ONTOLOGIES BASED APPROACH FOR SEMANTIC INDEXING IN DISTRIBUTED ENVIRONMENTS

Claude Moulin
CNRS Heudiasyc, University of Compiègne, France

Cristian Lai
CRS4, Center for Advanced Studies in Sardinia, Italy

Keywords: Semantics and Ontology Engineering, Information Search and Retrieval, Distributed Data Structures, Distributed Systems.

Abstract: In this paper we present some issues relating the semantic indexing of resources in distributed decentralized systems. We discuss some matter regarding the navigation between indexed resources and the way to enhance our model for answering more generic queries. The argument is introduced with a brief scenario focusing on e-learning domain, even if our goal is more general purpose. We discuss the role of ontologies in semantic indexing, moving from centralized systems to the distributed paradigm. We explain how to distribute a common index of resources semantically identified, composing a distributed knowledge base. The paper highlights the importance of an ontology system for the key generation and the use of specific domain ontologies.

1 INTRODUCTION

In the last years, research in the field of Knowledge Engineering has addressed technologies able to create modular and more efficient knowledge-based systems. An important requirement is the improvement of response time and the quality of the discovery facilitating the selection of results more related to a specific request. Semantic indexing seems to be a valid candidate to fulfill this requirement. Main search algorithms based on sequences of keywords also try to get better results but they often lack of contextual information. Instead, selective search takes into account the meaning, the importance of the words within the resources content, mutual vicinity, etc. Moving from client-server paradigm to Peer to Peer (P2P) networks model, information retrieval problem becomes more articulated. Even if P2P networks architectures can not offer a hyper textual navigation between resources and lost the hierarchical organization of data, they keep some advantages such as the easy scalability of resources, the reliability on data transferring, and low operating costs. A P2P architecture will avoid both physical and semantic bottlenecks that limit information and knowledge exchange (Staab, Stuckenschmidt, 2006).

In this paper we face the problem of the semantic indexing of resources in decentralized networks, in which traditional discovery is based on documents title. Our studies address techniques able to create an index of managed resources, distributed and semantically annotated. Semantic structures are introduced through ontologies. Distribution is gained by P2P paradigm. The distribution of the index is performed by a Distributed Hash Table (DHT), and thus the generation of keys remains an important aspect. The semantic information strictly related to a resource is included in a semantic annotation and then inserted in the index.

In the case of textual documents annotation, two approaches can be followed: metadata are either inserted within the document or found outside the content as external resources. In our work we prefer extracting the semantics of annotation from ontologies. The solution we propose allows to fasten to a document the meta-information that describes its content and to publish it in the network. The URL of the resource tied up to the meta-information

Molin C. and Lai C. (2009).
ONTOLOGIES BASED APPROACH FOR SEMANTIC INDEXING IN DISTRIBUTED ENVIRONMENTS.
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 420-423
DOI: 10.5220/0002298604200423

allows its further access. Our solution lets open the choice of the ontologies required for the descriptions. The user has just to select ontologies really shared among communities (Gruber, 1993) else the discovery of the documents published in the network would be impossible. The only characteristic we use is the Unified Resource Identifiers (URI) of concepts, relations and instances of concepts. We also chose ontologies because we want that any software agent can reason on the structure of knowledge in order to build the most appropriate queries when searching for resources. Indeed, we consider that the resources published in the network may be used diversely. The resources we want to index are not always textual, so a manual indexing is necessary; some interesting information is not contained within the resource, thus the automatic indexing would not have effect on this. It's the case when somebody wants to express a point of view on a resource.

The paper is organized as follows: Section 2 describe a scenario where our indexing system can be used. Section 3 reports briefly the related works. Section 4 presents the core of our indexing system and the way keys are generated.

2 SCENARIO

We consider teachers writing didactic material for their courses and wanting to share them with other teachers or students thanks to a simple mechanism. They don't want to deal with heavy tools or to depend on centralized repositories. Teachers only have to semantically describe their material in relation with one or more ontologies. Obviously, tools are supplied for that and the use of ontologies is transparent to users. Then the documents are published in a P2P network in conjunction with their semantic description. Publication of semantically annotated document in P2P networks was also presented as a real challenge in (Davies et al., 2003). Students and other teachers can discover the resources making queries based on the concepts expressed from ontologies. In this scenario the choice of the ontologies is crucial and ontologies really shared among communities should actually be used.

