Onto-DOM: Organizational Knowledge Sources Integration through an Ontology-Based Approach

Mariel Alejandra Ale[1], Cristian Gerarduzzi[1], Omar Chiotti[2], Maria Rosa Galli[2]

[1] CIDISI – UTN – FRSF, Lavaise 610, Santa Fe, Argentina
male@frsf.utn.edu.ar
[2] INGAR – CONICET, Avellaneda 3657, Santa Fe, Argentina
{chiotti, mrgalli}@santafe-conicet.gov.ar

Abstract. Nowadays, there is a large number of Knowledge Management (KM) initiatives implemented in organizations, which often fail to manage the natural heterogeneity of organizational knowledge sources. To address heterogeneity, documentation overload and lack of context we propose Onto-DOM, a question-answering ontology-based strategy implemented within a Distributed Organizational Memory. Onto-DOM is a portable question-answering system that accepts natural language queries and, using a domain ontology, transforms and contextualizes the query eliminating the inherent natural language ambiguity. At the same time, it recovers those knowledge objects that are most likely to contain the answer.

Keywords: Knowledge Management, Distributed Organizational Memory, Ontologies.

1 Introduction

There is already a large number of Knowledge Management (KM) initiatives implemented in organizations, which often fail to manage the natural heterogeneity of organizational knowledge sources. Instead, many approaches to KM have been only based on new information system technologies to capture all the possible knowledge of an organization into databases that would make it easily accessible to all employees [8]. The philosophy of regarding knowledge as a "thing" that can be managed like other physical assets has not been quite successful for several reasons related to tacit knowledge capture and tacit-to-explicit knowledge conversion. Therefore, we believe that an approach with a new conceptual basis is needed that emphasizes the semantics of organizational knowledge objects.

Our contributions are the following:

- We present a three-layer architecture for a Distributed Organizational Memory (Onto-DOM) that addresses two common problems in implementations with these characteristics: the documentation overload that implies for workers the knowledge elicitation to feed an Organizational Memory and the lack of context associated to tacit-to-explicit knowledge conversion.

XXXIV Conferencia Latinoamericana de Informática (CLEI 2008), pp. 569-578, 2008.

- We propose a question-answering ontology-based strategy implemented within the Organizational Memory that allows an automatic semantic treatment of heterogeneous organizational knowledge sources.
- By means of a real world scenario using documents related to the tourism area we describe the most important tasks of the strategy: document annotation in the Knowledge Representation Layer and query semantic treatment in the Information Retrieval and Processing Layer.

Onto-DOM employs domain ontologies in a number of key processes. The domain ontology is used as the core of the representation strategy of knowledge objects. This strategy selects ontological concepts as descriptors obtaining a homogeneous representation of objects structurally heterogeneous. The domain ontology is also used in query refinement, in the reasoning process (a process of generalization/specialization using ontology classes and subclasses) and in the similarity resolution. Experimental evaluation for the Tourism domain indicates that our strategy can automatically annotate documents with high precision and recall while is useful to eliminate natural language query ambiguity. We say that Onto-DOM is portable because the time needed to implement it in a new domain is minimum, requiring just a change of the associate domain ontology.

In section 2, we present our Onto-DOM architecture. In section 3, we discuss the knowledge representation strategy for semantic document treatment within Onto-DOM. In section 4, we present the question treatment strategy. Finally, remarks and future works are presented in section 5.

2 Distributed Organizational Memories and Domain Ontologies for KM

In their daily activity, organizations generate huge amounts of textual information along with less traditional non-textual information (audio, video and images). Making all this knowledge available requires a mechanism that retrieves a minimum of irrelevant information (high precision) while assuring that no relevant information is missed (high recall). A traditional solution was a keyword-based search where only those documents containing the keywords were retrieved. Nevertheless, documents often convey the required information without containing the exact keywords. This problem is normally addressed by expanding the query terms using co-occurrence techniques. As a consequence, recall is increased, but at the same time, precision is lost. A different approach to this problem is to classify documents using a semantic-based technique rather than doing it with a word-based or statistical technique.

