Expertise Extracting Within Online Shared Workspaces

Peyman Nasirifard and Vassilios Peristeras
Digital Enterprise Research Institute
National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
firstname.lastname@deri.org

Abstract. We present an approach for identifying people with expertise in certain areas from online research-oriented shared workspaces in three steps: a) Content analysis of stored documents (mainly scientific deliverables) within shared workspaces; b) Log files analysis; and c) Assigning relevant expertise to users using dynamic SPARQL queries.

1 Introduction and Related Work

Online shared workspaces (e.g. BSCW, Microsoft SharePoint) aim to ease collaboration among various types of users in different time zones (e.g. by sharing resources). When people collaborate using shared workspaces, they leave some latent document-based history. The document-based history contains the events that happened on a specific document (e.g. read event, create event). Most shared workspaces log such events and transactions and are able to export them in various formats (e.g. CSV). These log files which can be seen as fingerprints of users contain rich information and can be further analyzed and utilized using various reasoning techniques.

In this paper, we present an approach for assigning semi-automatically expertise elements to users of online shared workspaces taking to account the documents that they read, created or revised.

[3] discusses some possible ideas and challenges in expert finding. They discuss that semantics is an important factor in detecting expertise for selecting peer-reviewers. As an example, if someone is familiar with SIOC related material, perhaps s/he is also familiar with other pieces of the Semantic Web domain. The need for taxonomies for expertise has been also addressed in some other works like [6]. Similar to our approach, [5] introduces a topic-centric expert finding system from open access metadata and full text documents. [1] presents models for searching an organization’s document repositories for experts on a given topic using generative probabilistic models and in [2] they find similar experts by measuring similarity between term vectors. [4] uses also edit history of Wikipedia users in order to build expertise profiles for them.

Due to space limitation, we do not present a comprehensive list of related work. Our approach for finding expertise uses log files of shared workspaces to
exploit the internal interactions among users and to our best knowledge, this is the novel part of our work.

2 Approach

Our approach uses an online research-oriented shared workspace as a platform for mining expertise. We extract and assign expertise in three steps which are presented in the following. Figure 1 demonstrates the overall view of our approach.

The first step is key phrase extraction. We use documents and in particular scientific deliverables as inputs for key phrase extraction. The main reason behind this policy is the fact that deliverables are mature enough to be used for further processing, as they contain the final (research) results of project members.

Normally, (scientific) European Integrated Projects (IP) define some templates that are used by consortium as a way to standardize the structure of the documents. In the projects that we were involved and/or had a look at their deliverables (e.g. Ecospace, inContext, Nepomuk), we noticed that they followed the similar templates. As an example, each document (deliverable) has a version history that contains some metadata and contribution statistics of partners (who wrote/changed which parts). However, such information has potential to be used for further processing and perhaps finding expertise, but unfortunately, processing those unstructured data (e.g. in the form of tables) offers many challenges, such as in text processing and name ambiguity issues as well. Meanwhile, some users may unintentionally forget to update the tables or provide sufficient useful information. In addition to these arguments, such information is not rich enough and does not include fine-grained material and concepts that could be potentially extracted using content analysis techniques. In brief, the result of
first phase (key phrase extraction) is a list of key concepts of a document that are later assigned to users as expertise elements based on their interactions (log file analysis). For a better performance, we store the results, as we need to query them later. We decided to use RDF data model to be able to use SPARQL on top of that. We store each document name plus its extracted key phrases and their confidences in RDF repository. The confidence of a key phrase is a value that is generated by key phrase extractor and can be used for ranking the phrases. In other words, it is a value that determines whether the extracted phrase can be a real key phrase or not (based on Natural Language Processing (NLP) techniques). We have a filtering phase after extracting key phrases. The reason is rather based on experimental results and will be explained in section 4.

The second phase of expertise extracting is log file analysis. Log files of shared workspaces are rich structured data and can be further analyzed. Log files can be seen as a repository of transactions history within a project. They contain log records which keep the document-based events (e.g. read, create). We assume that each log record includes user ID, object ID and event name at minimum. In natural language, it can be translated to a user with specified ID performed the specified event on an object with specified ID. However, in most cases a log record contains more information such as temporal aspects of the record, description of the record, etc., but we do exclude these fields from further processing. Currently, we focus mainly on three document-based events which can affect expertise profiles: Create, Revise and Read. We transform each log record into a RDF statement in order to use SPARQL queries on top of that. To do so, we did a simple mapping between log records and RDF concepts. The user ID will be mapped to RDF:subject, the object ID will be mapped to RDF:object and the event name will be mapped to RDF:predicate. Obviously, a namespace will be added to all elements. We have a filtering phase before processing log records. The reason is mainly due to experimental issues and will be addressed in section 4.

