Concept Extraction Applied to the Task of Expert Finding

Georgeta Bordea

Unit for Natural Language Processing,
Digital Enterprise Research Institute,
National University of Ireland, Galway
georgeta.bordea@deri.org

1 Research Problem

The Semantic Web uses formal ontologies as a key instrument in order to add structure to the data, but building domain specific ontologies is still a difficult, time consuming and error-prone process since most information is currently available as free-text. Therefore the development of fast and cheap solutions for ontology learning from text is a key factor for the success and large scale adoption of the Semantic Web. Ontology development is primarily concerned with the definition of concepts and relations between them, so one of the fundamental research problems related to ontology learning is the extraction of concepts from text. To investigate this research problem we focus on the expert finding problem, i.e, the extraction of expertise topics and their assignment to individuals. The ontological concepts we extract are a person's skills, knowledge, behaviours, and capabilities.

For increased efficiency, competitiveness and innovation, every company has to facilitate the identification of experts among its workforce. Even though this can be achieved by using the information gathered during the employment process and through self-assessment, a person's competencies are likely to change over time. Information about people's expertise is contained in documents available inside an organisation such as technical reports but also in publicly available resources, e.g., research articles, wiki pages, blogs, other user-generated content. The human effort required for competency management can be reduced by automatically identifying the experts and expertise topics from text. Our goal is to explore how existing technologies for concept extraction can be advanced and specialised for extracting expertise topics from text in order to build expertise profiles.

2 Related Work

Expertise or competence management is a research topic from the area of knowledge management that is concerned with the "identification of skills, knowledge, behaviours, and capabilities needed to meet current and future personnel selection needs"[1]. In this thesis we focus on gathering the knowledge of an organisation in terms of scientific topics and technologies.

L. Aroyo et al. (Eds.): ESWC 2010, Part II, LNCS 6089, pp. 451–456, 2010.

Extensive work has been done for the task of expert finding using information retrieval techniques. In these approaches users look for experts in a collection of documents by giving a query with the topics of interest [2,3,4]. The proposed solutions are matching the user's query against the document collection in order to find the experts. This approach makes the assumption that the user is looking for an expertise topic, so it is not possible to find the expertise profile for a person. Other approaches rely on ontologies for competency management, building inference services ([5,6]). The ontologies are used for matchmaking services that bring together the skills demand and supply [7], but these methods can not be applied for domains where an ontology of skills is not already built. An approach to build an ontology of competencies has been proposed in [8], making use of an already built domain ontology, but this can not be applied for domains where an ontology is not defined or where new concepts are introduced often.

In [9] the relations between people and skills are extracted as a network of associations, both people and competencies being handled as entities. Although here the dynamic and automated support of expertise management is addressed, a deeper analysis of expertise topics is needed. Expertise should be analysed on several levels (knowledge, ability, centrality, and context). We analyse expertise on all these levels and we implement an integrated text mining strategy for each of them. In [10] an approach based on text genre specific lexico-syntactic patterns from scientific publications is investigated , but only a short list of context patterns manually identified is considered. In our work we propose an automatic method to identify context patterns that will increase the number of expertise topics that are considered ([11,12]).

3 Proposed Approach

The thesis is based on state of the art techniques from term and keyword extraction but applies these techniques to authors rather than documents. The novel aspect that our work covers is how to learn topic extraction patterns using web based knowledge sources (Linked Data[1], ontologies, etc.) as background knowledge. An important research challenge is the lexical disambiguation in the context of Linked Data, i.e., how can we reliably disambiguate extracted topics that have more than one possible interpretation in order to assign them to a unique URI. Our work is based on state of the art techniques in word sense disambiguation extended with a notion of semantic context as provided by web based knowledge sources. The thesis focuses on the topics illustrated in the following sections.

