.

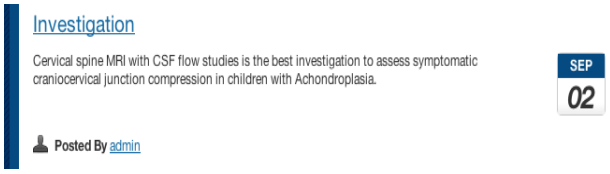

Fig. 1 Example of micro-contribution in SKELETOME

The main innovation of our research lies in the acquisition and management of the temporal and dynamic characteristics of expertise. Tracking the evolution of micro-contributions enables us to monitor the activity performed by individuals, which in turn, provides a way to show not only the change in personal interests over time, but also the maturation process (similar to some extent to the maturation process of scientific hypotheses, from simple ideas to scientifically proven facts) of an expert's knowledge. Moreover, by using well-grounded concepts from widely adopted vocabularies or ontologies, (published in the Linked Data Cloud), we facilitate easy mapping of individual expertise profiles to specific topics and comparison of profiles between individuals. As a result, the overhead imposed by performing co-reference entity resolution (to consolidate the expertise concepts to a shared understanding) is reduced to a minimum.

An additional significant consideration in collaborative knowledge-bases is the quality of the contributions. Traditional publications rely on peer review to maintain quality. In the context of scientific wikis, the scale and changeability of contributions makes conventional peer review difficult to scale[4]. Therefore, in our proposed approach, the quality of micro-contributions will be determined indirectly based on their persistence, extent of modification/correction or even deletion. These changes will be used to refine and adapt expert profiles over time.

In this paper, we propose a framework that is able to track Linked Data domain concepts in living documents using a fine-grained provenance ontology and assign them to an individual's expertise profile using IR-inspired techniques combined with an analysis of an associated and relevant collaboration/co-authorship network (e.g., Biomed Experts[5]). Firstly, we introduce an ontology that combines fine-grained provenance with versioning and change management to capture the dynamics of the knowledge/topics present in micro-contributions. Secondly, we lay the foundation for using instances of this ontology to build expertise profiles. Thirdly we implement and evaluate our approach through a system that recalculates and updates expertise profiles on the fly from a biomedical knowledge-base that changes over time. Finally we take into account the rate and type of change of the micro-contributions over time, to provide a measure of quality or validity of the micro-contributions, and adjust the expert profiles accordingly.

The remainder of the paper is structured as follows. In Section 2, we review related work. Section 3 describes the fine-grained provenance ontology(ies) that we have developed and our methodology for generating expertise profiles. Section 4 describes the evaluation process and Section 5 provides a discussion and conclusion.

2 Related Work

Expertise profiling is an active research topic in a wide variety of applications and domains, including bio-medical, scientific, education. In this section we present a brief overview of the related efforts, with particular accent on the Information Retrieval and the Semantic Web domains.

The two most popular and well performing types of approaches in TREC[6] (Text Retrieval Conference) expert search task are profile-centric and document-centric approaches. These studies use the co-occurrence model and techniques such as Bag-of-Words or Bag-of-Concepts on documents that are typically large and rich in content. Often a weighted, multiple-sized, window-based approach in an information retrieval model is used for association discovery (Zhu, Song and Rüger 2009) or the effectiveness of exploiting the dependencies between query terms for expert finding is proved (Yang and Zhang 2010). Other studies present solutions through effective use of ontologies and techniques such as spreading to link additional related terms to a user profile by referring to an ontology (Wordnet or Wikipedia) (Thiagarajan, Manjunath and Stumptner October 2008).

The *SubSift*[7] software, developed by the Institute for Learning and Research Technology (ILRT) at the University of Bristol, is a family of RESTful[8] Web services for profiling and matching text. SubSift (short for submission sifting) was originally designed to match submitted conference or journal papers to potential peer reviewers, based on the similarity between the paper's abstract and the reviewer's publications as found in online bibliographic databases. In this context, the software has already been used to support several major data mining conferences (Price, Flach, Spiegler, Bailey and Rogers 2010). SubSift uses traditional IR techniques such as TF-IDF, bag-of-words (BOW) and vector based modelling to profile and compare collections of documents.

Such traditional techniques work well with large corpuses as word occurrence is high and frequency is sufficient to capture the semantics of the document. However, when dealing with shorter texts such as micro-contributions within evolving knowledge bases, these traditional techniques are inadequate and unreliable. Their heavy dependency on statistical techniques (e.g., TF-IDF) cannot be applied to micro-contributions, as

[4] http://www.nature.com/ng/journal/v40/n9/full/ng.f.217.html

[5] http://www.biomedexperts.com/Portal.aspx

[6] http://trec.nist.gov/

[7] http://subsift.ilrt.bris.ac.uk

[8] http://www.ibm.com/developerworks/webservices/library/ws-restful/

such contributions don't provide sufficient context to capture the complete knowledge.

The *ExpertFinder* framework uses and extends existing vocabularies that have attracted a considerable user community already such as FOAF, SIOC, SKOS and DublinCore (Aleman-Meza, Bojars, Boley, Breslin, Mochol, Nixon, Polleres and Zhdanova 2007). Algorithms are also proposed for building expertise profiles using Wikipedia by searching for experts via the content of Wikipedia and its users, as well as techniques that use semantics for disambiguation and search extension (Demartini 2007). We intend to leverage these prior efforts to enable the integration of expertise profiles via a shared understanding based on widely adopted vocabularies and ontologies. This approach will also lead to a seamless aggregation of communities of experts.

