A KNOWLEDGE SEARCH FRAMEWORK FOR COLLABORATIVE NETWORKS

Rui J. Tramontin Jr., Ricardo J. Rabelo
Federal University of Santa Catarina, Department of Automation and Systems
GSIGMA – Intelligent Manufacturing Systems Group
Florianópolis (SC), BRAZIL
tramontin@gsigma.ufsc.br
rabelo@das.ufsc.br

Partners in a collaborative network (CN) must interact and share their knowledge. This paper focuses on how the knowledge exchanged in these interactions can be easily and precisely reached by CN partners through an knowledge search framework. In order to improve the search precision, ontologies are used to add metadata (semantic annotations) to documents as well as to provide support for semantic queries. This paper is also concerned to the CN-related requirements not addressed by current information retrieval solutions, like annotations consistency maintenance, multiple ontologies management, ontology mappings management and security issues. The proposed framework is being developed under the scope of the ECOLEAD Project as part of a broader ICT infrastructure for supporting CNs.

1 INTRODUCTION

During the last years, globalization, mass customization, and other new trends have led organizations to focus on new strategies in order to improve their agility, and thus to achieve a competitive advantage in the global market. Such strategies are increasingly based on collaboration forms classified as *collaborative networks*. According to Camarinha-Matos (2005), a collaborative network (CN) is constituted by a variety of entities (organizations and individuals) that are largely autonomous, geographically distributed, and heterogeneous. These entities work together in order to increase their access to business opportunities, share risks, reduce costs and achieve goals not achievable individually, and their interactions are supported by computer networks.

In order to take advantage of the CN paradigm, a number of challenges are faced by partners, coming from multiple perspectives, like technological, semantic, social and business (Chituc, 2005). These challenges include, but are not limited to: maintenance of privacy and visibility levels (security) as CN partners are autonomous; overcoming cultural and legal barriers as they work differently of each other and are placed in different countries; and dealing with semantic and interoperability issues, as partners and systems are heterogeneous. Above all these perspectives, a fundamental issue and challenge is the establishment of mutual trust among partners.

Working collaboratively implies to share information¹, resources and responsibilities among entities that jointly plan, implement, and evaluate a program of activities

¹ *Information* is defined as the result of manipulation and organization processes based on data gathered in a given context. *Knowledge* is defined as a complete set of information and its relations that lead people to take decisions and to create new knowledge (Huber, 1991).

to achieve a common goal (Camarinha-Matos, 2006). However, a facet of this is related to the sharing of knowledge, still very low exploited but which has gained relevance as partners realize how much they can benefit with this higher-level of collaboration. *Knowledge sharing* among CN partners is becoming of paramount importance in collaboration, being useful not only at the operational level – helping partners to achieve their goals –, but also at the tactic and strategic levels – helping managers to take decisions and plan future actions.

Collaborative tools, like CSCW², provide support to the interactions among partners and so enable knowledge sharing among them. Such tools also store the exchanged knowledge in several information sources, like documents, mailing lists, forums, blogs, etc. On the other hand, CSCW-like tools do not solve the problem of how to access, to search and to filter the knowledge – which is distributed (over CN members) and in diverse formats, sources and has different meanings –, in an efficient and smart way.

This paper proposes a knowledge search framework for CNs. Such framework is supported by ontologies and deals with CN-related requirements not addressed by current information retrieval solutions, like semantic annotations consistency maintenance, multiple ontologies management, ontology mappings management and security issues. The proposed framework is being developed under the scope of the ECOLEAD³ Project as part of an ICT infrastructure (ICT-I) for supporting CNs. Such infrastructure intends to support general (horizontal) requirements needed by any form of CN: people collaboration, systems interoperation, knowledge sharing, process synchronization, among others, in a secure, transparent, on demand and pay-per use way. The ECOLEAD ICT-I has been developed based on open platform-independent specifications and ICT standards, and in a Web Service Oriented Architecture (Rabelo, 2006).

Although the research on many aspects related to CN is of great value, this work is exclusively concerned with knowledge search aspects, such as semantic and security issues. Issues like trust building or interoperability are out of the scope.

The paper is organized as follows: section 2 describes the problem related to knowledge search in CNs; section 3 presents the envisaged framework; section 4 brings the implementation of a first prototype, and finally; section 5 presents the final remarks and the next steps.

2 PROBLEM DESCRIPTION

This section depicts the characteristics related to the problem of knowledge search in CNs and identifies the supporting techniques as well as the challenges and requirements involved.