3.1 Extracting Expertise Topics

Text genre specific lexico-syntactic patterns, i.e., frequently occurring patterns of particular lexical items or words in a certain syntactic sequence, are central to our approach. Some patterns are specific to a scientific area (e.g., for computer

[1] Linked Open Data: `http://linkeddata.org`

science: "implementation of", "algorithm for") while other patterns are used in any scientific domain (e.g., "approach for", "analysis of"). We first consider the list of context patterns identified in [10] and then we use the extracted expertise topics in order to identify other context patterns.

Using the syntactic description of a term, we discover candidate expertise topics in the vicinity of context patterns. Similar to the term weighting approach in information retrieval we use a combination of statistical measures to rank the candidate expertise topics, taking into consideration the frequency and the topic length. In addition we analyse the structure of a document and the relation between an expertise topic and the section of the document where it was extracted. We are using the Yahoo BOSS[2] search engine to filter out too general or too specific expertise topics by comparing the number of occurrences in the corpora and on the web. After extracting the linguistic realisations of the concepts from the text, we associate each expertise topic with background knowledge available from the Linked Data cloud. To achieve this, we disambiguate the terms that refer to several concepts using word sense disambiguation techniques. Different terms can refer to the same concept, therefore we explore the similarity of expertise topics to associate synonym expertise topics to a concept. In addition we plan to investigate the relation between expertise topics based on their co-occurrence and to explore different methods for expertise topic clustering.

3.2 Extracting Expertise Profiles

The expertise topics in a document are added to the expertise profile for the authors of the document, considering that they are subject matter experts as suggested in [13]. We assign a measure of relevance to each topic from the document collection. The measure of relevance for an author is computed using an adaptation of the standard information retrieval measure TF/IDF. The set of documents of a researcher is considered as a virtual document, and we measure the relevancy of each expertise topic over this virtual document. To identify a researcher's expertise level we take into consideration the performance indicators introduced in [14]: knowledge (coverage of the expertise graph), ability (practical skills associated with expertise) and transfer (centrality and application in different contexts).

4 Methodology and Current Contributions

We are planning to evaluate the proposed approach by implementing solutions for the problem of expertise topic and expertise profile extraction presented in Section 3. We will compare the extracted expertise topics with the terms extracted by systems that perform terminology extraction. We will compare our results from the computer science domain with the ACM topic hierarchy[3] and the

[2] Yahoo BOSS: `http://developer.yahoo.com/search/boss`
[3] ACM topic hierarchy: `http://www.acm.org/about/class/1998`

results from computational linguistics domain with the LT World[4] Ontology on Language Technology. Another method is to compare the expertise topics with the topics mentioned in the call for papers for the conference of each scientific publication. The evaluation of expertise profiles will be performed with a user study, asking a group of researchers to evaluate their own expertise profile, and also the expertise profiles for a set of well known researchers from their field. Advanced precision and recall measures (e.g. learning accuracy) will be used in this.

We currently use three different data sets for experiments: a corpus of scientific publications from Semantic Web conferences[5], a collection of articles published by researchers working in a web science research institute, and the ACL Anthology Reference Corpus[6]. The first dataset consists of 680 papers and 1692 researchers from 11 semantic web conferences starting from 2006 to 2009. The second dataset contains 405 scientific publications and 362 researchers. The ACL Anthology Reference Corpus is a much larger set that consists of 10921 scientific articles starting from 1965 to 2006 and 9983 researchers.

The first prototype for expertise mining is using 46 different context patterns manually identified by inspecting a set of publications. The system builds expertise profiles for an author as a list of ranked expertise topics. The top 5 expertise topics extracted from the Semantic Web Corpus are presented in Table 1(a). You can observe that the results are promising because the top expertise topics are relevant for the Semantic Web area. We also computed the expertise profile for researchers over the years. In Table 1(b) we can observe that the researcher was interested on topics related to data in 2006 but he was more interested in information retrieval in the next year.

