DC Proposal: Capturing Knowledge Evolution and Expertise in Community-driven Knowledge Curation Platforms

Hasti Ziaimatin

eResearch Lab, School of ITEE,
The University of Queensland, Australia
h.ziaimatin@uq.edu.au

Abstract. Expertise modeling 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 an all-encompassing, domain-agnostic model for expertise profiling in the context of dynamic, living documents and evolving knowledge bases. Our approach combines: (i) fine-grained provenance, (ii) weighted mappings of Linked Data concepts to expertise profiles, via the application of IR-inspired techniques on micro-contributions, and (iii) collaboration networks - to create and enrich expertise profiles in community-centered environments.

Keywords: Expertise profiling, Linked Data, Semantic Web, fine-grained provenance, micro-contributions

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. The task of expertise modeling 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 location 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. 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.

With the emergence of the Web of Data (which has evolved from the increasing use of ontologies, via the Semantic Web [1] and Web 2.0 [2]) there has been a significant shift from static documents to evolving documents. Wikis or diverse knowledge bases in the biomedical domain (e.g., Alzforum[1]) are examples of environments that support this shift by enabling authors to incrementally refine the content of the embedded documents to reflect the latest advances in the field. This gives the knowledge captured within them a dynamic character, and generating expertise profiles from this knowledge raises a new and different set of challenges. Both the IR and the 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 sufficient when tracking the evolution of both the knowledge and the expertise. As a result, such dynamic content requires a novel method for representing underlying concepts and linking them to an expertise profile. This method involves capturing and analyzing micro-contributions by experts through a fine-grained provenance model.

1.1 Aims and objectives

A comprehensive, fine-grained provenance model, able to capture micro-contributions in the macro-context of the host living documents, will facilitate expertise profiling in evolving knowledge bases or environments where the content is subject to ongoing changes.

The focus of the research will be on addressing the following questions:

- How can we model the provenance and evolution of micro-contributions in a comprehensive and fine-grained manner?
- How can we bridge the gap between Linked Data [3] domain concepts and their lexical representations, by taking into account *acronyms*, *synonyms* and/or *ambiguity*?
- What is the best IR-inspired model for consolidating the domain concepts present in micro-contributions and for computing ranked maps of weighted concepts describing the expertise profiles?
- How can we enrich expertise profiles using existing collaboration networks?
- What is the appropriate methodology for evaluating emerging expertise profiles from evolving micro-contributions?

These research questions can be further specified into the following objectives:

O1: development of a comprehensive model for capturing micro-contributions by combining coarse and fine-grained provenance, change management and ad-hoc domain knowledge;

[1] http://www.alzforum.org/

O2: development of a profile building mechanism by computing ranked maps of weighted (Linked Data) concepts and consolidating (Linked Data) concepts via IR-inspired techniques;

O3: development of a profile refinement mechanism by incrementally integrating the knowledge and expertise captured within given social professional networks.

The outcomes will be applied and evaluated in the context of the SKELETOME project[2] (a knowledge base for the skeletal dysplasia domain), the iCAT project[3] (a collaboratively engineered ontology for ICD-11) and the Alzforum (Alzheimer disease) knowledge base.

1.2 Significance and Innovation

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. Using well-grounded concepts from widely adopted vocabularies or ontologies, such as the ontologies published in the Linked Data Cloud, enables a straightforward consolidation of the expertise profiles. As a result, the overhead imposed by performing co-reference entity resolution (to consolidate the expertise concepts to a shared understanding) will be reduced to a minimum. From an academic perspective, a shift in the scientific publishing process seems to gain momentum, from the current document-centric approach to a contribution-oriented approach in which the hypotheses or domain-related innovations (in form of short statements) will replace the current publications. Examples of this new trend can be seen via nano-publications [4] or liquid publications [5]. In this new setting, mapping such micro-contributions to expertise will become essential in order to support the development of novel trust and performance metrics.

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 (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

[2] http://itee.uq.edu.au/~eresearch/projects/skeletome/

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

retrieval model is used for association discovery [6] or the effectiveness of exploiting the dependencies between query terms for expert finding is proved [7]. Other studies present solutions through effective use of ontologies and techniques such as *spreading* to include additional related terms to a user profile by referring to an ontology (Wordnet or Wikipedia) [8]. 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 no longer reliable. Their heavy dependency on statistical techniques (e.g., TF/IDF) cannot be applied on micro-contributions, because these don't offer sufficient context to capture the encapsulated knowledge.