2.1 Characteristics

The problem of knowledge search in CNs can be described through three main characteristics:

(I) There are multiple CNs, and each CN has its own shared knowledge available (documents, chat logs, mailing lists, etc.). Hence, **security** aspects should be taken into account in order to avoid unauthorized access from external members of the

² Computer Supported Cooperative Work

³ www.ecolead.org

CNs. For example, a post registered in a web forum of a given CN can be only accessible by its partners. Additionally, as CN partners are independent and autonomous, it is also desirable for them to be able to define access levels for the knowledge they are making available. This allows parts of a given document to be kept private for some partners, or even the whole document itself.

(II) CN partners can take part in several CNs simultaneously. As a generalization of the characteristic previously presented, partners are able to have access to knowledge from all CN they are involved in. Here, security mechanisms should again be used to guarantee access as well as to allow partners to perform **federated searches** in the related CNs. Federated search (also known as *distributed information retrieval*) links multiple search engines into a single virtual system (Si, 2005), allowing users to search in multiple systems with a single search request. For example, a partner participating in three CNs, A, B and C, can perform a search in all of them or in just one.

(III) Each CN contains its own way of representing its semantics (domain knowledge). The terms used in a given CN (or even in a single company) may have different meanings in other CNs. This aspect causes some problems when a partner wants to search in all CNs (s)he is involved in. In this case, **semantic translations** should be done in the original search query in order to convert it into different – but equivalent – queries. The results of every query should be translated back to be presented in a unified and understandable format.

2.2 Techniques

As knowledge is explicitly represented (stored in information sources) the application of Information Retrieval (IR) techniques is the most natural approach towards the implementation of a knowledge search framework. IR is a very broad research area and, in general, its main goal is to provide means for searching information in documents or searching for documents themselves. Traditional implementations of these techniques are the Web search engines, notably Google. Search engines are based on keyword queries, meaning that the retrieved documents contain the keywords specified in the query. Although such techniques provide some support for semantics, like search for synonyms and spelling correction, they are not enough to deal with the characteristics of knowledge search in CNs.

From another point of view, one of the new trends in the IR area is the usage of **ontologies** in order to improve the effectiveness of information search (Kiryakov, 2003), (Stojanovic, 2005), (Köhler, 2006), and thus to help retrieving knowledge. Ontologies are being used in a vast range of applications to represent semantics in a formal way. In this regard, one important assumption that is not only related to CNs, but also to the whole IR area, is that information sources express their meaning using a well-defined semantics. In this sense, it is assumed that the domain knowledge of CNs (mentioned in section 2.1, characteristic III) is formally represented by ontologies, whose importance in this context is three-fold:

1. Allows the definition of **semantic annotations** in information sources. With well defined semantics expressed in a machine interpretable way, information sources can be more precisely processed and searched. The idea of annotating document content with semantic information from domain ontologies was proposed

by the Semantic Web⁴ initiative (Berners-Lee, 2001), and it has been implemented in a number of applications. Annotations formally identify concepts and relations between concepts and are primarily intended for use by machines. One approach for that is to assign links between the document content and instances of an ontology (Figure 1), as defined by Kiryakov (2003).

Figure 1 – Semantic annotations pointing to some instances (in white) of an ontology fragment (in grey). [Based on Kiryakov (2003)]

2. Allows the definition of **semantic searches**. Based on ontologies, queries can express precisely what kind of knowledge is going to be searched, which in turn is retrieved based on the semantic annotations previously indexed. A simple example of semantic query is: search for documents that contain "*IT Companies located in a Location called 'Europe'*", i.e., documents containing annotations related to instances of the class "IT Company" that are located in "Europe", which in turn is an instance of the class "Location" (Figure 1). The result of this query would include the document presented in Figure 1, where "Comarch" is an "IT Company" located in "Poland" (which is part of "Europe").

3. Provides **semantic translation** support during *federated searches*. In this case, the search engine (also called meta-search engine) translates the original query to queries that are based on the ontologies used by the CNs involved in the search. Each new query is then executed in its related CN and the results are translated and presented to the user in a unified way. The query translation among the different contexts can be supported by **ontology mappings**. Ontology mapping (or matching) is a promising solution to the semantic heterogeneity problem, aiming at finding correspondences between semantically related entities of different ontologies, and thus enabling the data expressed in the matched ontologies to interoperate (OntologyMatching, 2006). This research area involves a large number of fields, ranging from machine learning, concept lattices and formal theories to heuristics, database schema and linguistics (Kalfoglou, 2003).