Table 1. Top expertise topics from the Semantic Web Corpus

(a) Top expertise topics

Expertise Topics
semantic web
social networks
Web services
semantic web services
SW objects

(b) Top expertise topics for Stefan Decker

2006	**2007**
Semantic Web	Semantic Web
RDF data	information retrieval
Semantic Web data	query processing
web pages	RDF data
inverted index	PDF documents

5 Conclusions and Future Work

This paper explored how existing techniques for concept extraction can be advanced and specialised for the application of expert finding. After presenting our approach and methodology, we described some preliminary results that show we

[4] LT World: `http://www.lt-world.org/`
[5] Semantic Web Corpus: `http://data.semanticweb.org`
[6] ACL Anthology Reference Corpus: `http://acl-arc.comp.nus.edu.sg`

extracted relevant Semantic Web expertise topics. The next step is to improve the current prototype by automatic context pattern discovery. We will then analyse the performance of the context patterns and the relation between the document structure and the expertise topics. To encourage the use of the extracted results for other applications we are planning to set up a SPARQL[7] endpoint for expertise data access, based on an expertise ontology that we will also develop. For the evaluation of the proposed algorithms we will participate in the SemEval[8] evaluation challenge for Automatic Keyphrase Extraction from Scientific Articles.

Acknowledgements

This work is supported in part by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2). We wish to thank Paul Buitelaar for his valuable contributions and supervision.

References

1. Draganidis, F., Metzas, G.: Competency based management: A review of systems and approaches. Information Management and Computer Security 14(1), 51–64 (2006)
2. Macdonald, C., Ounis, I.: Voting for candidates: adapting data fusion techniques for an expert search task. In: CIKM 2006: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 387–396. ACM, New York (2006)
3. Serdyukov, P., Rode, H., Hiemstra, D.: Exploiting sequential dependencies for expert finding. In: SIGIR 2008: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 795–796. ACM, New York (2008)
4. Craswell, N., Hawking, D., Vercoustre, A.M., Wilkins, P.: P@noptic expert: Searching for experts not just for documents. In: Ausweb, pp. 21–25 (2001)
5. Sure, Y., Maedche, A., Staab, S.: Leveraging corporate skill knowledge – from proper to ontoproper. In: Proceedings of the Third International Conference on Practical Aspects of Knowledge Management, pp. 30–31 (2000)
6. Kunzmann, C., Schmidt, A.: Ontology-based competence management for healthcare training planning: A case study. In: 6th International Conference on Knowledge Management, IKNOW 2006 (2006)
7. Colucci, S., Noia, T.D., Sciascio, E.D., Donini, F.M., Piscitelli, G., Coppi, S.: Knowledge based approach to semantic composition of teams in an organization. In: SAC 2005: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 1314–1319. ACM, New York (2005)
8. Posea, V., Harzallah, M.: Building a competence ontology. In: Proceedings of the Workshop Enterprise Modelling and Ontology of the International Conference on Practical Aspects of Knowledge Management, PAKM 2004 (2004)

[7] SPARQL: `http://www.w3.org/TR/rdf-sparql-query`
[8] SemEval: `http://semeval2.fbk.eu/semeval2.php`

9. Zhu, J., Goncalves, A.L., Uren, V.S., Motta, E., Pacheco, R.: Mining web data for competency management. In: Proceedings of Web Intelligence WI 2005, pp. 94–100. IEEE Computer Society, Los Alamitos (2005)
10. Buitelaar, P., Eigner, T.: Topic extraction from scientific literature for competency management. In: Proceedings of the Workshop Personal Identification and Collaborations: Knowledge Mediation and Extraction of the International Semantic Web Conference (2008)
11. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning taxonomic relations from heterogeneous sources of evidence. In: Ontology Learning from Text: Methods, Evaluation and Applications (2005)
12. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–545 (1992)
13. Becerra-Fernandez, I.: Facilitating the online search of experts at nasa using expert seeker people-finder. In: Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management (2000)
14. Paquette, G.: An ontology and a software framework for competency modeling and management competency. Educational Technology & Society 10, 1–21 (2007)