Extracting and Utilizing Social Networks from Log Files of Shared Workspaces

Peyman Nasirifard, Vassilios Peristeras, Conor Hayes and Stefan Decker

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

Abstract[1]. Log files of online shared workspaces contain rich information that can be further analyzed. In this paper, log-file information is used to extract object-centric and user-centric social networks. The object-centric social networks are used as a means for assigning concept-based expertise elements to users based on the documents that they created, revised or read. The user-centric social networks are derived from users working on common documents. Weights, called the Cooperation Index, are assigned to links between users in a user-centric social network, which indicates how closely two people have collaborated together, based on their history. We also present a set of tools that was developed to realize our approach.

1 Introduction and Background

Online shared workspaces (e.g., BSCW[2], Business Collaborator[3], and Microsoft SharePoint) provide necessary tools and technologies for users to share various objects, synchronize them and collaborate together. When people collaborate within shared workspaces, they leave some fingerprints. These fingerprints may vary from events that happen on a document (e.g., read, revise, delete) to inviting a new member to the shared workspace. Most shared workspaces log these fingerprints and are able to export them in different formats. These log files contain valuable information and reflect the behaviors of users.

From one perspective, online social networks can be divided into two main groups[4]: object-centric and user-centric (i.e., ego-centric). In object-centric social networks, an object (e.g., document, video, music) connects people together, whereas in user-centric social networks, users are directly connected to each other. In this paper, we present an approach to use log files of online shared workspaces for

[1] An earlier version of this work was published in the paper [Nasirifard, P., Peristeras, V.: Expertise Extracting Within Online Shared Workspaces. In: Proceedings of the WebSci'09: Society On-Line (2009)]

[2] http://www.bscw.de/

[3] http://www.groupbc.com/

[4] http://www.zengestrom.com/blog/2005/04/why_some_social.html

extracting social networks among users. We use the extracted object-centric social network for assigning expertise. As well as using the extracted expertise as a dynamic approach for building inter- and intra-organization level expertise profiles, it can be also used to build teams that require specific expertise. We approach the hidden user-centric social network as a weighted graph. We call these weights Cooperation Indices. Cooperation Index is a factor that determines how closely two people work together and it can be used as a light-weight recommendation system in access-control mechanisms or for finding proxies.

In this paper, we also present the prototypes that we have developed: Holmes extracts the user-centric social network and calculates Cooperation Indices from log files of the BSCW shared workspace and Expert Finder extracts and assigns expertise elements to users of the BSCW shared workspace.

Before explaining our approach and demonstrating the tools, we present a brief overview of Semantic Web technologies. The Semantic Web [1] is an effort to ease interoperability among applications by providing standards for data representation and exchange (e.g., ontologies). The Resource Description Framework (RDF), which provides the grammar for the Semantic Web, is an important factor to enable this approach. RDF is a data model and supports the notion of subject-predicate-object; an RDF Triple. Recently, some shared workspaces (e.g., BSCW, BC) have started to export data in RDF. We decided to use the RDF data model in our approach. This eases the extension of our approach to different shared workspaces, as they are or will be RDF-aware. Moreover, using RDF enables other application developers to use our data and results in their own applications and/or mash-ups. A query language is required to query the RDF data. There exists some query languages for RDF, the most well-known being SPARQL[5] which was recently released as a W3C Recommendation. SPARQL is used in our work.

2 Related Work

For extracting social networks, we use a closed world called online shared workspaces, where various users (of a single or multiple projects) are able to share documents and collaborate together. In particular, we use the log files of shared workspaces, where all document-based events are stored.

Studying the relationships among people in a subset of the open environment (e.g., an online community, forums, mailing lists) or in a closed world (e.g., email), where the access to data is restricted to a specific person or a group of people, has attracted some researchers. Culotta et al. [2] present a system that extracts the users' social network by identifying unique people from email messages and finding their homepages and filling out the fields of a contact address book. Adamic et al. [3] present social network analysis of the Club Nexus online community. Xobni[6], which is a Microsoft Outlook plugin, is a search and navigation tool for Outlook inbox. It is able to follow the email discussions and generate the social network of email senders and receivers. Chang et al. [11] used blogs as a means for social learning and analysis.

