An Ontology-Based Knowledge Management Platform

A.Aldea[2], R.Bañares-Alcántara[1], J.Bocio[1], J.Gramajo[2], D.Isern[2], A.Kokossis[3], L.Jiménez[1], A.Moreno[2], D.Riaño[2]

[1]Universitat Rovira i Virgili, Dept. Chemical Enginering. Tarragona, Catalonia, Spain
{*rbanares,jbocio,ljimenez*}*@etseq.urv.es*

[2]Universitat Rovira i Virgili, Dept. Computer Science and Mathematics. Tarragona, Catalonia, Spain
{*aaldea,jgramajo,disern,amoreno,drianyo*}*@etse.urv.es*

[3]Univ. of Surrey, Dept. Chemical and Process Eng. Guildford, Surrey, UK. *A.Kokossis@surrey.ac.uk*

Abstract

We describe the development of a knowledge management platform for web-enabled environments featuring intelligence and insight capabilities. The effort is the result of a FP5 project under the IST initiative involving 3 universities, a technology provider and 5 user companies. The main objective of the platform is to analyse, search and present information retrieved from the web) (or any other type of document). This is achieved through the use of *Multi-Agent Systems* and *ontologies*.

The automatic evolution of dynamic ontologies requires the action of a collection of agents to extract information and discover links using classification and learning techniques. These general-purpose agents will maintain a goal to periodically access the ontology and support search functions. Conceptually similar documents would get clustered into categories and information could then be retrieved by statistical approaches. Discovery of new knowledge would lead to modifications in the ontology by pruning irrelevant sections, refining its granularity and/or testing its consistency.

1 Introduction

The knowledge assets of a company consist of the knowledge regarding the products, markets, technologies and organisations of a business. Knowledge adds value through a set of business processes at strategic, tactical and organisational levels. In technology intensive companies the knowledge management challenges require a tentative and cautious review of the technological domains as well as venues to monitor and assess the way those domains evolve, emerge, mature, and decline. Benefits in utilising knowledge management practices include the enhancement of creativity and innovation, the strengthening of position, competence and responsiveness. The ability to grasp the dynamic profile of a discipline may lead to impressive gains of wealth and employment security. In contrast, slow reaction to developing dynamics may cost the viability of business and/or thousands of jobs. A good example of a dynamic and knowledge-intensive technological area is chemical engineering. It features a domain with hundreds of disciplines, accounts for a business sector valued over 1.3 trillion euros with 34,000 enterprises only in the EU, and is a domain where companies hardly maintain a national identity, as they span activities all over the world.

Engineers typically assess the evolution of their disciplines by reading journals, attending conferences or, quite often, by hearsay. Instead, the web (as well as other information resources) offers scattered and distributed information that is impossible to analyse manually. It has been estimated that the World Wide Web contains more than 300 million static objects [Bharat and Broder, 1998] accessible through 100 million internet hosts [Cameron, 2002]. In addition, organisations have intranets amounting to several million pages. The large majority of these documents are weakly structured. These repositories are usually searched by means of keyword-based search engines allowing a user to retrieve information by stating a combination of keywords. Documents downloaded from the web are indexed according to their contents, and only those matching the query (according to some metric) are returned to the user. The results of this type of search usually suffer from two problems derived from the nature of the query and the lack of structure in the documents: some of the retrieved documents are irrelevant, and some of the relevant documents may not have been retrieved (low precision and recall ratios). While search engines provide support for the automatic retrieval of information, the tasks of extracting relevant information and its further processing remain to be done by the human user.

The performance of a search engine can be improved by the use of an *ontology* [Fensel, 2001]. In its conventional form an ontology accounts for the representation of shared concepts in a domain by specifying a hierarchy of terms facilitating communication among people (collaboration) and applications systems (integration of tools).

In this paper, the automatic evolution of dynamic ontologies is supported by a *Multi-Agent System* (MAS) ([Wooldridge, 2002]), that uses classification and learning techniques to extract information and to discover new concepts from the retrieved pages. These general-purpose agents periodically access a user-given ontology and support search functions resulting in the retrieval of documents related to the ontology concepts. Conceptually similar documents will get clustered into categories; information will then be extracted by statistical approaches. Discovery of new knowledge will

lead to modifications in the ontology: pruning of irrelevant sections, addition of new branches, refinement of its granularity and/or testing of its consistency [Bañares-Alcántara *et al.*, 2003; Kokossis and Bañares-Alcántara, 2003].

In the next section we explain the two main paradigms which are the basis of our work: ontologies and multi-agent systems. We then present the proposed architecture for a knowledge management platform. Afterwards, we explain in detail the procedure of the search information module. Finally, we discuss the conclusions and the future work.