As more and more Web users participate in online discussions and micro-blogging, a number of studies have emerged, which focus on aspects such as content recommendation and discovery of users' topics of interest, especially in Twitter. Early results in discovering Twitter users' topics of interest are proposed by examining, disambiguating and categorizing entities mentioned in their tweets using a knowledge base. A topic profile is then developed, by discerning the categories that appear most frequently and that cover all of the entities (Michelson and Macskassy 2010).

WikiGenes[9] combines a dynamic collaborative knowledge base for the life sciences with explicit authorship. Authorship tracking technology enables users to directly identify the source of every word. The rationale behind *WikiGenes* is to provide a platform for the scientific community to collect, communicate and evaluate knowledge about genes, chemicals, diseases and other biomedical concepts in a bottom-up approach. WikiGenes links every contribution to its author, as this link is essential to assess origin, authority and reliability of information. This is especially important in the wiki model, with its dynamic content and large number of authors (Hoffmann 2008). Although *WikiGenes* links every contribution to its author, it doesn't associate authors with profiles. More importantly, it doesn't perform semantic analysis on the content of contributions for extracting expertise.

The feasibility of linking individual tweets with news articles has also been analysed for enriching and contextualizing the semantics of user activities on Twitter to generate valuable user profiles for the Social Web (Abel, Gao, Houben and Tao 2011). This analysis has revealed that the exploitation of tweet-news relation has significant impact on user modelling and allows for the construction of more meaningful representations of Twitter activities. However, as with other traditional IR methods, this study applies bags-of-words (BOW) and TF-IDF methods for establishing similarity between tweets and news articles. Such traditional techniques work well with large corpora as word occurrence is high and frequency is sufficient to capture the semantics of the document. However, when dealing with shorter texts such as micro-contributions within evolving knowledge bases, which don't offer sufficient context to capture the encapsulated knowledge, these traditional techniques are no longer reliable.

The *Saffron*[10] application provides users with a personalised view of the most important expertise topics, researchers and publications, by combining structured data from various sources on the Web with information extracted from unstructured documents using Natural Language Processing techniques (Monaghan, Bordea, Samp and Buitelaar 2010). However, as with other outlined studies, it also relies on a large corpus of static documents; i.e. the Semantic Web Dog Food (SWDF)[11] corpus. Although in ranking expertise, *Saffron* makes a distinction between the frequency of an expertise topic occurring in the context of a skill type and the overall occurrence of an expertise topic, it ranks topics based on term frequency, which is a reliable measure only when a large corpus of documents is analysed. Saffron also extends information about people by crawling the Linked Open Data (LOD) (Bizer, Heath and Berners-Lee 2009) from seed URLs in SWDF. The meaning of the SWDF and crawled data represented using Semantic Web technologies is consolidated to build a holistic view of the social graph of an expert.

Existing social networks such as *BiomedExperts* (BME)[12] provide a source for inferring implicit relationships between concepts of the expertise profiles by analysing relationships between researchers; i.e., co-authorship. BME is the world's first pre-populated scientific social network for life science researchers. It gathers data from PubMed[13] on authors' names and affiliations and uses that data to create publication and research profiles for each author. It builds conceptual profiles of text, called Fingerprints, from documents, websites, emails and other digitised content and matches them with a comprehensive list of pre-defined "fingerprinted" concepts to make research results more relevant and efficient.

ResearcherID[14] is a global, multi-disciplinary scholarly research community where users can update their profile information, build their publication list using Web of Science and Web of Knowledge search services or uploading a file, and select to make their profile public or private. Registered as well as non-registered users can search the *ResearcherID* Registry to view profiles and find potential collaborators. Researchers can register for a unique researcher ID number in order to eliminate author misidentification and view an author's citation metrics instantly.

Interweaving traditional news media and social media is the goal of research projects such as *SYNC3*[15] , which aims to enrich news events with opinions from the blogosphere. *Twitris 2.0* is a Semantic Web platform that connects event-related Twitter messages with other media

[9] http://www.wikigenes.org/

[10] http://saffron.deri.ie/

[11] http://data.semanticweb.org/

[12] http://www.biomedexperts.com/

[13] http://www.ncbi.nlm.nih.gov/pubmed/

[14] http://www.researcherid.com/

[15] http://www.sync3.eu

such as YouTube videos and Google News (Jadhav, Purohit, Kapanipathi, Ananthram, Ranabahu, Nguyen, Mendes, Smith, Cooney and Sheth 2010). It incorporates Twarql, an open source infrastructure for the detection of DBpedia entities, presented to capture the semantics of major news events (Mendes, Passant and Kapanipathi 2010). The Twitter network is also analysed to capture tweets that correspond to late breaking news (Sankaranarayanan, Samet, Teitler, Lieberman and Sperling 2009).

Although these studies target micro-blogs and perform analysis on short text, there are fundamental differences between micro-contributions in the context of evolving knowledge bases, contributions to forum discussions and Twitter messages; namely, online knowledge bases don't have to be tailored towards various characteristics of tweets such as presence of @, shortening of words, usage of slangs, noisy postings, etc. Also, forum participations are a much richer medium for textual analysis as they are generally much longer than tweets and therefore provide a more meaningful context and usually conform better to the grammatical rules of written English. More importantly, Twitter messages do not evolve, whilst we specifically aim to capture expertise in the context of evolving knowledge.

