Expertise Mining for Enterprise Content Management

Georgeta Bordea, Sabrina Kirrane, Paul Buitelaar, Bianca Pereira

Unit for Natural Language Processing, Digital Enterprise Research Institute,
National University of Ireland, Galway;
Storm Technology, Ireland;
Knowledge Engineering Group, Federal University of Rio de Janeiro, Brazil
name.surname@deri.org

Abstract

Enterprise content analysis and platform configuration for enterprise content management is often carried out by external consultants that are not necessarily domain experts. In this paper, we propose a set of methods for automatic content analysis that allow users to gain a high level view of the enterprise content. Here, a main concern is the automatic identification of key stakeholders that should ideally be involved in analysis interviews. The proposed approach employs recent advances in term extraction, semantic term grounding, expert profiling and expert finding in an enterprise content management setting. Extracted terms are evaluated using human judges, while term grounding is evaluated using a manually created gold standard for the DBpedia datasource.

Keywords: Expert Finding, Expertise Topics, Enterprise Content Management, Semantic Term Grounding

1. Introduction

The Enterprise Content Management (ECM) concept has been slowly evolving over the past two decades. Today, this all encompassing term is commonly used to refer to enterprise document management, content management, records management, collaboration, portal technologies, workflow and search (Dilnutt, 2011). The business benefits attributed to the deployment of an ECM system are compliance, efficiency, customer service and lower costs (Scott, 2011), (Dilnutt, 2011). Nevertheless, the aforementioned benefits are highly dependent on the effectiveness of both the taxonomy and metadata used to describe the enterprise content (Munkvold and Hodne, 2006), (Scott, 2011).

Significant groundwork goes into the initial enterprise content analysis and platform configuration, therefore organisations regularly seek help carrying out this activity from specialist consultants known as Information Architects. Because Information Architects are not necessarily experts in the organisation's business domain it is important that they identify key individuals to be involved in analysis interviews that will result in the construction of the ECM taxonomy.

Recent advances in expert finding (Balog et al., 2006), (Macdonald and Ounis, 2006), (Serdyukov et al., 2008), automatic term recognition and automatic taxonomy construction (Roberto Navigli and Faralli, 2011), (Kozareva and Hovy, 2010), as well as the increasing richness of structured data openly available on the Linked Data cloud [1] address some of these challenges. But these approaches still suffer from various limitations such as exact matches of expertise topics, lack of expert profiles needed in the selection process and generality of extracted terms and taxonomical relations.

Our research focuses on automatic techniques to support the initial content analysis, taxonomy generation and the selection of experts who can validate the knowledge obtained from enterprise repositories. Expertise mining complements the traditional task of expert finding with expertise topic extraction and expert profiling to automatically link expertise, Information Workers and documents. The ECM taxonomy can be used both for organising the enterprise contents and for improving expert finding. We integrate data driven approaches for expert search with knowledge resources such as domain taxonomies and Linked Data expertise traces. Our approach is implemented in the expert finding service Saffron [2].

2. Related work

Expert finding can be modelled as an information retrieval task by performing a full text search for experts instead of documents. A large body of work was encouraged by the Text REtrieval Conference (TREC)[3] with the introduction of the expert finding task which provided common grounds to evaluate and assess methods and techniques. Language models (Balog et al., 2006), latent topic models (Rosen-Zvi et al., 2010) and voting models (Macdonald and Ounis, 2006) can be used for this task.

Expert profiling is an essential part of an expert search system, but this task received considerably less attention in literature than the expert finding task. Building a topical profile facilitates the selection of experts, by providing additional context with respect to expertise topics. Following (Balog and Rijke, 2007) we use "expert profile" to refer to competencies, knowledge and skills but not to the background information of an expert (e.g., affiliation, education, contact) as in (Latif et al., 2010).