⁴ <http://www.w3.org/2001/sw/>

2.3 Requirements and challenges involved

Based on the presented characteristics and the related techniques, some particular requirements of CNs have been identified:

1. Maintenance of multiple ontologies. In order to support semantic translations between CNs, mappings between ontologies should be identified and managed. Semi-automatic facilities are required.
2. Ontology evolution. In a dynamic environment such as CNs, ontologies tend to evolve and so the consistencies of both semantic annotations and ontology mappings should be maintained.
3. Document evolution. In a similar case, when a document evolves, the consistency between this document and its annotations should be assured.
4. Automatic annotation. It is a fundamental aspect considering a large collection of documents, as it occurs in CNs. Automatic or semi-automatic support for annotation can be provided using IE (information extraction) and other related NLP (natural language processing) techniques. Based on the assumption that there will be multiple ontologies, automatic annotation techniques should deal with ambiguities.
5. Usability. The process of search should be easy-to-use for non-technical users. The main drawback of semantic queries is that they are more suitable for specialists that know ontologies than for ordinary users. Additionally, users must know which ontologies they can use and which classes, properties and relations are more appropriate for the search.
6. Security. It is an orthogonal requirement that should be as transparent as possible when controlling the access to ontologies, information sources or the scope of federated searches.

Although some of these requirements are fulfilled by existing solutions, some others still have open research questions or represent significant research challenges. For example, there are several techniques and implementations for automatic annotation of documents; however, as stated by Uren (2006), there is little support for ontology evolution and consistency maintenance. Another relevant challenge is related to the usability of semantic searches, i.e., search queries should be constructed in an easy way. Finally, it is important to highlight that current contributions for each of these requirements were made in isolated areas, not considering the inter-dependencies and integration among them.

3 THE ENVISAGED FRAMEWORK

Based on the problem description presented in the previous section (its characteristics, techniques and requirements) a conceptual framework for knowledge search was envisaged. It provides four main functionalities implemented by one or more services⁵: ontology management (*ontology browser* and *ontology editor* services), document indexing (*document indexer* service), simple search (*search engine*) and federated search (*meta-search engine* and *semantic translator*). These functionalities are detailed below.

1. Ontology management: it is considered a preparatory (off-line) stage for the other functionalities. It is to be used by domain experts to build and to manage

⁵ In this context the term *service* means a piece of software that provides a given functionality, regardless language and technologies involved on its implementation.

ontologies for representing knowledge in a given CN. This functionality can be divided in two sub-functionalities: *ontology browsing*, to allow read-only operations, like loading classes, properties and instances of ontologies; and *ontology edition*, to add/modify/remove classes, properties and instances. The reason for this division is that edition operations imply on ontology evolution, and consequently the maintenance of consistencies (of semantic annotations and ontology mappings).

2. Document Indexing: this functionality is used by information sources in order allow their documents to be searched in the future. This is provided by the *document indexer* service, which generates semantic annotations automatically, considering the ontologies used by the CN where the information provider is taking part. The annotations are further indexed.

3. Simple Search: using the *search engine* service, CN partners define semantic queries according to the ontologies adopted by the respective CN (provided by the *ontology browser* service).

4. Federated Search: users taking part in several CNs should use the *meta-search engine* service in order to perform federated searches. Users can choose the context of the search (among the CNs they are involved), define the semantic query, and then perform the search. All needed translations are made transparently by the *semantic translator* according to the ontologies of each involved CN, and the search results are translated back to the ontologies used in the original query. Moreover, the *meta-search engine* should use security mechanisms to control the search scope.

Mappings between ontologies must be defined and maintained in order to support automatic semantic translations. This aspect is not specified in this model because it can be implemented in several ways (Freitas, 2005), which has impact on how mappings should be identified and maintained, as well as how the *semantic translator* should work.

Another aspect not covered at the conceptual level is the semantic annotation approach. There are several techniques related to this issue, classified in two broader categories, *pattern-based* and *machine learning-based* (Reeve, 2005). In this sense, the model is not bound to any technique, which should be chosen during its implementation.

Besides the previous aspects, issues related to the *usability* of the semantic search are not covered by the framework. This subject is more related to the implementation level, such as graphical user interfaces.