Resources of various types can be handled and are not necessarily textual. Semantic annotations can represent objective information about resources (nature, concepts of scientific domain, etc.) but can also represent a point of view on documents (difficulty level, usefulness in some context, etc.).

This example of semantic indexing in P2P networks is planned to be integrated in a large project concerning Organizational Memories. In this kind of memories, resources can also be semantically indexed on ontologies and can be annotated. An ontology used for a memory is also a part of this memory and is used for navigation. Centralized memories may have some disadvantages when they need to be filled up. Such a distributed system could be a solution to this issue.

3 RELATED WORKS

Using ontologies in distributed systems like P2P networks is closely considered by scientific community. In 2002 Edutella project (Nejdl et al, 2001), handled by Sun Microsystems, did a first approach for the association of semantics to educational content through an open source infrastructure based on RDF metadata for the interoperation between different schemes (IEEE/LOM, IMS Learning Design (http://www.imsglobal.org/learningdesign/index.html), ADL SCORM), performing a mapping among them. The SWAP project (Ehrig et al, 2003), managed by the University of Karlsruhe, pays a special attention to topics related to Semantic Web. Its aim is to allow computers to actually comprehend the meaning of its processed data. Using the model of ontologies, the project allows to develop a technology in the area of knowledge management and P2P. Complex structures can be easily encoded in a set of RDF triples. In (Della Valle et al, 2006) is supposed that RDF should become the bases of the Semantic Web. Nevertheless, RDF isn't enough; it does not supply a sufficient expressive ability to represent the whole knowledge schema. DHT based overlays systems offer an interesting alternative to existing information system architectures. We propose to express the semantic classification through concepts articulated by ontologies that describe the specific domain and to formulate such expression by the OWL formalism.

4 SEMANTIC INDEXING

We generally consider two kinds of models for indexing resources: boolean (see Salton et al., 1982) and vectorial. In the boolean model, the index of documents is an inverse file which associates to each keyword of the indexing system the set of

documents that contain it. A user's query is a logical expression combining keywords and Boolean operators (NOT, AND OR). The system answers with the list of documents that satisfy the logical expression. The keywords proposed by the user are supposed contained in the index.

In the vectorial model a document is represented by a vector whose dimensions are the keywords of a vector and the coordinates correspond to the weight of the document in each dimension. A request is also a vector of the same nature. The system answers a request with the list of documents which present a similarity with the request thanks to a specific measure based on the vectors coordinates.

In centralized indexing both models are available. However, in our case the index must be distributed and the numbers of queries sent to the index when searching for resources should minimized, because they are time consuming. The model of a distributed index is necessarily Boolean. Our model should be able to satisfy logical expressions and we have to construct different keys for keeping this feature.

Our solution also allows a file sharing but the semantic information is not contained in the title of the resource. It's contained in keys that describe the resources. Ontology element identifiers are the essential part of the indexing keys.

4.1 Ontological Elements

Our solution lets open the choice of the ontologies requested for the descriptions. However, users of the community should share them else the discovery of the documents published in the network would be impossible. The resource provider is responsible of the choice of the ontology describing the concepts. In our example at least two ontologies are pointed out, one for the theoretical domain of the resource (theory of language) and another for the description of the resources. It's generally the case when indexing resources for e-learning. In case of a manual semantic indexing, first is necessary to select the ontologies used for building the indexing key. A key may contain several concepts belonging to different ontologies. For having homogeneity in our knowledge representation and for preserving agent reasoning based on ontologies, we use the LOM ontology developed at "Université de Technologie de Compiègne" for representing learning objects. Studying the previous example and analyzing the selected ontologies, we are led to the following conclusion: a user can select a concept (the concept of grammar is described in the first ontology) or an instance of a concept (Exercise is an instance of the concept of Learning Resource Type, and Difficult is an instance of the concept of Difficulty Level).

Within the ontology, an element is completely defined by its unique URI. It is enough to insert the URIs of ontological elements for characterizing a document in the key that indexes it. From this information any software agent can discover the type of the element inside a key and can decide to build new queries if some do not give satisfying results.

4.2 Distributed Knowledge Base

We aim at creating a community, i.e. a set of nodes of a P2P network, and at distributing a common index of resources semantically identified. The data structure that has been considered suitable for that is a Distributed Hash Table (DHT) (Stoica et al., 2006). The index is composed of entries that are pairs of data (key, value). Each node of the network contains a part of the whole data structure. The publication, or indexing, is the operation of insertion of a resource inside the DHT. More exactly it is the operation of inserting a new index entry in the DHT. The discovery is the operation which allows to find some resources in the network that correspond to a research key.