In previous expert profiling studies, knowledge areas were either assumed to be known (Balog and Rijke, 2007), created by users through tagging (Serdyukov et al., 2011)

[1] Linked Data cloud: a freely available collection of structured data from different domains that provides us with a gateway to additional information about expertise topics and people `http://richard.cyganiak.de/2007/10/lod/`

[2] Available at `http://saffron.deri.ie/lrec`

[3] `http://trec.nist.gov/`

Table 1: Domain specific general terms

algorithm	framework	software
analysis	implementation	strategy
approach	mechanism	study
design	method	system
development	model	technique
device	problem	technology
execution	program	theory

or extracted through simple methods that require domain knowledge such as manually selected seed words and raw occurrences on the web (Nakajima et al., 2009). It is not only expertise topics that are important for expert search but also the relations between terms. Taxonomical relations are a valuable resource for expert finding (Cameron et al., 2010), but a taxonomy has to be manually built for this purpose.

3. The expertise mining approach for expert search

Expertise topics are defined as the lexical realisation of a knowledge area, while the expert profile of an individual is defined as a ranked list of expertise topics along with supporting evidence, the list of documents used for the extraction (Balog and Rijke, 2007). To build an expert profile we first identify expertise topics, by analysing a corpus of documents with good coverage of the domain, as described in section 3.1. These are further included in the expert profiles of individuals associated with each document as discussed in detail in section 3.2.

3.1. Expertise topic extraction

Expertise mining explicitly identifies skills and competencies (which we call expertise topics), similar to competency management approaches, but it avoids their limitations (i.e., manual gathering of data, quickly outdated profiles) through automatic extraction techniques. Our approach initially makes use of core domain words extracted from either the documents themselves or from external sources such as domain thesauri. These domain specific general terms are high level concepts, representative for a domain, which are used to seed contextual extraction patterns.

As we are dealing with content from an IT organisation we make use of the ACM Computing Classification System[4] to manually identify domain specific general terms for the computer science domain. A list of about 80 such terms, a subset of which is shown in table 1, is extracted by a domain expert. Only the ACM subjects nouns, filtered by a stop word list and sorted in descending order of their frequency are considered.

A syntactic description of terms (i.e., nouns or noun phrases) is used to discover candidate expertise topics in the context of each domain specific general term. Two types of context patterns are used: noun phrases that include a domain specific general term or noun phrases introduced by the following pattern.

T Prep C

Where *T* stands for any domain specific general term, *Prep* stands one of the following prepositions: *for*, *to*, *of*, *on* and *C* stands for the candidate.

The features used for ranking expertise topics include length, frequency, acronyms and embeddedness (i.e., how many times is the expertise topic included in longer expertise topics). An external web search engine is used to filter candidates that are too specific, too general or misspelled from the final result list. We will not go into the details of the expertise topic extraction method as it is discussed in more detail elsewhere (Bordea and Buitelaar, 2010b). This approach is evaluated in the context of the keyphrase extraction task, achieving competitive results both compared with baselines and with other participating systems (Bordea and Buitelaar, 2010a).

3.2. Expert finding and profiling

In contrast to previous work on expert profiling, our method automatically extracts expertise topics which are added to expert profiles of individuals that authored or accessed a document. No distinction is made between individuals associated with a document, assuming that all the authors of a document have the same level of expertise. We make use of a relative measure of association strength that considers a person's expertise in comparison with the contributions of all the other organisation members. Each expertise topic is assigned a measure of relevance, computed using an adaptation of the standard information retrieval measure TF-IDF, called *TFIRF* (Bordea and Buitelaar, 2010b). A taxonomy of expertise topics allows us to analyse more sophisticated methods of expert profiling, such as coverage of different expertise topics related to an area and application of knowledge in different contexts.

4. Grounding expertise topics on the LOD cloud

Additional background knowledge, outside of a domain corpus, can be a useful source of information both for identifying domain concepts as well as for finding additional evidence of expertise. The Linked Open Data (LOD) cloud is a rich and continuously growing source where we can discover additional knowledge with respect to a domain (ontologies, thesauri) or other expertise traces such as scientific publications or patent descriptions. A first step in the direction of exploiting this potential is to provide an entry point in the LOD cloud through DBpedia, one of the datasources most widely connected in the cloud. Two naive but promising approaches for semantic term grounding on DBpedia are described and evaluated in section 5.2. Our goal is to associate as many terms as possible with a concept from the LOD cloud through DBpedia URIs and concept descriptions. Initially we find all candidate URIs using the following DBpedia URI pattern.