The architecture of the conceptual framework is shown in Figure 2. Although the services are not bound to any deployment configuration, most of them should be instantiated to specific CNs. The purpose of such distribution is to allow each CN to have its own domain knowledge, documents and indexes managed by dedicated services, and thus allowing searches inside the CN scope (performed by the *search engine* service). Federated searches are provided by the *meta-search engine* and supported by the *semantic translator*. There is a single instance for each one of these services, regardless the number of CNs.

Figure 2 – Architectural view of the conceptual framework.

The services' behavior and their distribution proposed in this framework are just recommendations and can be extended or reconfigured according to specific implementation needs, limitations or by design decisions. Next section presents the implementation aspects of a prototype that is currently being developed.

4 PROTOTYPE IMPLEMENTATION

As mentioned before, the framework proposed in this work has been developed under the scope of the ECOLEAD Project as part of an ICT infrastructure (ICT-I) for supporting CNs in a (Web) Service Oriented Architecture. The implementation of such framework is called Knowledge Search Services (*K. Search*). Following the recommendations applied for the whole ICT-I implementation, these services have been developed on top of existing solutions and some value/innovation for CNs is being added. In this sense, *K. Search* services are built on top of an IR platform called KIM Platform, which provides APIs for automatic (pattern-based) semantic annotation, indexing and retrieval of documents (Popov, 2003). The architecture of *K. Search* is shown in Figure 3.

Besides *K. Search* services themselves (on top of KIM), there are some client applications, namely: information providers for publishing documents; CN applications that perform searches; and an ontology GUI editor for ontology management. There is a special web client used to perform semantic queries upon the *K. Search* services. It is being implemented as a Java portlet that provides an easy-to-use GUI for general purpose searches. This portlet can be extended or customized to be used by particular applications. Figure 4 shows two screenshots of this portlet (showing a query and its matching entities, and the resulting document).

Figure 3 – K. Search Architecture.

As the envisaged functionalities should be transparent and easy-to-use, some KIM API functionalities were merged into a single K. Search operation. For instance, the publishing of a document (provided by K. Search) involves storing, annotation and indexing (provided by KIM). The current version of K. Search is limited as KIM Platform covers only a subset of the envisaged requirements: it works only with a single ontology, provides automatic generation of annotations and indexing of them, and it is possible to make semantic queries. These functionalities are being extended in order to cover the remaining requirements. For instance, current K. Search implementation provides a mechanism for consistency maintenance (not provided by KIM), consisting of re-annotating and re-indexing all documents in background. Another improvement regarding *K. Search* services that is under development is the integration with DRACO (Sowa, 2007), a security framework also in development in the scope of ECOLEAD Project.

Figure 4 –K. Search portlet screenshots.

5 CONCLUSIONS AND NEXT STEPS

This paper focused on the problem of knowledge search in CNs and proposed a supporting conceptual framework that takes into account requirements and challenges related to CNs. The proposed framework defines some recommendations related to the expected behavior of the envisaged functionalities and can be extended or reconfigured according to implementation needs or limitations.

A first version of a prototype (*K. Search* services) has been developed and it is already available, but it requires a number of improvements in order to cover the identified requirements. This prototype, along with the whole ICT-I, will be evaluated under the scope of the ECOLEAD project in a series of test-cases close to real CNs. Next activities regarding *K. Search* development comprise:

- (a) Implementation of the *meta-search engine* service, for allowing federated searches, regarding both security and semantic aspects.
- (b) Multiple ontologies management and functionalities for identification of similarities between ontologies and mappings management. MAFRA⁶ Toolkit and the system proposed by Sánchez-Alberca (2005) are under evaluation and seem to be good candidates for implementing this feature.
- (c) A client Ontology GUI Editor (based on Protégé⁷) will be developed to allow users to manage ontologies. As seen in the implementation architecture, this tool will be the client of the ontology management services (browsing and edition).
- (d) Final refinements in the search portlet and in the consistency maintenance mechanism.

Although the conceptual framework is not bound to any implementation model, the further developments under the scope of this work may require future refinements in the conceptual model.

As a final remark, it is important to highlight that the knowledge search task is demanding and complex. Hence, there are many open issues and challenges that require deeper efforts, like natural language processing techniques for interpretation of higher-level queries (defined in natural language), and the automatic identification of similarities between ontologies.

