Ontology integration in a multilingual e-retail system

Maria Teresa PAZIENZA(i), Armando STELLATO(i), Michele VINDIGNI(i),
Alexandros VALARAKOS(ii), Vangelis KARKALETSIS(ii)

(i)   Department of Computer Science, Systems and Management,
University of Roma Tor Vergata, Italy
{pazienza, stellato, vindigni}@info.uniroma2.it

(ii)   Software and Knowledge Engineering Lab., Institute of Informatics
and Telecommunications, N.C.S.R. "DEMOKRITOS". Athens, Greece
{alexv, vangelis}@iit.demokritos.gr

Abstract

The advent of e-commerce and the continuous growth of the WWW led to a new generation of e-retail stores. A number of commercial agent-based systems have been developed to help Internet shoppers decide what to buy and where to buy it from. In such systems, ontologies play a crucial role in supporting the exchange of business data, as they provide a formal vocabulary for the information and unify different views of a domain in a safe cognitive approach. Based on this assumption, inside CROSSMARC (a European research project supporting development of an agent-based multilingual information extraction system from web pages), an ontology architecture has been developed in order to organize the information provided by different resources in several languages. CROSSMARC ontology aims to support all the different activities carried on by the system's agents. The ontological architecture is based on three different layers: (1) a meta-layer that represents the common semantics that will be used by the different system's components in their reasoning activities, (2) a conceptual layer where the relevant concepts in each domain are represented and (3) a linguistic layer where language dependent realizations of such concepts are organized. This approach has been defined to enable rapid adaptation into different domains and languages.

1   Introduction

The continuous growth of the information on the Web and the proliferation of e-commerce sites are becoming overwhelming for consumers, who should acquaint themselves with a huge number of sites, dissimilar more in the amount of provided information, presentation styles, and overall organization, than in their contents. Extracting semi structured data from e-retail sites (and in general from the Web) is a complex task. Images, texts and other media that contain the relevant information, are organized in a way suitable to catch the human attention more than to be perceived as a rigorous and intelligible structure. The extraction task becomes even harder in our multi-lingual society, as web pages are typically written in different languages. Moreover, new product types are likely to appear in the market as technology is being evolved. This makes the customisation of existing systems/resources to new unforeseen scenarios an expensive and labour-intensive effort. In such systems, ontologies play a crucial role in supporting the exchange of data, providing a formal vocabulary for the information and unifying different views of a domain in a safe cognitive approach (Pazienza & Vindigni, 2002).

In this paper, we describe the knowledge model that has been adopted in CROSSMARC, an e-retail product comparison multi-agent system, currently under development as part of an EU-funded project, aiming to provide users with product information fitting their needs.

CROSSMARC technology operates in four languages (English, Greek, French and Italian) and is applied in two different product domains: computer goods and job offers.

2   CROSSMARC Architecture

The overall CROSSMARC architecture (see Fig. 1) is composed of a data processing layer (involving several language processing components), a database where relevant extracted information is stored and maintained, and a presentation layer (user interface).
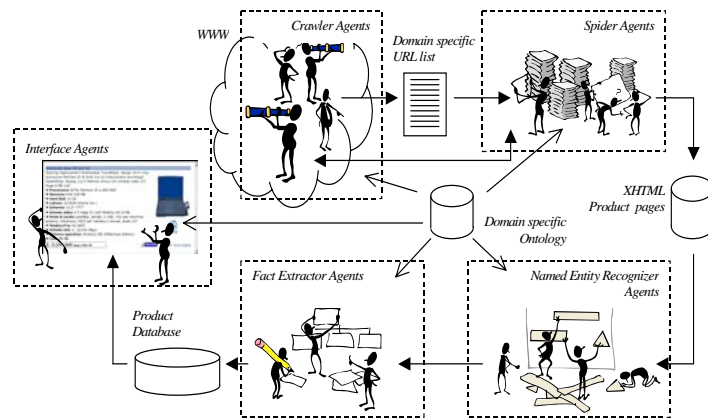

Fig. 1: architecture of the CROSSMARC system

Agents in the system could be broadly divided in three categories:

- retrieval agents, which identify domain-relevant Web Sites (focused crawling) and return web pages inside these sites (web spidering) that it is likely to contain the desired information;
- Information Extraction (IE) agents (a separate one for each language) which process the retrieved web pages, performing Named Entity Recognition and Classification (NERC) and Fact Extraction (FE), and finally populate a database with the extracted information;
- interface agents which process user's queries, perform user modelling, access the database and supply users with product information in their native language.

This architecture has been chosen to facilitate customisation into new languages, application domains, and/or services and to support the asynchronous activities mentioned above.