2 Background

In this project we propose the combination of two paradigms: one for building distributed systems and the other for knowledge representation.

2.1 Information agent systems

Agent paradigm is a promising technology for information retrieval. *Agents* provide some advantages with respect to traditional systems such as scalability, flexibility, autonomy, sociability, proactivity, etc. [Shah *et al.*, 2003]. Information agents provide access to information sources on behalf of a user or other agents [Weiss, 1999; Wooldridge, 2002].

Agents can collect information from the web by taking advantage of semantic annotations in a document, i.e. additional information provided by the document creator. Annotations are machine processable and add structure and/or semantics to the document (meta-information). However, most of the information stored in electronic format is expressed in XML (structured information), HTML (semi-structured information) or as text files (unstructured information). Web services and wrappers can be used to obtain information and to parse it into a structured format, for example into a database [Staab, 2003].

Existing systems of web data extraction (such as tools based on natural language processing) ([Laender *et al.*, 2002]) may be easily included in a multi-agent system building a *wrapper* ([Brenner *et al.*, 1998]) around it. A wrapper is able to translate requests from another agents in the representation of this existing system.

Several projects applying MAS to information retrieval such as [Corchuelo *et al.*, 2002; Gibbins *et al.*, 2003; Gómez *et al.*, 2001] demonstrate that agents can provide domain independence and flexibility to this type of systems. On the one hand ontologies provide the knowledge representation, on the other hand agents perform the actions.

2.2 Ontologies

An ontology is a vocabulary of entities, classes, properties, functions and their relationships. Ontologies are meant to provide an understanding of the static domain knowledge that facilitates knowledge sharing and reuse. [Fensel, 2001] identifies four different types of ontologies:

i) *Domain ontologies*, representing a target domain, as engineering, medicine, etc.

ii) *Generic or Common Sense ontologies*, capturing general knowledge about time, space, events, etc.

iii) *Method ontologies*, describing specific tas, as diagnosis.

iv) *Metadata ontologies*, describing the content of on-line information sources.

Several ontology representation languages have been developed in the last few years [Gómez-Pérez and Corcho, 2002]: XML (Extended Markup Language, see www.xml.com), RDF (Resource Description Framework, see www.w3.org/RDF/), DAML+OIL (see www.daml.org), etc. According to [Alexaki, 2000] RDF seems to be well positioned to become the standard to represent ontologies in the future.

Ontologies have been found to be useful for:

a) Retrieving the appropriate information from documents by providing a structure to annotate the contents of a document with semantic information [Alani *et al.*, 2003; Gibbins *et al.*, 2003].

b) Integrating the information from various sources by providing a structure for its organisation and facilitating the exchange of data, knowledge and models [AgentCities.NET, 2000; OntoWeb, 2002].

c) Ensuring consistency and correctness by formulating constraints on the content of information [OntoWeb, 2002].

d) Creating libraries of interchangeable and reusable models [AgentCities.NET, 2000; OntoWeb, 2002].

e) Supporting inference to derive additional knowledge from a set of facts [Gómez *et al.*, 2001; On-To-Knowledge, 1999].

The use of ontologies implies the study of the following topics:

- *Design and implementation*. An ontology is a representation of a domain. Thus, there could be many points of view for the same concept.
- *Validation*. We must evaluate the ontologies and make changes iteratively until an expert in the area thinks that the representation is accurate and complete.
- *Representation*. We must use a representation language to save ontologies. [Gómez-Pérez and Corcho, 2002] analyse some available languages and show their advantages and disadvantages.
- *Storage*. We need a representation to maintain a repository. We will use the ontologies to reason about a domain, therefore we also need a query language.

3 The agent-based knowledge management architecture

We propose a distributed architecture to perform the retrieval of information from the web and its further retrieval. As shown in Fig. 1, the system has three main parts: the *user interface*, the *search module* and the *knowledge generation module*.

In the next sections we describe in more detail this architecture and its main features.