5.1 Acknowledgments

This work has been partially supported by the Brazilian council of research and scientific development – CNPq. It has been developed in the scope of the Brazilian IFM project (www.ifm.org.br) and the European IST FP-6 IP ECOLEAD project (www.ecolead.org). Special thanks also to Mr. Carlos E. Gesser and Mr. Leandro Loss for their contributions during the design and implementation of the prototype.

6 REFERENCES

1. Berners-Lee, T.; Hendler, J.; Lassila, O.; 2001. The Semantic Web. In: Scientific American, May 2001 issue.
2. Camarinha-Matos, L. M.; Afsarmanesh, H.; Ollus, M.; 2005. ECOLEAD: A Holistic Approach to Creation and Management of Dynamic Virtual Organizations. In: Proceedings of the Sixth IFIP Working Conference on Virtual Enterprises (PRO-VE'05). pp. 3-16.

⁶ <https://sourceforge.net/projects/mafra-toolkit>

⁷ <http://protege.stanford.edu/index.html>

3. Camarinha-Matos, L. M.; Afsarmanesh, H.; 2006. Collaborative Networks: Value Creation in a Knowledge Society. In: Proceedings of PROLAMAT'06, Shanghai, China, 14-16 2006.
4. Chituc, C. M.; Azevedo, A. L.; 2005. Multi-Perspective Challenges on Collaborative Networks Business Environments. In: Proceedings of the Sixth IFIP Working Conference on Virtual Enterprises (PRO-VE'05). pp. 25-32.
5. Freitas, F.; Stuckenschmidt, H.; Noy, N. F.; 2005. Ontology Issues and Applications. In: Journal of the Brazilian Computer Society, Number 2, Volume 11, November 2005. ISSN 0104-6500.
6. Huber, G. P.; 1991. Organizational learning: The contributing processes and the literatures. *Organization Science*, pp. 88-115.
7. Kalfoglou, Y.; Schorlemmer, M.; 2003. Ontology Mapping: The State of the Art. In: *The Knowledge Engineering Review Journal*, 2003.
8. Kiryakov, A.; Popov, B.; Ognyanoff, D.; Manov, D.; Kirilov, A.; Goranov, M.; 2003. Semantic Annotation, Indexing, and Retrieval. In: 2nd International Semantic Web Conference (ISWC2003).
9. Köhler, J.; Philippi, S.; Specht, M.; Rüegg, A.; 2006. Ontology based text indexing and querying for the semantic web. In: *Knowledge-Based Systems*, Volume 19, Issue 8, December 2006, pp. 744-754
10. OntologyMatching; 2006. <http://ontologymatching.org/>
11. Popov, B.; Kiryakov, A.; Kirilov, A.; Manov, D.; Ognyanoff, D.; Goranov, M.; 2003. KIM – Semantic Annotation Platform. In: 2nd International Semantic Web Conference (ISWC2003).
12. Rabelo, R. J.; Gusmeroli, S.; Arana, C.; Negellen, T.; 2006. The ECOLEAD ICT Infrastructure for Collaborative Networked Organizations. In: Proceedings of the Seventh IFIP Working Conference on Virtual Enterprises (PRO-VE'06). pp. 451-460.
13. Reeve, L.; Han, H.; 2005. Survey of semantic annotation platforms. In: SAC '05: Proceedings of the 2005 ACM symposium on Applied computing. pp. 1634-1638. ACM Press, New York, NY, USA.
14. Sánchez-Alberca, A.; García-García, R.; Sorzano, C. O. S.; Gutiérrez-Cossio, C.; Chagoyen, M.; Fernández-López, M.; 2005. AMON: A Software System for Automatic Generation of Ontology Mappings. In: Workshop on Building and Applying Ontologies for the Semantic Web (BAOSW'05).
15. Si, L.; Callan, J.; 2005. Modeling Search Engine Effectiveness for Federated Search. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). pp. 83-90. ACM Press, New York, NY, USA.
16. Sowa, G.; Śnieżyński, T.; 2007. Technical Report (Deliverable) D61.4b - Security framework and architecture.
17. Stojanovic, N.; 2005. On the query refinement in the ontology-based searching for information. In: *Information Systems*, Volume 30, Issue 7, November 2005, pp. 543-563.
18. Uren, V.; Cimiano, P.; Iria, J.; Handschuh, S.; Vargas-Vera, M.; Motta, E.; Ciravegna, F.; 2006. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. In: *Journal of Web Semantics*. 4(1): 14-28 (2006).