[5] http://www.w3.org/TR/rdf-sparql-query/
[6] http://www.xobni.com/

Nurmela et al. [4] studied the log file of a groupware environment and demonstrated how the social network analysis approach can be used as a method to evaluate the social level structures and processes of a group studying in a Computer Supported Cooperative Learning (CSCL) environment. De Choudhury et al. [5] use social context to predict the information flow and introduce some parameters that play important roles in information flow. Demsar et al. [13] present coFinder, which crawls the Web for finding potential collaboration opportunities.

Finding experts and expertise have been also studied in many domains and platforms such as emails [6], Wikipedia [7], mailing lists [8], online communities [9], question-answering services [10], etc. There are many use cases for finding appropriate experts (e.g., recommendation systems for scientific and industrial activities). Our approach uses log files of online shared workspaces for extracting expertise.

3 Extracting and Using Social Networks

In this section, we present our model for extracting (object-centric and) user-centric social networks from log files of shared workspaces.

3.1 Object-Centric Social Networks

As stated in [12], a log file is composed of several log records and in each record, we assume that user ID, event name and object ID exist as minimum. User ID is the unique identification of a person that performs an event on an object that is also uniquely identifiable. For example, a log record in natural language can be *Person with ID 123 revised the document with ID 456*. In addition, log records can contain more information, such as description of the records, temporal aspects (e.g., time-stamps) of log records, etc.

Building an object-centric social network from log file is quite straightforward. We translate the log records into RDF triples and store them in RDF store. In order to do this, we map the main elements of the log records to RDF concepts. The user ID of a record is mapped to RDF subject; the event is mapped to RDF predicate and the object ID is mapped to RDF object.

We approach the log file or extracted RDF triples in a document-centric perspective, which results in virtual clouds containing a document in the middle and several users around the document, who have performed various events on that document, as illustrated in Fig. 1. We use dynamic SPARQL queries for building such clouds from RDF repository. A document-centric perspective of a log file does not make sense, unless it is used for a useful use case.

3.2 Using Object-Centric Social Networks

As object-centric social networks, these document-centric clouds may be used as a means for extracting *expertise*. We define *expertise* as a piece of knowledge that has

been acquired by a person in the past. We extract and assign expertise in three steps. For more information, refer to [12].

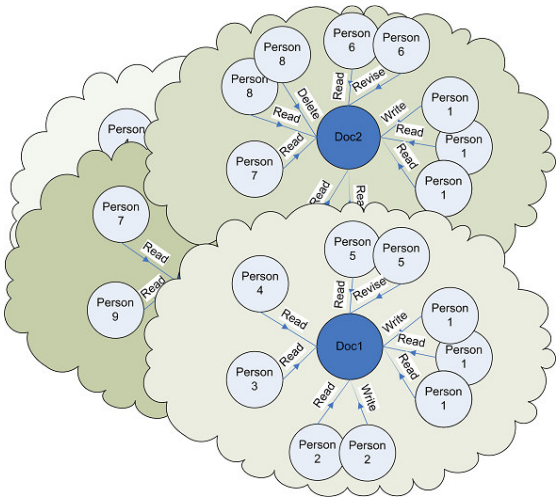

Fig. 1. Document-centric perspective of log file

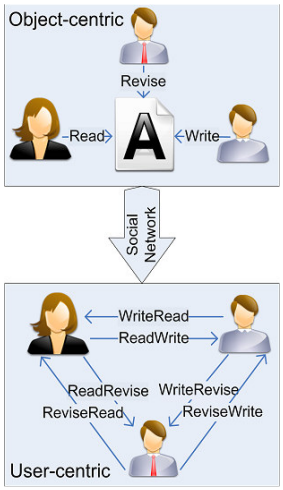

Fig. 2. From object-centric to user-centric social network

3.3 User-Centric Social Networks

At this point we need to create a user-centric perspective, i.e., social relationships among users, where actions on objects (e.g., documents) connect people together. To obtain the desired user-centric social network, we remove the objects and connect people directly together based on the events that the users performed on a specific object. In other words, we combine the RDF predicates in order to build a user-centric social network, where people are directly connected together. Fig. 2 shows the overall approach of building a user-centric social network using an object-centric one.