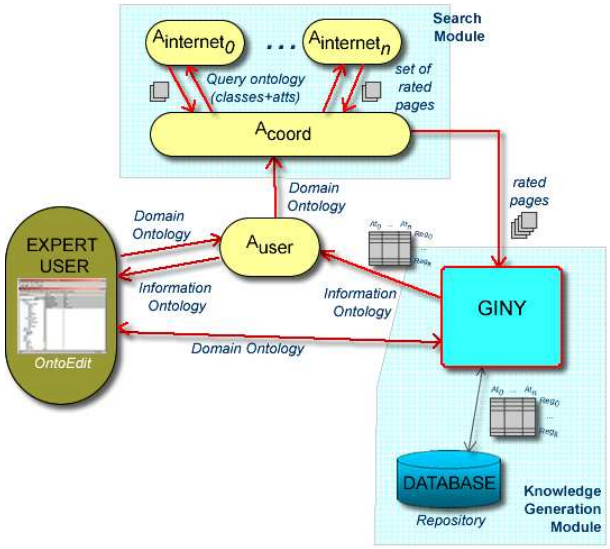

Figure 1: Dynamic Knowledge Management System

3.1 The user interface

The user interface consists of two parts: a *Graphical Interface* and a *User Agent*. The first one allows the user to specify the actions to be performed by the system, and to visualise the search results. Additionally, the user may retrieve an ontology from a repository or generate one by using an editor such as OntoEdit [Sure *et al.*, 2002]; this ontology is called *domain ontology*.

The second part, the *User Agent* (A_{user}), represents a user of the system. It stores all personal data, such as user preferences or profiles.

3.2 The search module

This module has an agent-based system whose main goal is to retrieve a collection of relevant web pages, rate and sort them according to different criteria.

In a Multi-Agent System, each agent is an autonomous entity with its own beliefs and goals. In the proposed system we define two types of agents:

- *Coordinator Agent* (A_{coord}). The search process has to be coordinated around some information agents. The coordinator has two tasks: splitting the *domain ontology* and merging the results.
- *Internet Agents* ($A_{internet}$). These agents encapsulate the analysis of pages from the web with a given criterion.

This module is explained in more detail in §4.

3.3 The knowledge generation module

As an output of the previous module, we have a set of web pages for each class of the *domain ontology*. The main goal of the knowledge generation module is to analyse these pages in order to discover instances of the classes of the *domain ontology*. For instance, if we have a class company, that has the properties company_address, company_fax_number, company_name, company_phone_number and company_contact_email, the module will try to discover values for each property.

We define *information ontology* as a *domain ontology* populated with the instances discovered by this module.

The GINY system [Gramajo and Riaño, 2002] is being implemented to analyse *structured*, *semi-structured* and *unstructured* documents by translating XML files into a table, analysing tables in HTML format and applying several strategies (e.g. regular expressions and WordNet [Fellbaum (ed.), 1998]) to text files respectively. All discovered instances are stored in a repository for further updates [Gramajo and Riaño, 2002].

4 The search process

4.1 Architecture

Information agents are elements that collect information from the web [Klusch, 2001]. In our case, the information agents are *Internet Agents* that try to discover web pages that are interesting for a given scenario.

As shown in Fig. 1, the search information module is composed by a *Coordinator Agent* and a set of *Internet Agents*. The search is monitored by the *Coordinator Agent* (A_{coord}) that receives requests from the *User Agent* (which sends a *domain ontology* to A_{coord}).

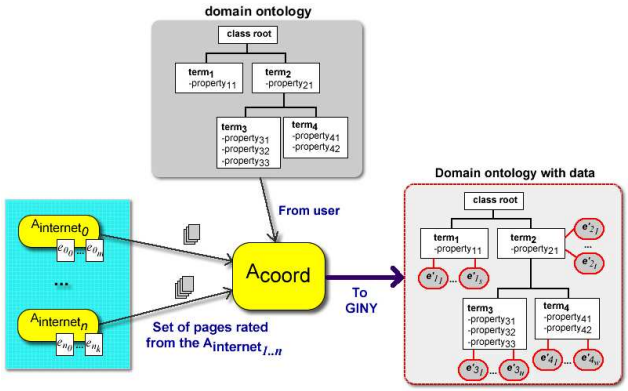

Figure 2: Merging of results by the *Coordinator Agent*

The *domain ontology* received by the coordinator is a hierarchy of classes and properties that is split in several parts; each part, that can contain one or more classes, is sent to an *Internet Agent*. After performing the search, each *Internet Agent* will return the results to the coordinator. Then, all these results will be merged using the original ontology (see Fig. 2). This process removes repeated and similar[1] elements and composes the final *domain ontology* with the retrieved web pages, which will be the core input to the knowledge generation module.

Each *Internet Agent* ($A_{internet}$) receives a part of the ontology (a class or a set of classes, called *query ontology*) and

[1] In [Dhyani *et al.*, 2002] there are different types of Web Page Similarity metrics such as content-based, link-based and usage-based.

performs the search within a deadline specified by the user. When the deadline is reached, all *Internet Agents* return an ordered set of web pages.