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

[4]ACM Computing Classification System: `http://www.acm.org/about/class/1998/`

Table 2: Perfect agreement results for expertise topic extraction

Answer	Top	Middle	Bottom
Good	**0.79**	0.18	0.09
Bad	0	0.06	**0.30**

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

http://dbpedia.org/page/Natural_language_processing

To ensure that only DBpedia articles that describe an entity are associated with an expertise topic we discard category articles and we consider only articles that match the *dbpedia-owl:title* or the final part of the candidate URI with the topic. Multiple morphological variations are discovered for an expertise topic, all of them are considered to increase the number of candidate DBpedia URIs found.

5. Evaluation

The enterprise dataset under analysis consists of 11,319 files subdivided into 3,319 folders and is composed of both structured and unstructured documents.The corpus contains a combination of word documents, excel spreadsheets, power point presentations, pdf documents and plain text files that span several years from 2003 to 2009 inclusive. In our first experiment we evaluate the extraction of expertise topics through a user study, then we present the results of expertise topics grounding on DBpedia.

5.1. Experiment 1: Expertise topic extraction

A user study with three participants who are domain experts is set up to evaluate the topic extraction method. Due to limited availability and strict time constraints, the domain experts are asked to evaluate a reduced list of 100 topics from top, middle and bottom of the ranked list of expertise topics. We expect a high number of correct topics at the top of the list and a high number of incorrect topics at the bottom of the list. Domain experts are given a list of shuffled topics selected in the following way: 34 topics from top ranked topics, 33 from middle ranked topics and 33 from bottom ranked topics.

The three judges are instructed to rate the expertise topics for the given domain by selecting one of three possible options for each topic: "good", "bad" and "undecided". Table 2 gives an overview of the user study results where all three annotators were in agreement. We only present topics considered correct and incorrect including the position where the topics appear in the ranked list (i.e. Top, Middle, Bottom). Almost 80% of the topics that are ranked high by our system are confirmed to be correct by all the three judges but only 30% of the topics ranked low are confirmed bad.

The kappa statistic is used to measure the agreement between the three judges. Only the expertise topics that are judged as good or bad (62 out of 100) are considered, ignoring all topics that are ranked as "undecided" by at least one participant. Kappa is in the moderate agreement range (0.61), but much higher agreement (perfect agreement for almost 80%) can be observed for the expertise topics at the top of the ranked list. The agreement rate is much lower for the topics ranked lower, indicating that the human judges have more difficulties distinguishing the quality of lower ranked topics.

Table 3: DBpedia URI extraction results

Approach	True P	False P	True N	False N
A1	93	4	82	7
A2	90	1	85	10

Table 4: Precision and recall for DBpedia URI extraction

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

5.2. Experiment 2: DBpedia grounding

Our second experiment aims at comparing two approaches for grounding expertise topics, the first approach (A1) using the expertise topic string as it appears in the corpus for matching URIs and the second approach (A2) using its lemmatised form. Stemming was also considered but this approach performed lower as stems increase the term ambiguity[5].

In order to evaluate our DBpedia URI discovery approach we built a small gold standard dataset by manually annotating the top 186 expertise topics with DBpedia URIs. Only about half of them have a corresponding concept in DBpedia, as we are dealing with a general knowledge datasource that has a limited coverage of specialised technical domains. The number of positive/negative matches is shown in table 3, where N and P stand for negative and positive results respectively. Although both approaches have similar results in terms of F-score, the A2 approach based on the lemmatised form of the expertise topics achieves better precision as can be seen in table 4.