Semantic Web technologies are also used to represent and manage provenance of data from Wikipedia and other wikis by providing, a lightweight ontology based on the W7 model. This abstract model is further refined using SIOC and the Actions module (Orlandi, Champin and Passant 2010). This research focuses on providing an RDF representation of provenance for Wikipedia pages - to identify trust values for pages, identify experts based on the number of contributions and other criteria such as the users' social graphs. However, it doesn't capture the semantics of contributions or a fine-grained provenance model to represent the grounding of contributions in the underlying content.

3 Methodology

The methodology proposed in this paper is based on the hypothesis that a comprehensive, fine-grained provenance model (able to capture micro-contributions in dynamic living documents), will enable accurate expertise profiling from evolving collaborative knowledge bases.

Consequently, we focus on three particular objectives:

- Development of a comprehensive model for capturing micro-contributions by combining coarse and fine-grained provenance, change management and ad-hoc domain knowledge;
- Development of a profile construction mechanism by computing ranked maps of weighted Linked Data concepts and consolidating Linked Data concepts via IR-inspired techniques;
- Development of a profile refinement mechanism by:
    - incrementally integrating the knowledge and expertise captured within given social professional networks.
    - Assessing revisions to micro-contributions over time (corrections/deletions)

Fig. 2 depicts the building blocks of our methodology, that are described in detail in the remainder of the paper.

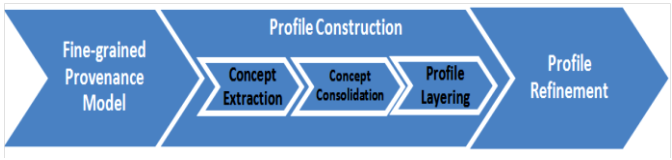

Fig. 2 Research methodology - building blocks

3.1 Fine-grained Provenance Model

This section provides a detailed overview of the fine-grained provenance model, starting with the requirements that underpin the design.

3.1.1 Requirements

Modularization (Groza, Handschuh, Breslin and Decker 2009): Modularization represents a key requirement for ontologies in order to achieve re-use and evolution (Rector 2003). With this aim in mind, we decouple domain knowledge and processes from our proposed fine-grained provenance model. This leads to a model that supports evolution and integration with ontologies from a variety of domains.

Identification and Revision (Groza, Handschuh, Breslin and Decker 2009): Contributions are represented as chunks of text, which capture the contribution semantics and are constituents of the host living documents. We identify the individual elements/contributions (text chunks) that comprise a document. In addition, by keeping track of the revisions made on these elements, we can track the evolution of micro-contributions, which enables us to adapt expertise profiles over time through multi-layered profiles. Tracking the evolution of micro-contributions enables us to monitor the activity performed by individuals, which in turn, provides a way to show not only the change in personal interests over time, but also an individual's maturation or regression process.

Support for Domain Knowledge and Specific Complementary Models (Groza, Handschuh, Breslin and Decker 2009): Our proposed model provides support for embedding domain knowledge. Ontologies from a variety of domains can be plugged into the model dynamically and benefit from the resulting multi-layered expertise profiles. We have also complemented our model with specific models for capturing coarse and fine-grained provenance and change management aspects of evolving knowledge.

3.1.2 Framework Overview

We propose a framework that combines coarse and fine-grained provenance modelling using the SIOC ontology (Breslin, Decker, Harth and Bojars 2006), with change management aspects captured by the SIOC-Actions module (Champin and Passant 2010). The Annotation Ontology (Ciccarese, Ocana, Castro, Das and Clark 2010) is then used to bridge the textual grounding and the ad-hoc domain knowledge, represented by concepts present in the Linked Data cloud, via domain-specific ontologies. The model is further complemented with the Simple

Knowledge Organization System (SKOS)[16] ontology to define concepts emerging from contributions to evolving knowledge.Our objective has been to reuse and extend existing, established vocabularies from the Semantic Web that have attracted a considerable user community or are derived from de facto standards. This focus guarantees direct applicability and low entry barriers (compared to developing an entirely new ontology from scratch). The final model has a layered structure where micro-contributions annotate the contributed text and are linked via the same annotations to domain knowledge. Instances of the model are not only useful for expertise profiling, but can also act as a personal repository of micro-contributions, to be published, reused or integrated within multiple evolving knowledge bases. Fig. 3 depicts a high-level overview of our fine-grained provenance framework.

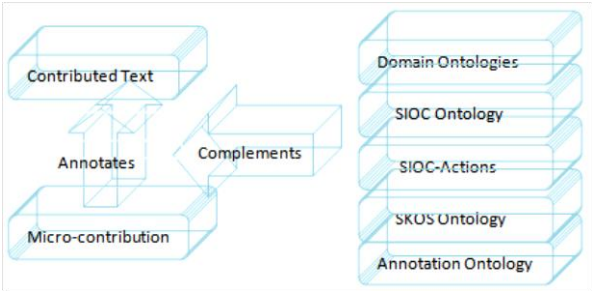

Fig. 3 High-level overview of the fine-grained provenance framework

3.1.3 Fine-grained Provenance Ontology

In this section we describe the Fine-grained Provenance ontology and its key concepts and properties. The change management process identifies the textual representation of micro-contributions in the context of evolving knowledge. The fine-grained provenance ontology captures the semantics of micro-contributions by annotating the textual representation of micro-contributions using domain specific ontologies and the Annotation ontology (Ciccarese, Ocana, Castro, Das and Clark 2010). This conceptual representation of contributions is used to create expertise profiles that provide a temporal representation of expertise through a multi-layered structure. As mentioned previously, where possible, we re-use and extend existing and well established vocabularies in the Semantic Web, rather than "re-inventing the wheel". Fig. 4 illustrates the structure of the Fine-grained Provenance ontology. Concepts have been colour coded to clearly represent complementing ontologies.