Internet Agents perform two main functions: *selection of URLs* and *ranking*:

a) The *selection process* retrieves the more interesting pages related to a concept of the *query ontology*. Initially, the agent begins the search with a set of pages given by an expert in the area (i.e. with previous knowledge), or with the set of pages retrieved by a search engine such as Google[2] (i.e. without previous knowledge). The class name and the root class name are used as keywords in the search. The agent analyses these pages and their links up to a user-given depth.

b) The *ranking* process is the most important feature of the *Internet Agents*. The rate of a web page p, shown in Eq. 1, is a function, $rate(p)$, that depends on local information (C) from the web page (frequency of terms, position of terms) [Dhyani *et al.*, 2002], domain information (O) from the ontology (relation with other terms), parametric information (SP) from the user (preferences, profile), and global information (RP) from previous results of retrieved pages.

$$rate(p) = f(C, O, SP, RP) \quad (1)$$

We are currently investigating the best relation between all these elements in order to find an *optimal* function f.

A first version of the multi-agent search engine has been implemented. A subset of a biosensor domain ontology has been used to test the system ([Bañares-Alcántara *et al.*, 2003]). Preliminary results obtained are quite encouraging, and show that the combination of *Internet Agents* and a *Coordinator Agent* in a client-server topology has better performance when the agents are deployed in several machines and these agents have not too many classes to consider.

4.2 Advanced search features

As Eq. 1 shows, each page has a rate that depends on a set of search parameters and the ontology. Unfortunately, sometimes this calculation does not estimate an adecuate rate because it is possible that a page does not contain enough concepts or attributes. One step forward would be:

- to use more than one ontology in one search, and
- to define the concept an *extended page*, i.e. a page and its neighbourhood of linked pages (up to a given depth) and evaluate its relation with the ontology.

We propose different research lines:

i) *Weighted Ontology*
Sometimes an ontology is not expressive enough to be used for search because the hierarchy (although well defined) gives the same weight to all classes. For instance, imagine that we have designed an ontology with three classes: biosensors, companies (a subclass of biosensors), and devices (also a subclass of biosensor); if we are only looking for information about companies, the system will spend too much time searching information about biosensors and devices. To avoid this, we could assign a weight to each class to better focus the search. Following the example, we could assign a weight of 0.2 to both, biosensors and devices, and a weight of 0.6 to companies (the concept we are more interested in).

ii) *Foreground/background ontology*
The previous idea could also be extended to ontologies. We could assign different weights to the ontologies and combine the results depending on these values. In Fig. 3 we show an example. In this case we would like to find information mainly about the area of biotechnology, but also, we would like to know whether a relation can be found with environmental (legislation) and chemical engineering.
In this case, we call the biotechnology ontology, the *foreground ontology* and legislation and chemical engineering, the *background* ontologies.

iii) *Extended page search*
It is possible that a web page does not contain all the concepts being searched, but if we consider a web page and its neighbours (*extended page*), the search could be improved. For instance, if a web page A contains information about the concept biosensor and page B contains information about the concept companies, if A has a link to B, we could assign a higher rate to page A. In this case, we consider only one level of neighbourhood, but we could extend this idea to any number of levels.

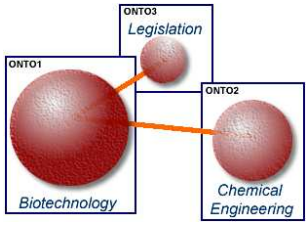

Figure 3: A view of the relation of ontologies during search

5 Conclusions and future work

This paper describes a distributed knowledge management system. The main objective of the platform is to improve the capabilities of industries to monitor, predict and respond to technological, product and market trends and changes. The retrieval and analysis of information from the web (or any other type of resource) are achieved through the use of multi-agent systems and ontologies. Ontologies are used to specify the queries, whereas the search is done by a set of *Internet Agents*, each of them acting on a part of the original ontology. Each agent in the system plays a set of roles defined in §3. With this architecture, the system could be updated easily in the future by adding new features or improving exisiting ones. Once the search engine is completed and validated, it will be connected to the knowledge generation module to automatically extract information relevant to the user form the web pages.

[2] www.google.com

In the next phase of the project, we will work on the organisation of the retrieved information. This information will be merged into a single *information ontology* where the discovery of new knowledge will take place. New concepts, properties or values of properties will be extracted from the retrieved web pages and added to the existing it information ontology, so the ontology will be dynamically updated.

Also, we are planning to study whether some type of coordination among *Internet Agents* could improve the results. For instance, to avoid repeated accesses to the same web page by different agents.

6 Acknowledgements

This work has been funded by the EU Project *hTechSight* IST-2001-33174. We would like to thank all the partners for their feedback, especially Dr. P. Linke at U. Surrey.