The relationships between people are defined by a combination of events on objects. Generally, real-life social relationships are transitive. We noticed that we may also make the event-based relationships transitive by enabling users to traverse across document-centric clouds. Thus, the depth of event-based relationships is also important (i.e., moving from one cloud to another one). As an example, if user A has created a document which has been revised by user B and user B has read another document which has been deleted by user C, the depth of possible relationship between user A and user B (CreateEvent;ReviseEvent[7] is one, whereas the depth of the possible relationship between user A and user C is two (CreateEvent;ReviseEvent;ReadEvent;DeleteEvent), as two different documents were in the middle which were connected via user B. Note that all relationships are unidirectional. We do not store the user-centric social network in the RDF repository as these are built on-the-fly using SPARQL.

[7] We use semicolon (;) as a separator between events.

3.4 Assigning Weights to User-Centric Social Network

The extracted user-centric social network needs to be weighted. We call each weight of the user-centric social network a Cooperation Index; an index that determines how closely two people work together. The higher the index is, the greater the collaboration history between two people. To calculate the index, we assign user-defined weights to the relationships between people and we sum up the frequency of relationships with consideration of weights. Due to space limitations, we do not present formal definitions, but just a use case scenario.

In the following, we provide an example for calculating Cooperation Index (for depth one). In our example, we have two users: Alice and Bob. They work together using the BSCW shared workspace and they take part in some document-based events (e.g., read, create) which then BSCW exports in CSV format. For simplicity, we just considered 8 records of the log file. Listing 1 shows this piece of log file. Of course, the greater the number of entries, the more accurate the Cooperation Index will be.

Listing 1. A piece of a sample log file

```
2007-03-10 17:17:55;337777;CreateEvent;11;D1;2;Bob;
2007-03-11 17:19:15;333481;CreateEvent;13;D3;1;Alice;
2007-03-16 09:13:22;335481;ReadEvent;13;D3;2;Bob;
2007-03-17 12:17:56;385481;ReviseEvent;13;D3;2;Bob;
2007-03-17 13:17:45;337431;ReadEvent;12;D2;2;Bob;
2007-03-17 14:19:35;332581;ReviseEvent;12;D2;1;Alice;
2007-03-17 16:10:25;346541;ReadEvent;12;D2;1;Alice;
2007-03-18 13:25:15;312431;ReviseEvent;11;D1;1;Alice;
```

Each log record/entry starts with temporal information; followed by event ID, event name, object ID, object name, user ID and user name. For simplicity, we did not present other elements of the actual BSCW log records.

Suppose that Alice wants to calculate her Cooperation Index at depth one with Bob. To do so, we should calculate the possible relationships between Alice and Bob taking into account the documents in the middle. In the following, such relationships are presented:

- Relationship regarding to D1: *ReviseEvent;CreateEvent*
- Relationships regarding to D2**:** *ReadEvent;ReadEvent* and *ReviseEvent;ReadEvent*
- Relationships regarding to D3: *CreateEvent;ReadEvent* and *CreateEvent;ReviseEvent*

In the next step, we should set user-defined weights for the relationships. Here users can decide what types of relationship are more important depending on the context of the specific common project or collaboration. In other words, it is the user that decides what types of relationship should have more effect and influence on calculation. In our example, Alice assigns the following weights to her possible relationships with others:

- CreateEvent;ReviseEvent = 0.4
- ReviseEvent;CreateEvent = 0.2
- ReviseEvent;ReviseEvent = 0.4

Due to space limitation, we did not present the relationships with the weight zero. Now, based on the relationships between Alice and Bob and also the weights assigned by her, we calculate the Cooperation Index by counting the frequency of the relationships between Alice and Bob for depth one with consideration of weights. We reach the value 0.6 for this Cooperation Index.