Complementary Ontologies

We draw on the following ontologies to build our fine-grained Provenance model:

The SIOC Ontology (Breslin, Decker, Harth and Bojars 2006), is used to describe the semantics of content generated within collaboration platforms; i.e. the macro-context of the host living documents.

The SIOC-Actions Module (Champin and Passant 2010), a module of the SIOC ontology, or *sioca* for short, is used to represent the way in which experts manipulate various digital artifacts that constitute the contents of collaboration platforms, from forum posts to content updates. In other words, this module describes the semantics of the change management process.

The Annotation Ontology (Ciccarese, Ocana, Castro, Das and Clark 2010) is used to bridge the lexical grounding and the ad-hoc domain knowledge, represented by concepts present in the Linked Data Cloud, via domain-specific ontologies. It captures annotations of terms identified in micro-contributions and their corresponding concepts.

Simple Knowledge Organization System (SKOS) is used to define concepts emerging from annotations performed on contributions to evolving knowledge.

Core Components

We discuss the core components of the Fine-grained Provenance ontology with respect to the way in which micro-contributions are represented. Where applicable, components are prefixed with terms representing the ontology in which they have been defined

Textual Representation

sioca:Action - The central notion of the SIOC-Actions module that represents a time stamped event involving a user and a number of digital artifacts (Champin and Passant 2010), used to represent a modification to a sioc:Item.

Contribution – A subclass of sioca:DigitalArtifact and the sioca:product of a sioca:Action, represents a contribution made to the contents of the underlying host living document; i.e., sioc:Item. The change management process captures the textual representation of a contribution.

TextChunk – The lowest granularity item, able to capture variable sized information chunks as constituents making up the host document, which is modelled as a sioc:Item. It is important to note that a sioca:Action, which sioc:modifies a sioc:Item produces a sioca:product; i.e. Contribution and a sioca:byproduct; i.e.sioc:Item, which is essentially a new version of the modified sioc:Item. This model captures changes and enables tracking the evolution of a sioc:Item. A sioc:Item will therefore have a sioc:next_version and sioc:previous_version, which are also sioc:Items.

Pointer – Represents an abstract concept for modelling the localization of the textual content i.e. Contribution and TextChunk in sioc:Item