References

[AgentCities.NET, 2000] AgentCities.NET. IST project IST-2000-28384 Agentcities, 2000. http://www.agenticities.net/.

[Alani *et al.*, 2003] H. Alani, S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis, and N. Shadbot. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1):14–21, 2003.

[Alexaki, 2000] S. Alexaki. Managing rdf metadata for community webs. In *2nd International Workshop on the WWW and Conceptual Modelling (WCM'2000)*, pages 140–151, Salt Lake City, USA, 2000.

[Bañares-Alcántara *et al.*, 2003] R. Bañares-Alcántara, A. Kokossis, A. Aldea, L. Jiménez, and P. Linke. A knowledge management platform to extract and process information from the web. In *Process System Engineering (PSE'03)*, Kunning, China, 2003.

[Bharat and Broder, 1998] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engine. In *7th WWW Conference*, Brisbane, Australina, 1998.

[Brenner *et al.*, 1998] W. Brenner, R. Zarnekow, and H. Wittig. *Intelligent Software Agents. Foundations and Applications*. Springer Verlag, Berlin, Germany, 1998.

[Cameron, 2002] I. Cameron. Web-based cape systems - now and the future -. In *CAPE Forum*, Tarragona, Catalonia, Spain, 2002.

[Corchuelo *et al.*, 2002] A. Corchuelo, J.L. Arjona, and A. Ruiz. Automatic extraction of semantically-meaningful information from the web. *UPGRADE*, 3(3):44–51, 2002.

[Dhyani *et al.*, 2002] D. Dhyani, W. Keong N., and S.S. Bhowmick. A survey of web metrics. *ACM Computing Surveys*, 34(4):469–503, 2002.

[Fellbaum (ed.), 1998] C. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusets, US, 1998. ISBN 0-262-06197-X.

[Fensel, 2001] D. Fensel. *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*. Heidelberg, Germany, 2001.

[Gibbins *et al.*, 2003] N. Gibbins, S. Harris, and N. Shadbolt. Agent-based semantic web services. In *The Twelfth International World Wide Web Conference (WWW2003)*, Budapest, Hungary, 2003. ACM Press.

[Gómez *et al.*, 2001] M. Gómez, C. Abasolo, and E. Plaza. Domain-independent ontologies for cooperative information agents. *Lecture Notes in Artificial Intelligence*, 2128:118–129, 2001.

[Gómez-Pérez and Corcho, 2002] A. Gómez-Pérez and O. Corcho. Ontology languages for the semantic web. *IEEE Intelligent Systems*, 17(1):54–60, 2002.

[Gramajo and Riaño, 2002] J. Gramajo and D. Riaño. A knowledge management platform to extract and process information from the web. In *5th Joint Conference on Knowledge Based Software Engineering*, Maribor, Slovenia, 2002.

[Klusch, 2001] M. Klusch. Information agent technology for the internet: A survey. *Journal on Data and Knowledge Engineering, Special Issue on Intelligent Information Integration*, 36(3), 2001.

[Kokossis and Bañares-Alcántara, 2003] A. Kokossis and R. Bañares-Alcántara. Dynamic information management for web-enabled environments in the chemical process industries. In *Foundations of Computer-Aided Process Operations (FOCAPO'2003)*, Coral Springs, Florida, US, 2003.

[Laender *et al.*, 2002] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira. A brief survey of web data extraction tools. *ACM SIGMOD Record*, 31(2):84–93, 2002.

[On-To-Knowledge, 1999] On-To-Knowledge. IST project IST-1999-10132 On-To-Knowledge, 1999. http://www.ontoknowledge.org/.

[OntoWeb, 2002] OntoWeb. IST project IST-2000-29243 OntoWeb, 2002. http://www.ontoweb.org.

[Shah *et al.*, 2003] U. Shah, T. Finin, A. Joshi, R.S. Cost, and J. Mayfield. Information Retrieval on the Semantic Web. In *10th International Conference on Information and Knowledge Management*. ACM Press, 2003.

[Staab, 2003] S. Staab. Web services: Been there, done that? *IEEE Intelligent Systems*, 18(1):72–85, 2003.

[Sure *et al.*, 2002] Y. Sure, S. Staab, and J. Angele. Ontoedit: Guiding ontology development by methodology and inferencing. In *International Conference on Ontologies, Databases and Applications of Semantics (ODBASE'02)*, Irvine, USA, 2002.

[Weiss, 1999] G. Weiss. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT Press, Cambridge, Massachusets, US, 1999. ISBN 0-262232030.

[Wooldridge, 2002] M. Wooldridge. *An Introduction to Multiagent Systems*. John Wiley and Sons Ltd, Chichester, England, 2002.