Some organizational KM systems proposals focus on the application of information technologies for the capture, storage, and retrieval of organizational knowledge. In our approach we propose Organizational Memories (OMs) to support knowledge effective representation, use, handling and conservation over time and space, whenever possible - without human intervention [1].

Croasdell et al. define OMs as the means by which knowledge from the past is brought to bear on present activities resulting in higher organizational effectiveness [5]. Knowledge is naturally distributed across the organization and it is necessary to

represent and retrieve knowledge objects in the same way. To this aim we propose to divide the organization in several knowledge domains and associate every domain with its own OM. Every OM has an interface that enables knowledge retrieval from other domain OM if necessary. Allowing connection between individual OMs creates a Knowledge Network that fosters knowledge sharing and reuse within the organization [2][3].

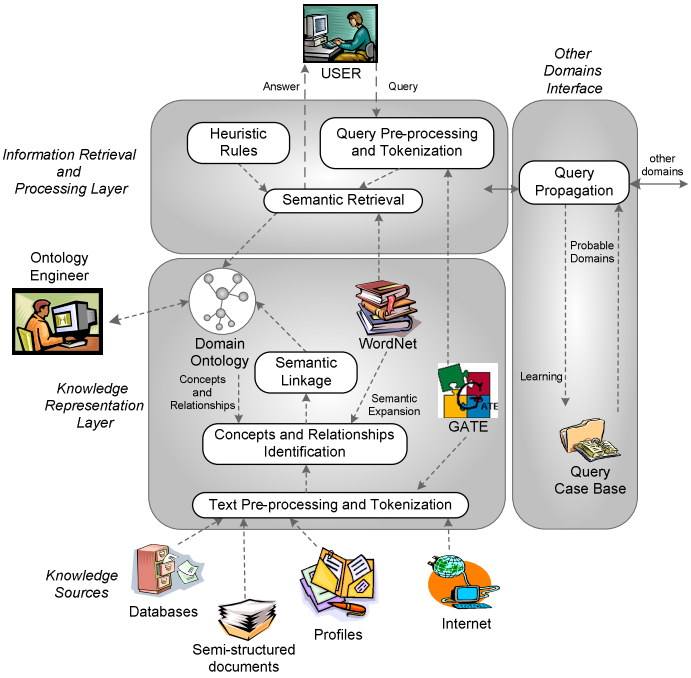

Fig. 1. Onto-DOM architecture

In this particular type of OM, the characteristics, attributes, and semantics of the knowledge objects, as well as the relationships among them are represented through a domain ontology. Ontologies aim to capture domain knowledge in a generic way and provide a commonly agreed understanding of a domain, which may be reused, shared, and operationalized across applications and groups [6].

An additional benefit of ontology modeling is context representation. Ontologies provide a domain model that allows knowledge objects to be seen in their context and this can be crucial for subsequent reinterpretation or use in a new task or project. As shown in Figure 1, our Onto-DOM architecture has three main components:

- Information Retrieval and Processing Layer: it is responsible for user query analysis, query transformation into a matching format and information retrieval.
- Knowledge Representation Layer: this component is responsible for the knowledge extraction and representation from heterogeneous sources.
- Other Domains Interface: It is responsible for propagating the user query to OMs in different domains that can provide an answer. In order to accomplish this task

the module implements a learning mechanism based on user´s feedback to propose possible target domains.

Another important advantage provided by ontologies can be seen in the Information Retrieval area, where the availability of an ontology allows for replacing the traditional keyword-based retrieval approaches by more sophisticated ontology-based retrieval mechanisms [7][9]. In fact, ontologies are often presented as silver bullets for the Semantic Web and are expected to bring several benefits to Information Retrieval related to recall and precision, user assistance in query formulation, and retrieval from heterogeneous knowledge sources.

In the next sections we will describe the implementation of Onto-DOM´s most important layers: knowledge representation and information retrieval.

3 Knowledge Representation Layer

As we said before, our goal is to represent in a homogenous way knowledge sources that are heterogeneous in nature (more specifically we began our experiments with natural language documents).