[16] http://www.w3.org/TR/skos-primer

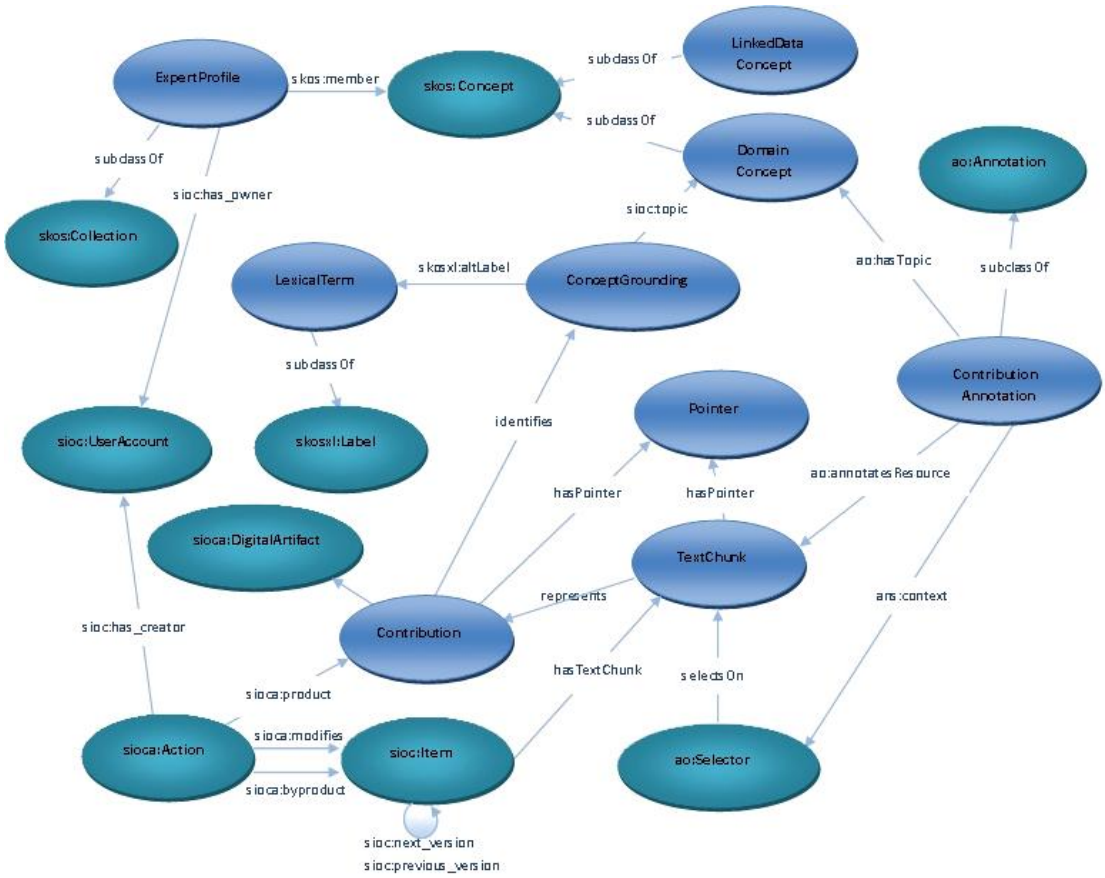

Fig. 4 Fine-grained Provenance Ontology – Concepts and relations

Conceptual Representation

ContributionAnnotation – A subclass of ao:Annotation defined in Annotation Ontology Core, instances of this class describe the common features of annotations performed on contributions and represent an annotated term within a contribution. A ContributionAnnotation has a number of ao:hasTopics, each of which represents a distinct annotation for the underlying term; e.g. represent annotations for the term from multiple ontologies.

ao:Selector – A core concept of the Annotation Ontology, used to describe the context of an annotation and represent an annotated term in a contribution. The Annotation Ontology defines a variety of selectors for different types of documents and content types e.g. based on XPointer for selecting a chunk of text in an XHTML document. A Selector based on an offset and range is used in our proposed Fine-grained Provenance ontology.

skos:Concept – A SKOS concept can be viewed as an idea or notion; a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive rather than restrictive[17].

Profile Representation

LinkedDataConcept – A subclass of skos:Concept, which represents a concept from the Linked Data Cloud and encapsulates the overall weight of the concept across all contributions for an expert.

ConceptGrounding – Represents a domain specific concept resulting from the concept consolidation process and its textual grounding identified in a contribution. A ConceptGrounding has a DomainConcept as its sioc:topic.

LexicalTerm – A subclass of skos-xl:Label[18], instances of this class represent the textual grounding of a domain-specific concept in a contribution. A lexical grounding is associated with a weight, representing its semantic importance in the context of the underlying contribution.

DomainConcept - A subclass of skos:Concept, represents an annotated concept in a contribution from a domain-specific ontology.

ExpertProfile – A subclass of skos:Collection, used to represent an expert profile. Every expert will have a number of ExpertProfile instances associated with them, where each instance represents a layer of the multi-layered expertise profile model (i.e. session, short-term, long-term) in a time period. An ExpertProfile instance includes a number of skos:Concept instances which are allocated to the profile according to concept currency, frequency and persistency criteria (Shtykh and Jin 2009), details of which are explained in the following section. Where possible, concepts from the Linked Data cloud are used in expertise profiles. Where Linked Data concepts are not applicable, domain-specific concepts will be included in the profiles.

[17] http://www.w3.org/TR/skos-reference/

[18] http://www.w3.org/TR/skos-reference/skos-xl.html

Ontology Design Patterns

The *Partition Pattern*[19] is a logical pattern that introduces axioms, which model a partition of concepts. A partition is a general structure, which is divided into several disjoint parts. With respect to ontologies the structure is a concept that is divided into several pair-wise disjoint concepts. This pattern reflects the simplest case where a named concept is defined as a partition of concepts. We have used this pattern to represent an ExpertProfile as a partition of profiles; i.e., skos:Collection, each representing the session, short-term and long-term profiles for an expert. Applying this pattern results in the following axioms:

```
EquivalentClasses(ExpertProfile,
ObjectUnionOf(Session ShortTerm LongTerm))
DisjointClasses(Session ShortTerm LongTerm)
```

Using SKOS Concept

A SKOS concept can be viewed as an idea or notion; a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive. The notion of a SKOS concept is useful when describing the conceptual or intellectual structure of a knowledge organization system (KOS), and when referring to specific ideas or meanings established within a KOS[20]. Applications that require finer granularity will greatly benefit from SKOS being a Semantic Web vocabulary. SKOS can indeed be seamlessly extended to suit the specific needs of a particular KOS community while retaining compatibility with applications that are based on the core SKOS features. This can mostly be done by specializing existing SKOS constructs into more-specific ones. Users can create their own properties and classes and attach them to the standard SKOS vocabulary elements by using the rdfs:subPropertyOf and rdfs:subClassOf properties from the RDF Schema vocabulary [RDF-PRIMER]. Therefore, in the Fine-grained Provenance ontology, concepts are represented as subclasses of skos:Concept. We have created a subclass of the skos-xl:Label (LexicalTerm) to represent weighted lexical groundings of a concept in an underlying contribution. We have also created a subclass of skos:Collection; i.e., ExpertProfile to represent the multi-layered expertise profile model.

3.2 Profile Construction

A temporal dimension is introduced to user profiles by splitting and combining (generalising) all concepts on a time line. User profiles are therefore multi-layered – each layer reflects user interests within a certain period. There are four layers – static, session, short-term, and long-term (Shtykh and Jin 2009). Each layer consists of concepts, which are the components of profiles that represent user contextual information by topic (Fig. 5).