The CROSSMARC ontology represents the shared domain model for the purpose of extracting structured descriptions of e-retail products. It embodies a language neutral vocabulary of the domain, which is a formal description of relevant concepts that describe a specific product. In the overall processing flow, the ontology plays several key roles, as it is depicted in Figure 1:

- During Web page collection (focused crawling + spidering of web sites), it comes in use as a "bag of words", that is a rough terminological description of the domain in all the four languages that helps CROSSMARC crawlers and spiders to identify the interesting web pages.
- During Named Entity Recognition (NERC) it drives the identification and classification of relevant entities in textual descriptions (Grover et al, 2002). Also, ontology's structure is used in cross-lingual name matching. Each language-specific IE component identifies the ontological concepts for the entities that it recovers.
- During Fact Extraction and Normalization, entities identified in the NERC phase are correlated and aggregated to form language-independent specific product descriptions exploiting the ontology structure. All the values of product features in the final description are normalised to their canonical representation provided by the ontology.

- During the final presentation of the results to the end user, exploiting the correlation among the four lexicons and the ontology it is possible to provide cross-lingual descriptions of the same product. Thus, results are adapted to the language preferences of the end user who can also compare uniform summaries of product descriptions from pages in different languages.

3   Knowledge model

The realisation of CROSSMARC functions demands background knowledge at different levels (i.e. lexical, ontological and task oriented). Therefore, the ontological architecture is organized around three different layers:

- a meta-concetpual layer, which represents the common semantics that will be used by the different components of the system in their reasoning activities,
- a conceptual layer where relevant concepts of each domain are represented, and
- an instances layer where language dependent realizations of such concepts are organized.

The current ontology architecture is implemented in Protégé 2000 (Noy et, al. 2000), an ontology engineering environment that supports ontology development and maintenance. Protégé-2000 adopts a frame-based knowledge model, based on classes, slots, facets, and axioms. Classes are concepts in the domain of discourse, organized in a taxonomic hierarchy, while slots describe their properties. Facets and axioms specify additional constraints. In Protégé-2000, individuals are instances of classes and classes are instances of metaclasses.

In CROSSMARC metaclasses are used to constrain the intended interpretation of classes, slots and instances in the ontology. The basic idea is to define specific metaclasses to represent our model and use them to specify how the different elements are connected together (for instance, that a *feature* could have one or more *attributes* ranging over some *values*). In this way, the extension of CROSSMARC reasoning capabilities to a new semantic type requires the extension of metaclasses to allow for such new type, the declaration of how it fits the rest of the knowledge representation, and its instantiation it in the ontology.

The meta-conceptual layer of CROSSMARC defines how linguistic processors will work on the ontology, enforcing a semantic agreement by characterizing the ontological content according to the adopted knowledge model. The Protégé metaclasses hierarchy has been extended introducing a few metaclasses. These are used in the Conceptual level to assign computational semantics to elements of the domain ontology. Basically the metaclass extension provides a further typization to concepts, adding a few constraints for formulating IE templates.

The conceptual layer is organized around the three different modalities introduced in the previous section (see Fig. 2-c). All the conceptual model is rooted under three main classes: DOMAIN-TEMPLATE, DOMAIN-ONTOLOGY and DOMAIN-LEXICON. Each of these represents a specific knowledge aspect in the overall organization.

The DOMAIN-TEMPLATE component is made of specific elements that bring computational semantics on the concepts in the ontology. This semantics is strictly related to the IE task and is provided by the definition of three meta-elements: *Feature*, *Attribute* and *Value*. Essentially, the DOMAIN-TEMPLATE component describes a product offer as being composed by a set of features; this roughly corresponds to a structural description of "part-of" relationships (for instance, in the computer goods domain, they represent components as screen, CPU, devices, and so on). Each feature is characterized by a certain number of specific attributes that could range over some domain (for instance, the CPU feature could be characterized by the "Processor Type" and the "Processor Speed" attributes). DOMAIN-TEMPLATE is the superclass of each IE template in a domain. Its subclasses further characterize the kind of template they model.

DOMAIN-ONTOLOGY roots all the domain-relevant concepts. No computational specific semantics is adopted here, as this class (and its subclasses, the domain concepts) should only obey to a more general knowledge model. In this way knowledge engineers could model here in a

natural declarative form concepts, attributes and relations they feel relevant to describe the domain. A specific subclass of the DOMAIN-ONTOLOGY is the ABSTRACTION class, under which there are general abstract concepts as Measurement Units and numeric ranges.

Support for lexical information (i.e. language dependent realizations of ontology concepts) is rotted at DOMAIN-LEXICON class and its subclasses, ENGLISH, FRENCH, ITALIAN and GREEK LEXICON. Each lexical instance of these is composed by three slots: REFER-TO, a reference to a FEATURE, DOMAIN-CONCEPT or ATTRIBUTE; SYNONYM that holds multiple synonym terms describing a concept, feature or attribute; REG-EXPR, same as the previous but with regular expressions. Subclassing the four sublanguages makes easier to partition lexical information among the languages (as these contents should be filled by different information providers).
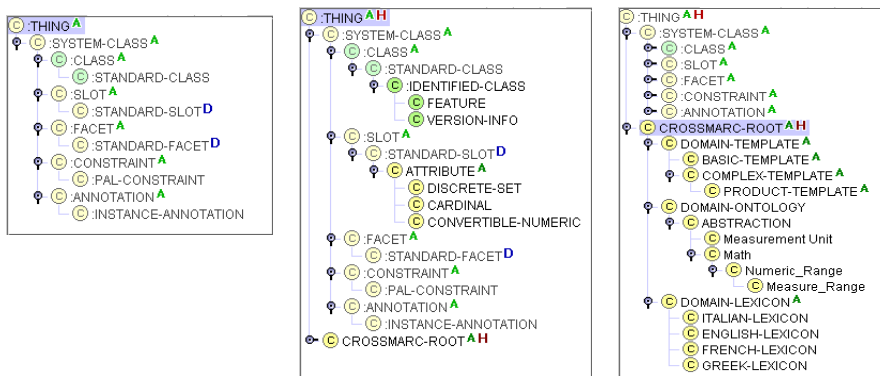