We propose a strategy for semantic document representation where ontologies are used as the main structure for the classification process. Our proposal relies on the hypothesis that domain ontologies contain all the relevant concepts and relationships in a given domain even though the way in which ontologies are built up in the domain is out of the scope of this paper. To illustrate our strategy, we present an example using an extended version of the Travel[1] ontology that contains more than 120 concepts from the tourism area and an extract of a web page[2] of the same domain.

3.1 Tokenization and Lexical-Morphological Analysis for Concepts Identification

This task is divided into two main phases: the tokenization of the text and, the lexical-morphological analysis of each token. Tokenization consists of dividing the text into single lexical tokens and involves activities such as sentence boundary detection, simple white space identification, proper name recognition, among others. After tokenization, a lexical-morphological analysis has to be done using a POS (Part-of-Speech) tool. In our case, we use the POS tagger provided by GATE[3] (General Architecture for Text Engineering) which specifies if a term is a verb, an adjective, an adverb, or a noun.

Usually, the decision on whether a particular word will be used as a representative term is related to the syntactic nature of the word. In fact, nouns frequently carry more semantics than adjectives, adverbs, and verbs [4]. As, in our case, representative terms

[1] available at http://protege.stanford.edu/plugins/owl/owl-library/index.html (for the extended version send a request to male@frsf.utn.edu.ar)

[2] available at http://www.vacationidea.com

[3] available at http://gate.ac.uk

will be determined by ontological concepts, which are nouns, we will focus on this syntactic category within the tagged text.

In this sense, ontological concepts can be seen as possible classifying categories. At this stage, if the noun is not directly found in the ontology, using the synonyms set and hyperonymic/hyponymic structure provided by WordNet[4], we semantically expand every noun identified in the text and perform a new search in the domain ontology. By doing this, we do not only identify exact ontological concepts occurrences but also derivations of the same word or even a synonym. Up to this point, we are not interested in the meaning of each possible concept and that is why the presence of more than one sense for each noun in WordNet is not a problem.

For example, the concept "food" has been found with WordNet assistance. In this particular case, by using WorldNet's hypernym relationship we found out that "meal" (a concept present in the text) is a kind of "food", which is a concept in the ontology. In other cases, this tool helps us to mark as ontological concept occurrences the presence of synonyms, and in this way, if the noun is not found directly in the ontology, WordNet allows us to expand the matching possibilities taking advantage of related concepts (synonyms, hypernyms, etc.).

3.2 Semantic Document Representation

At this point, we navigate through the domain ontology using the properties structure in order to find relationships among previously identified concepts. By doing this, we expand the possible document descriptors using intermediate ontology levels and contextualizing those concepts that, in another way, could not be related to other concept among those that were identified in the previous step.

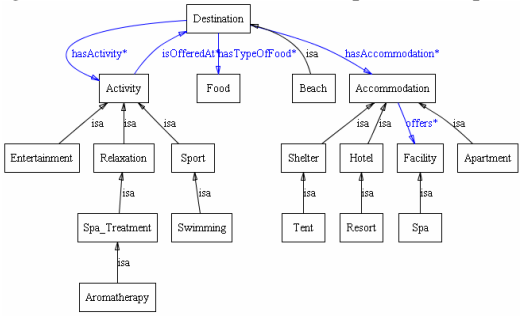

Fig. 2. Ontology representation

We take advantage of ontological relationships and knowledge contained in the domain ontology in order to perform a more accurate and contextualize representation of the document. As a result, we finally obtain the subset of the domain ontology that best models the document semantic content (Figure 2). Figure 3, shows the knowledge representation prototype from where ontology engineers can choose the descriptors for each document along with the methodology used to obtain each

[4] available at http://wordnet.princeton.edu

descriptor (straight finding, synonyms, hypernyms, etc.). This semantic document classification will enable new, semantically enhanced, access methods.