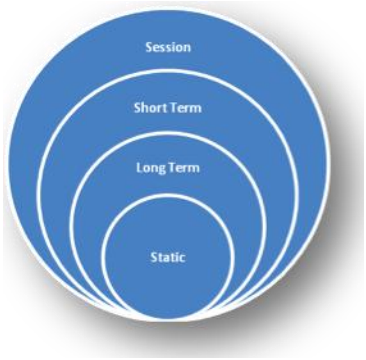

Fig. 5 Multi-layered user profile model (Shtykh and Jin, 2009)

The Profile construction phase comprises three steps: (i) concept extraction, where micro-contributions annotate the contributed text and are linked to their textual grounding in the underlying contribution, (ii) concept consolidation, where the most appropriate IR-inspired model will be used to consolidate the domain concepts present in micro-contributions and compute ranked maps of weighted concepts describing the expertise profiles; and (iii) profile layering, i.e., building the actual expertise profile by using the temporal dimension intrinsically associated with micro-contributions. The following sections outline the profile construction phase with reference to the fine-grained provenance ontology.

3.2.1 Concept Extraction

A *sioca:Action* modifies a *sioc:Item* through operations such as update, addition or deletion. This produces a *Contribution* (subclass of *sioca:DigitalArtifact*) and a new version of the modified *sioc:Item*. Micro-contributions annotate the contributed text and are linked via the same annotations to the domain knowledge. The process of annotation links a contribution to the domain knowledge through identifying concepts from complementing domain-specific ontologies. Annotations are performed on the *TextChunk* that encapsulates a contribution. An annotation is represented by an instance of *ContributionAnnotation* (subclass of *ao:Annotation*), which associates a term with concepts from domain-specific ontologies; i.e. *DomainConcept*. While our proposed approach is domain-agnostic, we intend to implement our methodology in the context of the SKELETOME project by using the NCBO Annotator[21] for annotating expert contributions.

3.2.2 Concept Consolidation

Annotations provide us with a conceptual representation of a contribution; however, as a number of domain-specific ontologies will be used for annotating contributions, there may be several annotations for a given term, each relating to a different ontology. In order to analyse the most suitable annotation for a given term, we need to determine the most suitable ontology for annotating the corresponding contribution. In order to determine the most appropriate ontology for annotating a

[19] http://ontologydesignpatterns.org/wiki/Submissions:Partition

[20] http://ontologydesignpatterns.org/

[21] http://www.bioontology.org/annotator-service

contribution, we calculate the similarity of concepts from domain-specific ontologies identified through annotations on the underlying contribution. The most suitable ontology will be the one for which annotated concepts are the most similar. This solves the concept disambiguation issue. At term level, this also solves the word sense disambiguation problem, as an annotation for the unintended sense of a term is likely to have the least similarity to other annotations by the same ontology.

Concept consolidation will be performed by comparing different IR models; i.e., vector space models (Salton, Wong and Yang 1975), latent semantic indexing (Papadimitriou, Raghavan, Tamaki and Vempala 2000) and conceptual spaces (Gärdenfors 2004).

Mapping concepts from heterogeneous ontologies is usually performed manually by domain experts or accomplished by computer programs via comparing the structures of the ontologies and the linguistic semantics of their concepts. A different approach compares Cosine similarity and Jaccard coefficient, two vector-based similarity measures and a variation of the Market Basket model to compare semantic similarity between ontologies. A document corpus related to the field of the domain-specific ontologies and the consideration of the corpus hierarchical information is used in concept similarity comparison. The proposed market basket model appears to outperform the other two similarity measures (Cheng, Lau, Pan, Law and Jones 2008). Our innovation comes from determining the most optimal method for concept consolidation in the context of micro-contributions; i.e., in the absence of a large corpus and taking into consideration the structural composition of the host living document in order to analyse relatedness of concepts.

Once the most suitable annotations are determined, we associate annotated terms with their representing concepts through instances of *ConceptGrounding*. Annotated terms are ranked based on their semantic importance within the text using the lexical chain linguistic phenomenon. Additionally, an n-dimensional concept vector space model is used to estimate the semantic importance of each concept and its lexical grounding within a document (Kang and Lee 2005).

We also extract evidence of expertise from the Semantic Web based on the association of an annotated term with an expert by using the *Sindice*[22] Semantic Web search engine. This analysis measures the relation strength between an expert and an annotated term representing an expertise topic/concept.

3.2.3 Profile Layering

This step builds the actual expertise profile by using the temporal dimension intrinsically associated with micro-contributions. User profiles are therefore multi-layered – each layer reflects user interests within a certain period. There are four layers – static, session, short-term, and long-term. The multi-layered user profile model is maintained by observing concept creation dynamics. The model update is not constrained by predefined parameters, such as the fixed time period after which the update occurs but driven by natural dynamics of changing user interests. This mechanism is used to find a user's *n* past profiles and their concepts to determine the areas of expertise (Shtykh and Jin 2009). Where possible, concepts from the Linked Data cloud are associated with expertise profiles. Where Linked Data concepts are not applicable, domain-specific concepts will be associated with the profiles.

3.3 Profile Refinement

As a final step, we plan to incorporate additional information acquired by analysing existing collaboration networks (such as BiomedExperts) and refine profiles based on the collaboration structure and collaborators' expertise. In other words, implicit relationships between concepts in expertise profiles can be inferred from relationships between experts; e.g. co-authorship relationships. With regard to collaboration structure, we will take into consideration the type of collaboration (e.g. co-authorship) and its strength. For a given expert, we will retrieve a dense sub-graph of his/her collaborators, by measuring the connectedness of direct neighbours to the expert node (clustering coefficient). We will establish the collaboration strength (strong, medium, weak, extremely weak, unknown) by measuring the minimum path length that connects two nodes in the network. The expert's profile will then be refined based on the profiles of experts with whom there is a strong collaboration. The refinement is performed by taking into consideration the similarity in expertise and the type and strength of collaboration between experts.