A different approach is adopted by the ExpertFinder framework, which uses and extends existing vocabularies on the Semantic Web (e.g., FOAF, SIOC) as a means for capturing expertise [9]. 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 [10]. We intend to leverage these studies in order to enable the straightforward consolidation of expertise profiles to generate a shared understanding by using widely adopted vocabularies and ontologies. This will also lead to a seamless aggregation of communities of experts.

In the context of micro-blogging, a proposed framework includes the capacity to link entities within each blog post to their disambiguated concept on the Semantic Web [11]. However, this approach relies on the author to manually annotate the entities in the post. Our proposed approach could alleviate this manual linking of the entities. 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 frequently and cover the entities [12]. Although this study analyzes short texts, there are fundamental differences between micro-contributions in the context of online knowledge bases and Twitter messages. These differences include: shortening of words, usage of slang, noisy postings and the static nature of twitter messages.

Finally, other studies focus on methods for finding experts in social networks and online communities via a conceptual framework that uses ontologies such as FOAF, SIOC or SKOS [13]. A propagation-based approach to finding experts in a social network is proposed that makes use of personal local information and relationships between persons in a unified approach [14]. Most studies focus on connecting experts based on their profiles and social networks. However, our focus is on leveraging existing collaboration networks and enriching expertise profiles based on the relatedness of an expert's profile and the profiles of his/her collaborators.

3 Approach and Methodology

The objectives listed in Sect. 1.1, and depicted in Fig.1 represent the building blocks of our research methodology. In the following sections we detail the provenance and change management model for micro-contributions, the profile building and the profile refinement phases.

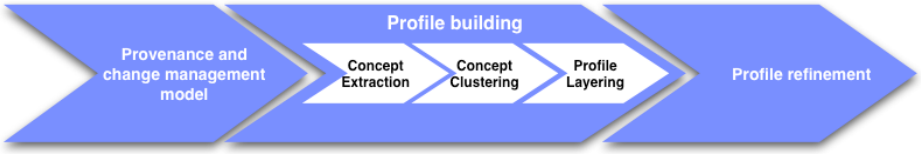

Fig. 1 Research methodology - building blocks

3.1 Provenance and change management model

The first step of our research involves the development of a comprehensive provenance and change management model for micro-contributions. The resulting ontology will combine coarse and fine-grained provenance modeling (using the SIOC ontology [15]) with change management aspects captured by the generic W7 model [16] and SIOC actions [17]. The Annotation Ontology [18] will then be used to bridge the lexical grounding and the ad-hoc domain knowledge, represented by concepts present in the Linked Data Cloud, via ontologies such as SNOMED. As a result, the final model will have a layered structure where *micro-contributions* will *annotate* the contributed *text* and will be linked via the same *annotations* to domain knowledge. Instances of the model will not only be useful for expertise profiling, but will also act as a personal repository of micro-contributions, to be published, reused or integrated within multiple evolving knowledge bases.

3.2 Profile building

The Profile building phase comprises three steps: (i) concept extraction, i.e., extracting the domain concepts from micro-contributions, (ii) concept clustering, i.e., consolidating the domain concepts around particular centroids, and (iii) profile layering, i.e., building the actual expertise profile by using the temporal dimension intrinsically associated with micro-contributions.

Concept extraction. This step bridges the gap between the domain knowledge and evolving documents. While in general our methodology is domain-agnostic, in the context of our target use cases, we intend to utilize the NCBO Annotator[4] to identify and annotate concepts within the micro-contributions. For every identified concept, we will define its corresponding lexical chain of terms from the results of the annotation. Assuming that terms having more relations with other terms are semantically more important, we will attach a weight to each term based on its relations with other terms such as identity, synonymy and their weights in the lexical chain. We will use WordNet [19] to determine the relatedness of terms. The concept weight is then obtained by summing the scores of all terms in the lexical chain. The output of this step is therefore, a weighted vector of terms in the lexical chain representing each concept identified by the annotator in a micro-contribution.