To extract descriptions or definitions of concepts we rely on the *dbpedia-owl:abstract* property or the *rdfs:comment* property in the absence of the former. For now we are only interested in English definitions, therefore only triples tagged with *lang='en'* are considered. Even though English descriptions are available for a larger number of topics this tag is not always present, therefore we can only retrieve them for a smaller number of topics. Expertise topics that include an acronym (e.g. "NLG system" instead of "Natural Language Generation system") are more difficult to annotate with a URI as acronyms are often ambiguous.

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

Other general purpose data sources such as Freebase [6] or domain specific data sources can be linked in a similar manner. Another complex problem that we do not address in this work is the disambiguation of concept URIs but this has a limited impact on the results when dealing with specialised technological vocabulary .

6. Conclusions and future work

The work described in this paper is a first step in the realisation of a set of tools to assist in automatic content analysis for ECM platform configuration. The ranked list of topics, extracted by the expertise mining algorithm provides the information architect with a deeper understanding of the business domain of organisation.

Our approach allows Information Architects to gain a high-level view of the enterprise content and to identify the key employees that need to be involved in the ECM platform analysis interviews. In summary, both the expertise topic extraction and expert profile construction will streamline the creation of a taxonomy which is truly representative of the information contained in the ECM system. The expert finding solution described in this work can be integrated into existing ECM platforms, providing end users with the ability to explore semantic relationships between expertise topics, employees and enterprise documents.

Acknowledgment

This work is supported by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

7. References

Krisztian Balog and Maarten De Rijke. 2007. Determining expert profiles (with an application to expert finding). In *IJCAI 2007*, pages 2657–2662. Morgan Kaufmann Publishers.

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

Georgeta Bordea and Paul Buitelaar. 2010a. Deriunlp: A context based approach to automatic keyphrase extraction. In *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*.

Georgetas Bordea and Paul Buitelaar. 2010b. Expertise Mining. In *Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland*.

Delroy Cameron, Boanerges Aleman-Meza, I. Budak Arpinar, Sheron L. Decker, and Amit P. Sheth. 2010. Formal models for expert finding in enterprise corpora. In *Proceedings of the Fourth International Conference on Semantic Computing (ICSC), 2010 IEEE*, pages 333 – 340, Pittsburgh, USA.

Rod. Dilnutt. 2011. Surviving the Information explosion? *IEE Engineering Management*, 118(2):59, February.

Zornitsa Kozareva and Eduard Hovy. 2010. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1110–1118, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atif Latif, Muhammad Tanvir Afzal, and Klaus Tochtermann. 2010. Constructing experts profiles from linked open data. In *Proceedings of the 6th International Conference on Emerging Technologies (ICET)*, pages 33 – 38.

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

Bjørn Erik Munkvold and Anne Kristine Hodne. 2006. Contemporary Issues of Enterprise Content Management : The Case of Statoil. *Journal Scandinavian Journal of Information Systems*, 2(18).

Shinsuke Nakajima, Jianwei Zhang, Yoichi Inagaki, Tomoaki Kusano, and Reyn Nakamoto. 2009. Blog ranking based on bloggers knowledge level for providing credible information. In Gottfried Vossen, Darrell Long, and Jeffrey Yu, editors, *Web Information Systems Engineering - WISE 2009*, volume 5802 of *Lecture Notes in Computer Science*, pages 227–234. Springer Berlin / Heidelberg.

Paola Velardi Roberto Navigli and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1872–1877.

Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. 2010. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):1–38.

J.E. Scott. 2011. User Perceptions of an Enterprise Content Management System. In *hicss*, pages 1–9. IEEE Computer Society.

Pavel Serdyukov, Henning Rode, and Djoerd Hiemstra. 2008. Exploiting sequential dependencies for expert finding. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 795–796, New York, NY, USA. ACM.

Pavel Serdyukov, Mike Taylor, Vishwa Vinay, Matthew Richardson, and Ryen White. 2011. Automatic people tagging for expertise profiling in the enterprise. In *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 399–410. Springer Berlin / Heidelberg.

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