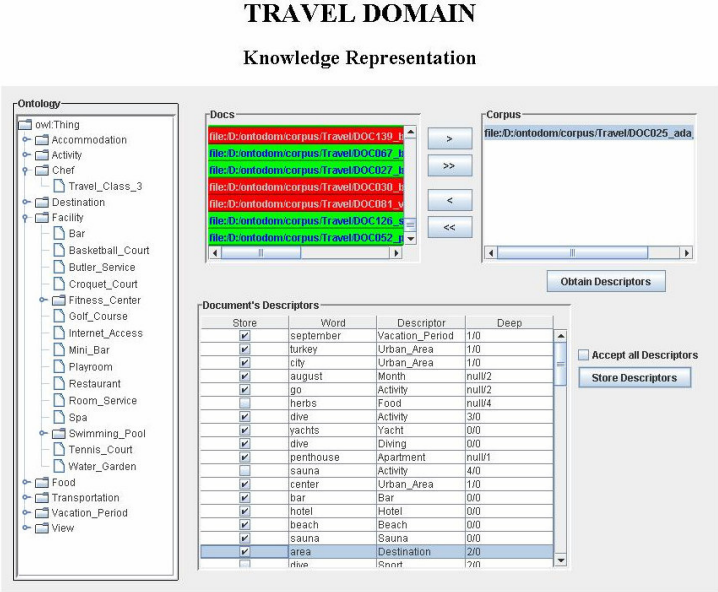

Fig. 3. Knowledge Representation Prototype Interface

3.3 Representation Evaluation

As a first step in the implementation process, we estimate the representation strategy performance applying the following metrics according to Yang´s [10] definitions: recall, precision, fallout, and accuracy[5].

Recall	Precision	Fallout	Accuracy
87%	70%	14%	86%

Recall is a measure of strategy performance in finding relevant concepts. Recall is 100% when every relevant concept is annotated. In theory, it is easy to achieve good recall simply annotating every noun in the text. Therefore, recall for itself is not a good measure of strategy quality. Precision, on the other hand, is a measure of strategy performance in not annotating non relevant nouns. Finally, fallout is the measure of how fast precision is reduced as recall is increased, in other words, it represents the portion of non-relevant concepts that were annotated. We analyzed the reason for the relative low value of recall measure and found that 82% of the not annotated relevant concepts correspond to names of vacation destinations that were either places not recognized by WordNet (i.e. Caicos) or types of destinations that were not taken into account in the domain ontology (i.e. islands, archipelago). We

[5] performed over 150 documents with 35.091 words

believe that recall can be improved by using common vocabulary domain lists and enriching the domain ontology.

4 Information Retrieval and Processing Layer

Most works on ontology-based question-answering tends to focus on simple query expansion or on exploiting the availability of a knowledge base linked to the ontology to provide a precise answer. In the first case, we believe that this is a limited use of ontology potential and, in the second case, a vast knowledge base must be learnt in order to provide adequate answers. The effort required to feed all organizational knowledge in a knowledge base is prohibitive. Moreover, if precise answers are required this process cannot be fully automated.

Ontologies ensure an efficient retrieval of knowledge resources by enabling inferences based on domain knowledge. This vision relies on the assumption that an ontology designed to describe a domain can both annotate and retrieve knowledge sources. In fact, this is not always the case because domain specialists usually build the ontologies and users do not always share or understand their viewpoints. Users might not use the right concepts – from an ontologist´s viewpoint – when writing a query, leading to missed answers. For example, a user might use "student lodging" instead of "hostel". Or, perhaps a user asking for a "hotel" might also appreciate the retrieval of documents about "resorts". Consequently we partially use the same strategy applied to document descriptors determination in the semantic query treatment.

In this case, Onto-DOM accepts natural language queries and, using the domain ontology, transforms the query by eliminating natural language ambiguity and recovering those knowledge objects that are most likely to contain the answer. In a sense, this layer tries to find similarity between the query and the ontological concepts.

Our strategy to determine similarity includes both conceptual and relationship similarity. The first step is to transform the query in a format that facilitates ulterior evaluations and, to this aim, we apply part of the same strategy for document representation. After this stage, we have not only nouns that match ontological concepts but we also keep the verbs in order to evaluate relationship similarity and wh-words that give us an idea of the type of answer expected (time, location, person, etc.).