For the final step, our main assumption is that if somebody creates or revises a document with topic X, then s/he has more expertise on topic X than a person that only reads that document. Towards this direction, we assign two different expertise granularities to a user: expert in and familiar with. A user is expert in topic X, if s/he created or revised a document that contains topic X. A user is familiar with topic Y, if s/he just read a document that contains topic Y. This step uses SPARQL queries which are built dynamically (using user IDs of shared workspace) as a mean for matching expertise based on the already-extracted information. We filter also the result of this step to restrict repetitive terms, as various documents may contain similar key phrases.

3 Prototype and Technical Issues

We developed a simple prototype for extracting and assigning expertise elements to users using BSCW shared workspace. As a use case, we used the data provided by the Ecospace project. Ecospace is an European Integrated Project (IP)
in the area of Collaborative Working Environments (CWE). As inputs for extracting key phrases, we used around fifty documents (mostly deliverables) of the Ecospace project in PDF format, which were submitted to four review sessions. On average, forty key phrases with various confidences (between 0.3 and 0.8) were generated for each deliverable. BSCW enables users to export the transactions history in CSV. As a testbed, we used the log files of the Ecospace project from March 2005 to December 2008 with more than 30,000 log records. More than 180 users (active and inactive) were extracted from log file. The prototype is very simple and enables end users to search through expertise and find relevant experts. We used Sesame 2.0\textsuperscript{1} RDF store. The prototype is accessible online\textsuperscript{2}.

4 Evaluations and Discussions

Due to unknown reasons, BSCW log files contain some noisy records. Noisy log records are those ones that do not include user ID, object ID or event name. In order to remove noisy log records, we did a filter (purify) phase before using them. To do so, we defined some patterns and those records that do not follow the patterns will be automatically removed from the process.

We did a simple experimental evaluation for our approach. We asked 12 members of the Ecospace project to have a look at their extracted expertise. All of them confirmed that the extracted phrases are relevant to the concepts that they are familiar with or expert in. But they also complained about several issues which are described in the following.

Meaningless expertise: We used a free open source key phrase extractor for our purpose. As far as we know, there exists no error-free automated keyword extractor. Thus, we used the confidence values as a threshold to get the phrases that have a bigger probability to be a meaningful key phrase. Meanwhile, we had to filter (purify) the extracted key phrases to remove a) top 572 common English terms\textsuperscript{3} (i.e. stop words); and b) repetitive terms.

Organization expertise profiles: One interesting result that we observed is the fact that in some cases - projects with many partners (organizations) - there is a person (broker/proxy) in an organization, who has the larger amount of activities in the shared workspace. In such cases, the broker may acquire more expertise than the real contributors of a document. To overcome this situation and in order to make the results more reliable, we may build an expertise profile for an organization by unifying the expertise of all members of that organization.

Similar phrases: Some phrases were conceptually the same, but reported several times. As an example, collaboration and collaboration concepts rather refer to the same concept. However, we filtered the expertise to remove repetitive phrases, but due to lack of semantics for terms (e.g. taxonomy), such behaviors were expected. One partial solution to this problem could be using WordNet and/or DBpedia to infer the semantics of the terms and merge relevant terms.

\begin{itemize}
\item \textsuperscript{1} http://www.openrdf.org/
\item \textsuperscript{2} http://purl.oclc.org/projects/expertui
\item \textsuperscript{3} http://members.unine.ch/jacques.savoy/clef/
\end{itemize}
Irrelevant expertise: Currently, all extracted expertise of a document is assigned to a user, who contributed just part of that. To tackle this issue and remove irrelevant expertise, we may utilize version history of the shared workspace to get the exact contribution of a user (e.g. by using diff).

5 Conclusion and Future Work

In this paper, we presented an approach for mining expertise from online shared workspaces using log files and content analysis of stored documents. We also presented briefly our prototype that uses BSCW. We believe that what we presented in this paper, is the first step of a long way for extracting expertise profiles from shared workspaces. There exist many directions and improvements for the future work. Clustering users based on their expertise is one motivating use case. In this case, we may identify groups of people that can possibly collaborate together (e.g. writing future proposals). Introducing temporal aspects (e.g. validity period of expertise) is one of the possible future improvements. We may enable end users to change the threshold of confidences in order to increase/decrease the scope of expertise.

Acknowledgments. We thank Alexander Schutz for his kind support and providing keyword extraction package. The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/11380 (Lion-2) and the European Union under Grant No. FP6-IST-5-35208 (Ecospace).

References