Concept clustering. In this step, the aim is to maintain a collection of concept clusters such that as each concept is presented, either it is assigned to one of the current clusters, or it starts off a new cluster, while two existing clusters are merged

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

into one [20]. We propose a real-time, incremental and unsupervised algorithm for clustering concepts resulting from the previous step. The concept-cluster similarity is measured using the classic cosine similarity between the reference point *Rp* (representative concept) of each cluster and a concept identified in the previous step. If the similarity value is higher than a predefined similarity threshold, the concept is assigned to the nearest cluster; otherwise incoming concepts are examined to create a new cluster. If the concept already exists in the identified cluster, its weighted vector is adjusted according to the weighted vector of the assigned concept; otherwise, the concept and its associated weighted vector will be assigned to the cluster. The weight associated with a concept in a cluster will therefore represent the overall weight of the concept across all micro-contributions for an expert. As concept weights in clusters are subject to change, we will recalculate a cluster's reference point (using a method similar to the Evolving Clustering Method (ECM) [21]), each time the weight of a concept is changed or a concept is added to a cluster. As this is a completely separate process, it will not affect the performance of our proposed clustering algorithm.

Profile layering. We introduce a temporal dimension to user profiles by splitting and combining concepts on a timeline. Thus, the user profiles will be multi-layered; static, session, short-term and long-term, with each layer reflecting a user's interests within a certain period. This approach will not only reflect the changeability of user interests but also maintain the steadiness of persistent preferences. Once concepts are extracted, weighted and clustered, we will detect any change of context and assign the latest as the current context (currency criterion) to the session layer. The short-term layer consists of the most frequently updated and used concepts (frequency criterion), which are in turn chosen from the most recent concepts in the concept currency list. The long-term layer is derived from the concepts of the short-term layer (currency and frequency criteria), whose *persistency factor, PF,* is high. PF is a measure to infer an expert's continuous interests by combining a concept's frequency count with its evidence of being a constituent of the user's short-term layer. The emergence of a new model for a user is not determined by predefined parameters, such as the fixed time period after which a new model should be created. It is rather driven by natural dynamics of changing user interests, signified by change of the concepts in terms of their ranking or new concepts in the short-term profile layer.

3.3 Profile Refinement

As a final step, we will look at existing collaboration networks and refine profiles based on the collaboration structure and collaborators' expertise. 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, through measuring the connectedness of direct neighbours to the expert node *(clustering coefficient)*. We will then establish the collaboration strength *(strong, medium, weak, extremely weak, unknown)* by measuring the minimum path length that connects two nodes in the network; i.e. *Geodesic*. 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 considering the similarity in

expertise through measuring the cosine similarity between the weighted vector of expertise concepts and the type and strength of collaboration between experts.

4 Evaluation

We will specifically evaluate the concept clustering, overall profile building and profile refinement phases using the use cases outlined in Section 1.1. *Precision* and *recall* will be used as metrics in all of the evaluation steps.

The proposed concept clustering method will be compared against the ECM algorithm [21] since (as with our method) it does not limit the number of clusters and the threshold value determines the ranges within which concepts assigned to a particular cluster must lie. In ECM the threshold value is a distance, whereas in our proposed method, it's a similarity. Therefore in our experiments, the distance threshold for the algorithm will be converted into similarity. We will process the same use cases with ECM and use cosine similarity as a similarity measure. The evaluation will compare our method against ECM, using a number of different similarity thresholds. We will also record the following quantities as additional evaluation measures for this step: the number of concepts correctly assigned to the relevant cluster (*true positives*), the number of concepts incorrectly assigned to a cluster (*false positives*), the number of concepts correctly identified as irrelevant to all clusters, hence their assignment to the pool (*true negatives*) and the number of concepts relevant to a cluster, incorrectly identified as irrelevant (*false negatives*).

The evaluation of the profile building phase will be performed using a range of domain experts from each of our use cases. The chosen participants will have significantly different interests – in order to capture a variety of expertise with variable change rates. The profile resulting from the building phase will be compared with manually generated and maintained expert profiles over the same period. This comparison will specifically target new profile generation/update as a result of new concepts and concept rank changes, as well as the precision and recall of our method.