We go beyond taxonomic relationships (is-a) making use of semantic relationships to sharpen query comprehension. Essentially, we are trying to "understand" the question lying on the codified knowledge in the domain ontology, lexical resources as WordNet and GATE and the heuristics associated to the treatments of wh-words. For example: in the query "Where can I eat Vegetarian dishes?, after the first analysis we obtain the following useful information:

eat(Vegetarian, Food) (where, location)

In this case, the concept Food is derived from Dishes with the help of WordNet`s hyperonymy structure. Nevertheless, as we said before, our main objective is to go beyond a keyword search or the use of the domain ontology as a query expansion tool. To this aim, on the one hand, we will use the verbs detected in the query to look for semantic similarity related to relationships, and on the other hand, we will analyze the concepts related to those relationships to see if they belong to the expected type according to the wh-word.

Following the previous example we recover the ontological concepts identified in the query along with their neighbors, Restaurant and Chef (Figure 4). To decide if one of these neighbors is useful to represent the query (and not search only by Food and Vegetarian) we evaluate similarity between the verb in the query (eat) and the verbs in the relationships attached to the identified concepts (serve, specialize) using the synonym and correlate sets of WordNet.

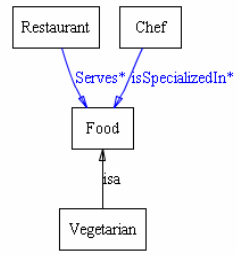

Fig. 4. Ontological concepts identified in the query (with their neighbors)

This analysis shows that "serve" has a higher semantic similarity with "eat" than "specialize".

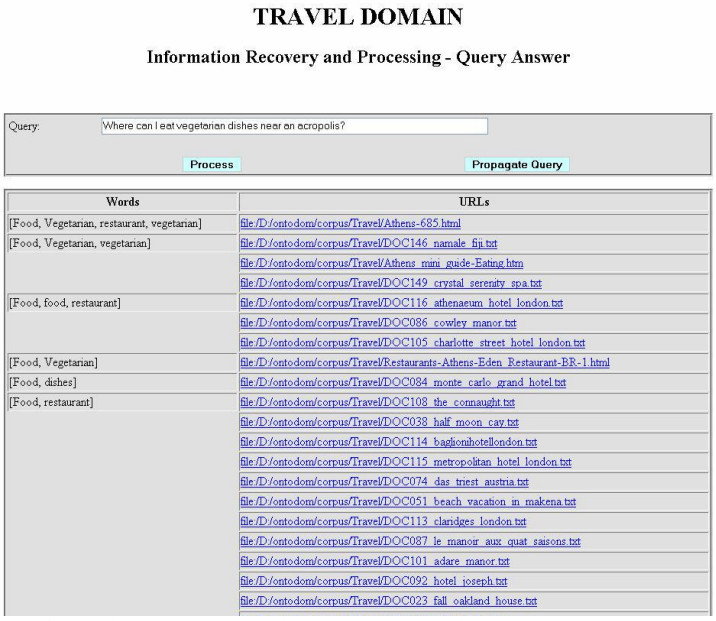

Fig. 5. Information Recovery and Processing Interface

To confirm this result, or as an alternative in case we are not able to obtain a conclusive result in the verbs comparison, we analyze the concepts at each end of the relationships (Restaurant, Chef) to see if they match with the expected type according to the wh-word. In this particular case, WordNet tells us that Restaurant is a Location (expected type according to the wh-word Where in the query) and Chef is a Person confirming that the portion of the domain ontology that best represents the query contains the concepts: Food, Vegetarian and Restaurant.

As regards to query evaluation the same results as those for document annotation are expected since the strategy being used is almost the same, adding in this particular case verbs treatment and the use of the ontological relationships (Figure 5). In this sense, our analyses have demonstrated that the queries, due to their short length, are much more sensible to the errors of the strategy. In these cases, a concept detection error attributable to the POS-tagging tool or the annotation strategy has a much greater impact than the same error in a document. To address this problem we are working in a domain independent heuristics set to improve query treatment.