4 Evaluation

In order to evaluate our approach we will use data from various knowledge bases in the biomedical domain. More specifically we plan to evaluate our methodology in the context of the SKELETOME project (skeletal dysplasia domain), the iCAT[23] project (a collaboratively engineered ontology for ICD-11[24]) and the AlzSWAN (Alzheimer's disease) knowledge base.

Expertise profiles resulting from the profile construction phase will be evaluated against user-generated and system-generated profiles. We will use the following metrics in our user and system-based evaluations:

Precision – the accuracy, i.e. the degree of similarity between concepts of the expertise profiles resulting from the implementation of our proposed methodology and concepts of the expertise profiles identified by (i) participating experts (user-based evaluation) and (ii) a nominated expertise profiling system (system-based evaluation).

Recall - the percentage of concepts of the expertise profiles resulting from the implementation of our proposed methodology to concepts of the expertise profiles identified by (i) participating experts (user-based evaluation) and (ii) a nominated expertise profiling system (system-based evaluation).

[22] http://sindice.com/

[23] http://icat.stanford.edu/

[24] http://www.who.int/classifications/icd/ICDRevision/

User-based Evaluation: expertise profiles resulting from the implementation of our proposed methodology will be evaluated using expertise profiles generated by a number of participating domain experts. These profiles will be generated and maintained manually by the participants over the same period as the profiles generated by our system. Experts with significantly different interests will be used, in order to capture a variety of expertise with variable change rates.

System-based Evaluation: expertise profiles resulting from the implementation of our proposed methodology will be evaluated using expertise profiles generated by the following systems:

- The *BiomedExperts* scientific professional network – this is the first literature-based scientific professional network containing 1,800,000 pre-generated profiles of life science researchers.
- The open source *SubSift* software (Price, Flach, Spiegler, Bailey and Rogers 2010), a family of RESTful web services for profiling and matching text. *SubSift* uses traditional IR techniques such as TF-IDF, bag-of-words (BOW) and vector based modelling to profile and compare collections of documents.
- The *Saffron* (Monaghan, Bordea, Samp and Buitelaar 2010) application, an expert profiling and expert finding system that combines structured data from various sources on the Web with information extracted from unstructured documents, i.e. the Semantic Web Dog Food (SWDF) corpus using Natural Language Processing techniques. *Saffron* makes use of existing structured data on the web such as social connections and extends information about people by crawling the Linked Open Data (LOD) from seed URLs in the SWDF. It builds a holistic view of an expert's social graph using consolidated data from the SWDF and crawled data, represented by Semantic Web technologies.

We plan to compare the results of our concept consolidation approach with the results of the *Biomedical Ontology Recommender Web Service* (Jonquet, Musen and Shah 2010). Given textual metadata or a set of keywords describing a domain, the Recommender suggests ontologies appropriate for annotating or representing the data. Appropriateness is evaluated according to three main criteria; coverage, the ontology that best covers the given data; connectivity, the ontology containing the terms that are most often mapped or referred to by other ontologies; size, the number of concepts in the ontology. The following metrics will be used for evaluating the concept consolidation process:

- Recall - the percentage of ontologies selected by the concept consolidation phase to ontologies suggested as appropriate by the Recommender
- Precision - the comparison of ranking of ontologies with respect to their suitability for annotating a given contribution

The results of our proposed method for calculating concept similarity will also be evaluated against the UMLS::Similarity[25] project, which determines the similarity between two Unified Medical Language System (UMLS) concepts.

5 Conclusion

In this paper, we have proposed a methodology for building expertise profiles from micro-contributions in the context of living documents and evolving knowledge bases. The methodology consists of three building blocks: (i) a comprehensive fine-grained provenance model for representing micro-contributions, (ii) a profile construction phase that includes expertise concept extraction and consolidation, and (iii) a profile refinement phase that takes into account existing social professional networks and revisions to the contributions over time.

Although the model that we propose is being applied in the context of biomedical knowledge-bases, it has been designed to be domain-agnostic - so that it can potentially be used for expertise modelling in any domain, where domain knowledge is captured via micro-contributions. Furthermore, the proposed approach is highly significant in that it enables more precise, time-dependent expertise profiling within the context of online community-generated knowledge environments.

Acknowledgements. The work presented in this paper is supported by the Australian Research Council (ARC) under the Linkage grant SKELETOME - LP100100156.

References

Abel, F., Gao, Q., Houben, G. J. and Tao, K. (2011). "Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web." The Semanic Web: Research and Applications: 375-389.

Aleman-Meza, B., Bojars, U., Boley, H., Breslin, J. G., Mochol, M., Nixon, L. J. B., Polleres, A. and Zhdanova, A. V. (2007). "Combining RDF vocabularies for expert finding." In Proceedings of the 4th European Semantic Web Conference (ESWC2007), number 4519 in Lecture Notes in Computer Science, Innsbruck, Austria, June 2007. Springer: 235–250.