For profile refinement, we will target existing collaboration networks specific to our domain, such as BiomedExperts[5] – an online community connecting biomedical researchers through the display and analysis of the networks of co-authors with whom each investigator publishes scientific papers. As in the previous step, we will select a number of experts from our use cases with a variety of interests and co-authorship activities. The refined profile will be compared against the expertise profiles generated by the collaboration network.

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. This methodology consists of three building blocks: (i) a comprehensive provenance and change management model for micro-contributions, (ii) a profile building phase

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

that includes expertise concept extraction and clustering, and (iii) a profile refinement phase that takes into account existing social professional networks. This research will lead to two significant outcomes: time-dependent expertise profiling, as well as adaptive and novel trust and performance metrics in incrementally changing knowledge environments.

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

References

1. Lee, T.B., J. Hendler, and O. Lassila, The semantic web. Scientific American, 2001. **284**(5): p. 34-43.
2. O'Reilly, T. and J. Musser, Web 2.0 principles and best practices. Retrieved March, 2006. **20**: p. 2008.
3. Bizer, C., T. Heath, and T. Berners-Lee, Linked data-the story so far. Int. J. Semantic Web Inf. Syst., 2009. **5**(3): p. 1-22.
4. Mons, B. and J. Velterop. Nano-Publication in the e-science era. Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse, co-located with ISWC 2009, Chantilly, Virginia, US, Oct. 2009.
5. Casati, F., F. Giunchiglia, and M. Marchese, Liquid publications, Scientific Publications Meet the Web. Technical Rep. DIT-07-073, Informatica e Telecomunicazioni, University of Trento, 2007.
6. Zhu, J., D. Song, and S. Rüger, Integrating multiple windows and document features for expert finding. Journal of the American Society for Information Science and Technology, 2009. **60**(4): p. 694-715.
7. Yang, L. and W. Zhang. 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, 2010
8. Thiagarajan, R., G. Manjunath, and M. Stumptner. Finding experts by semantic matching of user profiles. in Technical Report HPL-2008-172, HP Laboratories. October 2008.
9. Aleman-Meza, B., et al., Combining RDF vocabularies for expert finding. In Proceedings of the 4th European Semantic Web Conference (ESWC2007), Innsbruck, Austria. Springer, 2007: p. 235–250.
10. Demartini, G. Finding experts using wikipedia. Proceedings of theWorkshop on Finding Experts on the Web with Semantics (FEWS2007) at ISWC/ASWC2007, Busan, South Korea, November 2007.
11. Passant, A., et al. Microblogging: A semantic and distributed approach. Proceedings of Workshop on Scripting for the Semantic Web, 2008.
12. Michelson, M. and S.A. Macskassy, 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, 2010: p. 73-80.
13. Breslin, J.G., et al. Finding experts using Internet-based discussions in online communities and associated social networks. First International ExpertFinder Workshop, Berlin, Germany, 2007.
14. Zhang, J., J. Tang, and J. Li, Expert finding in a social network. Advances in Databases: Concepts, Systems and Applications, 2007: p. 1066-1069.
15. Breslin, J.G., et al., SIOC: An approach to connect web-based communities. The International Journal of Web-based Communities, 2006. **2**(2): p. 133-142.
16. Ram, S. and J. Liu, Understanding the semantics of data provenance to support active conceptual modeling. Active conceptual modeling of learning, 2007: p. 17-29.
17. Orlandi, F., P.A. Champin, and A. Passant, Semantic Representation of Provenance in Wikipedia. Proceedings of the SWPM 2010, Workshop at the 9th International Semantic Web Conference, ISWC-2010.
18. Ciccarese, P., et al., Ao: An open annotation ontology for science on the web. Proceedings of Bio Ontologies 2010, Boston, MA, 2010.
19. Fellbaum, C., WordNet: An electronic lexical database: The MIT press, Cambridge (MA US), 1998.
20. Charikar, M., et al. Incremental clustering and dynamic information retrieval. Proceedings of the 29th Annual ACM Symposium on Theory of Computing, 1997: p. 626-635.
21. Kasabov, N.K., Evolving connectionist systems: the knowledge engineering approach: London, U.K.: Springer-Verlag, 2007.