5 Final Remarks and Future Work

Our goal is the implementation of a Distributed Organizational Memory system to support KM activities, representing in a homogenous way knowledge sources that are heterogeneous in nature (more specifically documents). To this aim, we propose a strategy for semantic document representation where ontologies are used as the main structure for the classification process. Our proposal relies on the hypothesis that domain ontologies contain all the relevant concepts and relationships in the domain. Onto-DOM combines, in a novel way, a series of techniques to "understand" the natural language query and map it to the semantic annotation done in the organizational knowledge sources.

Our initial experiments have yielded reasonable results. These show that it is possible to automatically perform operations such us document integration to an Organizational Memory by semantic annotation and a richer information retrieval. During our experiments with the annotation strategy we have identified several factors that may contribute to uncertainty. One of the reasons for errors in ontology concept identification has to do with text preprocessing. This preprocessing includes a fully automatic noun markup that has an error rate that influence the effectiveness of the subsequent steps. These results could be improved using more sophisticated Natural Language Processing techniques. The domain ontology definition, which is currently restricted to a relatively small number of concepts, also contributes to a low recall rate.

Other identified problems are related to words association. For example, we found some documents that describe what things a place does not have (bars, theaters, cars, etc.) but the strategy classified these documents as describing places with those characteristics. We are currently seeking more advanced techniques to improve the analysis of negative expressions. In relation to query treatment we are currently developing a heuristics set to lower the impact of error rate in short sentences.

Finally, we have assumed the availability of a predefined domain ontology. This means that all documents will be treated according to that particular view of the world. However, in any realistic application scenario, new documents that have to be classified will generate the need for new concepts and relationships. The meanings of terms evolve or take on new meanings as organizational knowledge evolves. It is clear that we will have to find solutions to problems regarding the addition, change or elimination of ontological concepts. Further research would be directed towards the use of the annotation strategy to suggest ontology improvements.

Despite the issues to be solved as future work, our semantic representation and query treatment strategy has proved to be a useful approach to address two major problems in KM initiatives: documentation overload and lack of context. Our strategy is automatic and does not have a learning phase that has to be redone every time we move to a different domain; only a change of the domain ontology is needed. We believe that these characteristics make this strategy suitable for a DOM implementation where a large and variable number of domains are presented and where Knowledge Intensive Tasks´ knowledge needs are continuously changing. Finally, the query treatment strategy allows us to make further use of ontology advantages beyond query expansion.

References

1. Ale M., Chiotti O. and Galli M.: A Distributed Knowledge Management Conceptual Model for Knowledge Organizations, ICFAI Journal of Knowledge Management, ICFAI University Press, pp. 27-39 (2005)
2. Ale M., Gerarduzzi C., Chiotti O. and Galli M.: Onto-DOM: A Question-Answering Ontology-based Strategy for Heterogeneous Knowledge Sources, ICFAI Journal of Knowledge Management, ICFAI University Press, pp. 54-68 (2007)
3. Ale M., Chiotti O. and Galli M.: Enterprise Knowledge Management for Emergent Organizations: An Ontology-Driven Approach, in Knowledge Management Strategies: A Handbook of Applied Technologies, IDEA Group Publishing, pp. 218-249 (2008)
4. Baeza-Yates R. and Ribeiro-Neto B.: Modern Information Retrieval, Addison-Wesley, Wokingham, UK (1999)
5. Croasdell D., Jennex M., Yu Z. and Christianson T.: A Meta-Analysis of Methodologies for Research in Knowledge Management, Organizational Learning and Organizational Memory: Five Years at HICSS, 36th Annual Hawaii International Conference on System Sciences, pp. 110 (2003)
6. Fensel D.: Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Springer Verlag, Berlín (2001)
7. Guarino N., Masolo C. and Vetere G.: Ontoseek: Content-based access to the web, IEEE Intelligent Systems, Vol. 14, Issue 3, pp. 70-80 (1999)
8. Levine L.: Integrating Knowledge and Processes in a Learning Organization, Information System Management, pp. 21-32 (2001)
9. Richard-Benjamins V., Fensel D., Decker S. and Gómez-Pérez A.: (KA)2: building ontologies for the internet: a mid-term report, International Journal of Human-Computer Studies, Vol. 51, Issue 3, pp. 687-712 (1999)
10. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval (1999)