4 Prototypes

We have developed several prototypes to realize our approaches. The prototypes are based on a Service-Oriented Architecture (SOA). In SOA, business processes are packaged as services and are accessible via end points to end users. We used Apache CXF[8] in order to develop Web services. The User Interface (UI) of the prototypes is powered by JSP. We used OpenRDF[9] Sesame 2.0 as RDF store. The tools use the data provided by the Ecospace[10] project. 183 users were extracted from the log file.

As stated in [12], **Expert Finder** is a simple prototype for extracting and assigning expertise elements to users of BSCW shared workspace. Fig. 3 demonstrates some snapshots of Expert Finder. The prototype is accessible online[11].

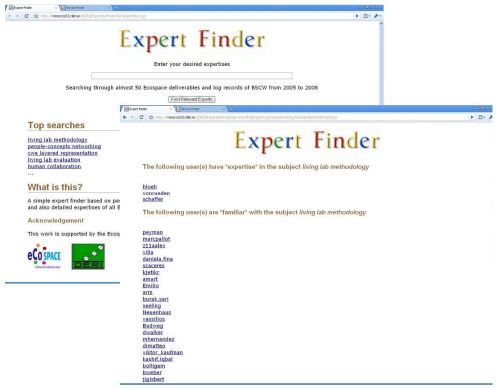

Fig. 3. Some snapshots of Expert Finder

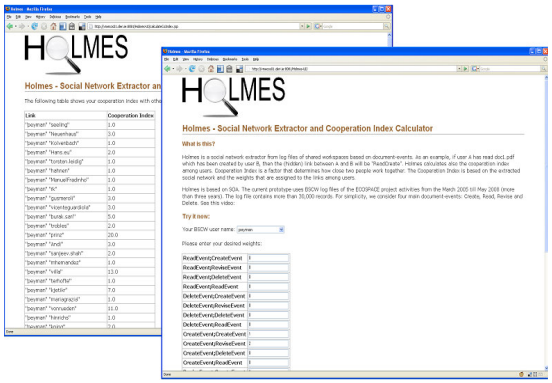

Fig. 4. Some snapshots of Holmes

Holmes extracts the user-centric social network from log files and calculates Cooperation Indices between users. Fig. 4 demonstrates some snapshots of Holmes. Ecospace users are able to select their user names from a select box, assign desired weights to their relationships and calculate the Cooperation Indices with the rest. To reduce the overhead for users, we assigned predefined weights to all relationships. In

[8] http://cxf.apache.org/
[9] http://www.openrdf.org/
[10] Ecospace is a European Integrated Project (IP) in the area of Collaborative Working Environment (CWE). For more information refer to http://www.ip-ecospace.org/
[11] http://purl.oclc.org/projects/expertui

our case, *CreateEvent;ReviseEvent* and *ReviseEvent;CreateEvent* have bigger weights. The Holmes prototype is accessible online[12].

5 Discussions and Evaluations

Using a query-based approach (i.e., SPARQL) enabled us to extract some interesting facts from existing log records. For example, by counting all triples that contain a certain user as a RDF subject and with the consideration of document-centric approach for social networks, it can be inferred who has read the most amount of documents during the life cycle of the project, who has revised the most, who has deleted the most, etc. These statistical results can be used to determine the most active persons during the time period of a project. The *active* person can be defined as a person, who carries out more document events than others taking into account the time intervals.

Besides statistical results, one of the interesting outputs was a dynamic approach for visualizing the social networks of users, based on document-events. To do so, we built a simple mash-up from RDF triples generated by queries to NetDraw[13] input format. NetDraw is a free social-network visualizer tool, which accepts plain text files as input. Due to space limitations, we do not present the snapshots.