Bizer, C., Heath, T. and Berners-Lee, T. (2009). "Linked data-the story so far." Int. J. Semantic Web Inf. Syst. **5**(3): 1-22.

Breslin, J. G., Decker, S., Harth, A. and Bojars, U. (2006). "SIOC: An approach to connect web-based communities." The International Journal of Web-based Communities **2**(2): 133-142.

Casati, F., Giunchiglia, F. and Marchese, M. (2007). Liquid publications, Scientific Publications Meet the Web. Technical Rep. DIT-07-073, Informatica e Telecomunicazioni, University of Trento.

Champin, P. A. and Passant, A. (2010). SIOC in action representing the dynamics of online communities, Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10, New York, NY, USA, ACM, 2010: p. 1-7.

Cheng, C. P., Lau, G. T., Pan, J., Law, K. H. and Jones, A. (2008). Domain-specific ontology mapping by corpus-based semantic similarity. Proceedings of 2008 NSF CMMI Engineering Research and Innovation Conference, Knoxville, TN, USA.

Ciccarese, P., Ocana, M., Castro, L. J. G., Das, S. and Clark, T. (2010). "An open annotation ontology for science on

[25] http://orbit.nlm.nih.gov/resource/umlssimilarity

web 3.0." Journal of BioMedical Semantics **2**(Suppl 2): S4.

Demartini, G. (2007). Finding experts using wikipedia, Proceedings of theWorkshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007, Busan, South Korea.

Gärdenfors, P. (2004). Conceptual spaces: The geometry of thought, The MIT Press, 2004.

Groza, T., Handschuh, S., Breslin, J. G. and Decker, S. (2009). "An Abstract Framework for Modeling Argumentation in Virtual Communities." Int. J. of Virtual Communities and Social Networking **1**(3): 35-47.

Hoffmann, R. (2008). "A wiki for the life sciences where authorship matters." Nature genetics **40**(9): 1047-1051.

Jadhav, A., Purohit, H., Kapanipathi, P., Ananthram, P., Ranabahu, A., Nguyen, V., Mendes, P. N., Smith, A. G., Cooney, M. and Sheth, A. (2010). "Twitris 2.0: Semantically empowered system for understanding perceptions from social data." Proc. of the Int. Semantic Web Challenge.

Jonquet, C., Musen, M. A. and Shah, N. H. (2010). "Building a biomedical ontology recommender web service." Journal of BioMedical Semantics **1**(Suppl 1): S1.

Kang, B. Y. and Lee, S. J. (2005). "Document indexing: a concept-based approach to term weight estimation." Information processing & management **41**(5): 1065-1080.

Lee, T. B., Hendler, J. and Lassila, O. (2001). "The semantic web." Scientific American **284**(5): 34-43.

Mendes, P. N., Passant, A. and Kapanipathi, P. (2010). "Twarql: tapping into the wisdom of the crowd." Triplification Challenge 2010 at 6th International Conference on Semantic Systems (I-SEMANTICS), Graz, Austria: 1-3.

Michelson, M. and Macskassy, S. A. (2010). "Discovering users' topics of interest on twitter: a first look." Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Data in conjunction with the 19th ACM CIKM Conference: 73-80.

Monaghan, F., Bordea, G., Samp, K. and Buitelaar, P. (2010). "Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food."

Mons, B. and Velterop, J. (2009). Nano-Publication in the e-science era. Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), CEURWS Washington DC, USA.

O'Reilly, T. and Musser, J. (2006). "Web 2.0 principles and best practices." Retrieved March **20**: 2008.

Orlandi, F., Champin, P. A. and Passant, A. (2010). "Semantic Representation of Provenance in Wikipedia." Proceedings of the SWPM 2010, Workshop at the 9th International Semantic Web Conference, ISWC-2010.

Papadimitriou, C. H., Raghavan, P., Tamaki, H. and Vempala, S. (2000). "Latent semantic indexing: A probabilistic analysis." Journal of Computer and System Sciences **61**(2): 217-235.

Price, S., Flach, P. A., Spiegler, S., Bailey, C. and Rogers, N. (2010). SubSift web services and workflows for profiling and comparing scientists and their published works, IEEE Sixth International Conference on eScience.

Rector, A. L. (2003). "Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL." In Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP), ACM Press: 121–129.

Salton, G., Wong, A. and Yang, C. S. (1975). "A vector space model for automatic indexing." Communications of the ACM **18**(11): 613-620.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D. and Sperling, J. (2009). "Twitterstand: news in tweets." Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems: 42-51.

Shtykh, R. Y. and Jin, Q. (2009). "Dynamically constructing user profiles with similarity-based online incremental clustering." International Journal of Advanced Intelligence Paradigms **1**(4): 377-397.

Thiagarajan, R., Manjunath, G. and Stumptner, M. (October 2008). Finding experts by semantic matching of user profiles. Technical Report HPL-2008-172, HP Laboratories.

Yang, L. and Zhang, W. (2010). A study of the dependencies in expert finding, Proceedings of the 2010 Third International Conference on Knowledge Discovery and Data Mining, IEEE Computer Society Washington, DC, USA

Zhang, J., Tang, J. and Li, J. (2007). "Expert finding in a social network." Advances in Databases: Concepts, Systems and Applications: 1066-1069.

Zhu, J., Song, D. and Rüger, S. (2009). "Integrating multiple windows and document features for expert finding." Journal of the American Society for Information Science and Technology **60**(4): 694-715.