4.3 System Ontology

We consider that a key used for indexing a document is based on a RDF semantic description, i.e. a graph pattern of triples representing subjects linked to objects by predicates. What is the meaning of "indexing a document on the concept of automaton" (in theory of language)? That corresponds to the fact that the content of the document has something to see with the concept. In RDF, the document is represented as a blank node of type Document and treating of the concept of "automaton". We had to create a system ontology for representing these data. The RDF representation of the document is then in N3 notation:

```
[] rdf:type syst:Document ;
   syst:hasInterest lt:Automaton.
```

The system ontology is represented by the *syst* prefix and the domain ontology by *lt*. "hasInterest" is an annotation property allowing to attach any type of elements (concept, relation or individual). The system ontology also contains the concept of Ontology, subconcept of Document for representing the ontologies used by the users for indexing. Such

an ontology is published in the network under the description:

```
[] rdf:type syst:Ontology
```

The discovery system first looks for all the ontologies used in the network for indexing and can load them in the navigation sub system. For indexing a document that occurs to be an ontology people can use the following description:

```
[] rdf:type syst:Document ;
   syst:hasType syst:Ontology.
```

4.4 Key Generation

An interesting feature of the navigating through an ontology is the way to suggest for accessing other resources. We can say that resources are close if they are indexed by close ontological elements. Concepts are close if one is a specialization of the other or if they are domain and range of the same property. From a selected concept, it's simple to find close concepts and then to build the query that allows to access close resources. Instances of concepts are close if they have the same type or if their types are close.

Due to the Boolean model of the distributed index, it' necessary to create different keys when publishing a document, so to be discovered with different requests. The indexing must allow the reasoning by subsumption. A request on a super-concept must allow the discovery of documents indexed on a sub-concept. The difficulty is to stop the reasoning about the transitivity of subsumption, for not indexing on a too general concept. We consider that only two levels are enough in this case.

It is also possible to index on *an attribute*. A document may show the interest for a country to have a population and if this notion is modeled by using an attribute, the description could be:

```
[] rdf:type syst:Document ;
   syst:hasInterest ex:hasPopulation.
```

5 CONCLUSIONS

In this paper we have presented a solution which aims at distributing semantically indexed resources on P2P networks. The distribution of the index is performed by a Distributed Hash Table. The semantic information strictly related to a resource and representing a point of view on the resource is inserted in the key used to index the resource. The semantic information comes from ontologies. Any ontology can be used by our system. The drawback of our solution is that the user has to navigate the suitable ontologies and this operation can be time consuming. Domain specific expert users have to look for interesting ontologies and to publish them in the network. Currently we are enhancing the tools used to manage the ontologies in a way to hide their underlined structures and to present them in a comprehensive way. Our solution can supports the building of online communities of users that want share easily digital resources. We consider that the building, storage and maintenance of ontologies are the duty of the community the user belongs to. The semantic indexing is strictly related to these issues.

REFERENCES

Davies, J., Fensel, D., and Van Harmelen, F., 2003. *Towards the Semantic Web: Ontology-Driven Knowledge Management*. New York : John Wiley & Sons.

Della Valle, E., Turati, A. and Ghigni, A., 2006. PAGE: A *Distributed Infrastructure for Fostering RDF-Based Interoperability.* Distributed Applications and Interoperable Systems. Berlin : Springer.

Gruber, T. R., 1993. *Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer.* Kluwer Academic Publishers.

Ehrig, M., Tempich, C., Broekstra, J., Van Harmelen, F., Sabou, M., Siebes, R., Staab, S. and H. Stuckenschmidt, 2003. *Swap - ontology-based knowledge management with peer-to-peer technology.* WIAMIS'03, London, pp. 557-562. World Scientific, London.

Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmer, M. and Risch, T.,2001. *Edutella: A p2p networking infrastructure based on Rdf.*

Salton, G., Fox, Edward A. et Wu, Harry, 1982. *Extended Boolean information retrieval*. Technical Report, Cornell University.

Staab, S., Stuckenschmidt, H., 2006. *Semantic Web and Peer-to-Pee: Decentralized Management and Exchange of Knowledge and Information*. Springer.

Stoica, I., Morris, R., Karger, D., Kaashoek, M. F. and Balakrishnan, H., 2006. *Chord: A scalable peer-to-peer lookup service for internet applications*. In Proceedings of the ACM SIGCOMM '01 Conference, San Diego, California, pp 149-160.