We conducted a simple experimental evaluation for expert finding approach. The result can be found at [12]. We also conducted a simple experimental evaluation for Cooperation Index approach. We asked 12 participants of the Ecospace project to have a look at their extracted social networks. All of them confirmed that the presented results were relevant to them. They had also some suggestions: Currently, for calculating the Cooperation Index, we considered four main document events (i.e., Create, Revise, Delete, and Read) and only relationships at a depth of one. These events can be simply extended to cover more document events as well as deeper depths. One important issue that may arise with more types of events and deeper depths is to combine events and assign weights to them, which in some cases can bring overhead for users. In a more complex model for calculating Cooperation Indices, different weights can be posed to documents based on their importance for the collaboration process. So, some documents may have bigger weights assigned by their creators and this could be then taken into account when calculating the Cooperation Indices.

6 Conclusion and Future Work

In this paper, we presented an approach for extracting and utilizing social networks from log files of shared workspaces. We used the extracted social network for two main use cases: Finding expertise and calculating Cooperation Index, which can be seen as a weighted user-centric social network. Cooperation Index is a benchmark that

[12] http://purl.oclc.org/projects/holmes
[13] http://www.analytictech.com/

determines how close two people collaborate together. We demonstrated also our prototypes (Holmes and Expert Finder).

Besides the points mentioned by the evaluators, we plan to benefit from temporal aspects of log files to enable users to calculate the Cooperation Indices within a specific time period. Currently, the Cooperation Indices are calculated during the life cycle of the project.

Acknowledgments. The authors thank Alexander Schutz and Marco Zuniga. The work presented in this paper has been funded in part by SFI under Grant No. SFI/08/CE/I1380 (Lion-2) and the EU under Grant No. FP6-IST-5-35208 (Ecospace).

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, May (2001)
2. Culotta, A., Bekkerman, R., McCallum, A.: Extracting social networks and contact information from email and the Web. In: First Conference on Email and Anti-Spam (2004)
3. Adamic, L.A., Buyukkokten, O., Adar, E.: A social network caught in the Web. First Monday, 8(6) (2003)
4. Nurmela, K., Lehtinen, E., Palonen, T.: Evaluating CSCL log files by social network analysis. In: CSCL '99: Proceedings of the conference on Computer support for collaborative learning. International Society of the Learning Sciences (1999)
5. De Choudhury, M., Sundaram, H., John, A., Seligmann, D.: Dynamic prediction of communication flow using social context. In: HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, pages 49-54. ACM, New York, NY, USA (2008)
6. Balog, K., De Rijke, M.: Finding experts and their details in e-mail corpora. In: WWW '06: Proceedings of the 15th international conference on World Wide Web (2006)
7. Demartini, G.: Finding experts using Wikipedia. In: Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics (2007)
8. Chen, H., Shen, H., Xiong, J., Tan, S., Cheng, X.: Social network structure behind the mailing lists. ICT-IIIS at TREC 2006 Expert Finding Track. In Fifteenth Text Retrieval Conference (TREC) (2006)
9. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, pages 221-230, New York, NY, USA (2007)
10. Liu, X., Croft, W.B., Koll, M.: Finding experts in community-based question-answering services. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, pages 315–316, New York, NY, USA (2005)
11. Chang, Y., Chang, Y., Hsu, S., Chen, C.: Social network analysis to blog-based online community. In: ICCIT '07: Proceedings of the 2007 International Conference on Convergence Information Technology (2007)
12. Nasirifard, P., Peristeras, V.: Expertise extracting within online shared workspaces. In: Proceedings of the WebSci'09: Society On-Line (2009)
13. Demsar, D., Mozetic, I., Lavrac. N.: Collaboration opportunity finder. In: Virtual Enterprises and Collaborative Networks, pages 179-186, Springer (2007)