Fig. 2: Protégé (a) and CROSSMARC metaclasses (b) and Conceptual Layer (c)

The instance layer represent domain specific individuals. It instantiates classes in the domain ontology; these instances fill the values for attributes of the domain templates. There are two kind of instances in a CROSSMARC knowledge base: concept instances and lexical instances. Concept instances are subject to change over time, for instance, as technology evolves and new manufacturers, products and components appear, while others will become obsolete. Similarly, lexical instances could be refined, added, or adjusted to reflect linguistic bias (for instance, today no vendor refer to *Windows 95* while writing *Windows* in a product description).

4   Ontology Maintenance Process

Ontology maintenance concerns the addition, removal or reorganization of entries in the ontology. The impact of changes over the different ontology layers is strongly dependent on the modifications to be applied: extending the number of ontology concepts is expected to make no impact at all on the design of CROSSMARC agents, nor on the database structure (as long as the FE will maintain its original output data model); other actions, such as deleting concepts or entire branches, could potentially invalidate already extracted information. Lexical entries could be added, deleted or modified without affecting the conceptual layer, and the lexicon scheme could be modified as long as a reference to the concepts is maintained. On the other hand, changes in the conceptual layer usually affect the lexicons: changes to the main fabric of the DOMAIN-TEMPLATE component (i.e. attributes and features), could obviously have a heavy impact on all of the CROSSMARC processing steps and reasoning capabilities.

The maintenance process is performed through six different phases (see fig. 3): Domain Experts examine the current status of the specific domain to identify possible changes (1) and report the most significant ones to the Knowledge Engineers (2) in the form of new and/or misused concepts;

Linguistic Content Providers give coherent lexicalisations to the changes reported by the Domain Experts (3) and discuss them with the Knowledge Engineers; these then prepare new models for the ontology (4), taking into account the modifications to the domain model, and submit them to the Ontology Administrator (5),who will release a new version of the ontology based on the proposals received from the Knowledge Engineers, and releases it to the community; finally (6) Knowledge Engineers adapt lexicalisations by Linguistic Content Providers to the concepts specified in the new version of ontology.
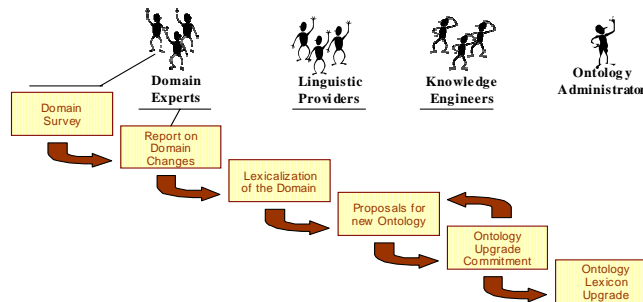

Fig. 3: Ontology Maintenance flow of processes

5   Conclusions

On the WWW the construction and use of ontologies have begun to replace the old-fashioned ways of exchanging business data in weak standardized formats with standard syntax (like XML/RDF) that adheres to semantic specifications given through ontologies. Ontologies help standardize the meaning of core concepts through different components and applications and facilitate the construction of services able to draw information from various sources in a uniform manner. However, any such standardization in the highly dynamic and expanding world of the Web is bound to face considerable difficulties, due to the considerable effort involved in adapting existing content to the new standards and following these standards in the construction of new content. The CROSSMARC ontology design and maintenance process aims to integrate these experiences by providing a general methodology that could be easily adapted across different domains and languages. The organization of the ontology has been designed to be applied to different domains without changing the overall structure, but simply changing relevant values; this has been obtained by decoupling the lexical component (language-dependent) from the conceptual one and inscribing the domain model in a widely assessed framework. Also, this architecture provides us with a language-independent and a homogeneous approach for presenting data in the graphical user's interface.

References

N. F. Noy, R. W. Fergerson, & M. A. Musen (2000). The knowledge model of Protege-2000: Combining interoperability and flexibility. *2nd International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pins, France.

M. T. Pazienza and M. Vindigni. (2002). Language-based agent communication. 13th *International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, Sigueza, Spain.

C. Grover, S. McDonald, D. Nic Gearailt, V. Karkaletsis, D. Farmakiotou, G. Samaritakis, G. Petasis, M.T. Pazienza, M. Vindigni, F. Vichot and F. Wolinski (2002): Multilingual XML-Based Named Entity Recognition for E-Retail Domains. *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain