RECIIS
Electronic Journal of Communication Information & Innovation in Health
[www.reciis.cict.fiocruz.br]
ISSN 1981-6286

Technological Advances

From MASTER-Web to AGATHE: the evolution of an architecture for manipulating information over the Web using ontologies

DOI: 10.3395/reciis.v2i1.137en

Fred Freitas
Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
fred@cin.ufpe.br

Luciano Cabral
Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
lsc4@cin.ufpe.br

Rinaldo Lima
Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
rjl4@cin.ufpe.br

Eunice Palmeira
Coordenação de Informática, Centro Federal de Educação Tecnológica de Alagoas, Maceió, Brazil
eunice@cefet-al.br

Guilherme Bittencourt
Departamento de Automação e Sistemas, Universidade Federal de Santa Catarina, Florianópolis, Brazil
gb@das.ufsc.br

Bernard Espinasse
Laboratoire des Sciences de l'Information et des Systèmes Universités d'Aix-Marseille, Domaine Universitaire de St Jerôme, Marseille, France
espinasse@univ-cezanne.fr

Sébastien Fournier
Laboratoire des Sciences de l'Information et des Systèmes Universités d'Aix-Marseille, Domaine Universitaire de St Jerôme, Marseille, France
sebastien.fournier@isis.org

Abstract
This article presents two architectures for information gathering systems on restricted Web domains, for example the academic or the biologic domain. This text processing is based on the use of domain-related ontologies employing them as a well-defined and understandable semantic model for the software. If, on one hand, the solution here presented cannot be scaled to the entire Web, on the other hand, the offered services are more versatile and precise and able to combine information with well-defined relationships distributed over the Web. The presented systems are still able to draw inferences about the information present in the Web about these domains. As a proof of concept, we present experiments with good results in two distinct domains, showing the feasibility and portability between domains of the presented solution besides presenting a high degree of reuse during the portability.

Keywords
Multi-agents, information agents, agent-oriented software engineering, cooperative manipulation of information, classification of information

Introduction

Currently, nearly every scientific publication is also made available in (different) electronic formats. The technological revolution undeniably increased the volume and availability of information, mainly in the internet (PALMEIRA et al., 2006).

This resulted in what we call information overload; the users must cope with the difficult task of searching for useful information among an enormous quantity of available documents. In order to minimize this problem, the search engines, among them Google, which practically dominated this market, where created based on techniques developed by the area of information retrieval (RIBEIRO-NETO et al., 1999).

Search engines like Google however (GOOGLE 2008) present a number of deficiencies mainly related to the way they were conceived. In the first place, it is very common that in reply to a key word-based search the user will receive a great quantity of useless or irrelevant pages. As the search engines are using statistical ranking algorithms for attributing relevancy to the pages, their search does not use semantics because these engines have only the capacity to represent the pages on lexical level (FREITAS, 2002). When comparing the services of the search engines with those of database management systems, the limitations of the first-mentioned become evident: databases can be easily searched because they store data about a restricted context, in a structured way, and without ambiguity. Therefore database management systems can provide the user with semantically clear and precise answers about entities and the relations between them, including combining and totalizing data, tasks that search engines are not able to perform. A user could, for example, search the system for articles in the field of "neural networks", published in scientific events held in Asia later than 2000.

On the other hand, since we are dealing with pages about a restricted domain (like the academic domain, for example) and using knowledge about this domain, in addition to a simple retrieval we can offer a more deep-reaching treatment of the information providing additional benefits to the user. These additional tasks are described in the next sub-section.

Other tasks related to information retrieval: classification and extraction of information

The tasks that can be performed further to information retrieval are:

• Classification of pages, a process that will determine to which category of pages a determinate page belongs. A page describing a congress, for example, will fall into the category Scientific Event, more specifically into the sub-category Congress;

• Extraction of information. According to KUSHMERICK (1999), extraction of information is "the task of identifying the specific fragments of a single document that constitute its core semantic content" From the sentence *"The Parliament was bombed by the guerrillas",* for example, processed as belonging to the domain Terrorism, three pieces of information were extracted: the kind of the terrorist act: bombing; the target of the terrorist act: the Parliament; and the authors of the terrorist act: the guerillas.

Complementarity between information retrieval, classification and extraction

The initial premise of this work was the hypothesis that retrieval, categorization and extraction can and should be performed in an integrated way. As shown in Figure 1, this would improve the performance of each of these processes.

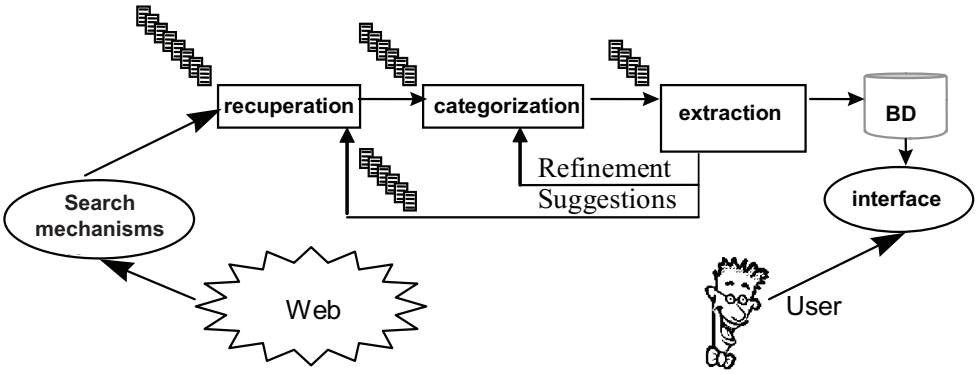


Figure 1 - Outline of the architecture of an extraction system, demonstrating the complementarity between retrieval, categorization and extraction.

Source: FREITAS, F. Sistemas Multiagentes Cognitivos para a Recuperação, Classificação e Extração Integradas de Informação da Web. Doctors' thesis, Federal University of Santa Catarina, 2002.

Retrieval systems can provide access to an initial set of pages about a certain domain, for example scientific events. Categorization systems should then select the pages that belong to the classes to be processed and, finally, extractors could capture the requested information, for example date and place, where the event took place. During the extraction process, it would be possible to find within the processed pages links to pages belonging to other, also processed classes, for example home pages of researchers. Such suggestions are presented in a semantic and safe context and thus the extraction process would help in the retrieval of information depending on the cooperation of the agents (for instance, the agent "scientific events" sends addresses of the pages of the scientific committee to the agent "researchers"). The extraction could still refine the categorization by disregarding pages not containing the data that characterize the processed class, for example events without stating a date.

It must be pointed out that the integrated execution of these three tasks is only possible thanks to an indispensable requisite: domain restriction. In other words, the system executing these tasks is only able to process pages belonging to a certain domain. Thus, for taking advantage of a solution of this nature, the solution should be portable between distinct domains.

Ontology-based portability between domains

The problem related to domain portability provokes many times drastic alterations in computer systems. On the other hand, the use of declarative programming solutions, i.e. solutions in which the necessary knowledge is located outside the system and not in its executable code, facilitates the implementation of portability.

A declarative solution currently very much in vogue is ontology. Ontologies are present in a great number of systems, tools and products for information manipulation and electronic commerce, represented as keyword hierarchies, concepts and many other forms.

Although the term "ontology" denotes the study of the nature of being, in computer science the term can be interpreted as a set of entities and their relations, restrictions, axioms and vocabularies. Ontology defines a knowledge domain or, more formally, specifies a conceptualization about it (GRUBER, 1995). Normally, an ontology is organized in concept hierarchies (or taxonomies).

Figure 2, for example, shows some classes of the ontology "Science" (FREITAS, 2001) and some of their relations. The class "Member of the Academic Staff" is a more specific subset of the class "Researcher" characterizing the relation known as inheritance. This class inherits the entire knowledge associated with the class Researcher such as the relations, restrictions and other items of knowledge of the latter class. The figure still contains other relations between classes as, for example, the relation HeadOF connecting the classes Member of the Academic Staff and Project, signaling that a project is coordinated by a Member of the Academic Staff. An example for restriction would to impose cardinality in this relation: only one Member of the Academic Staff participates in this relation.

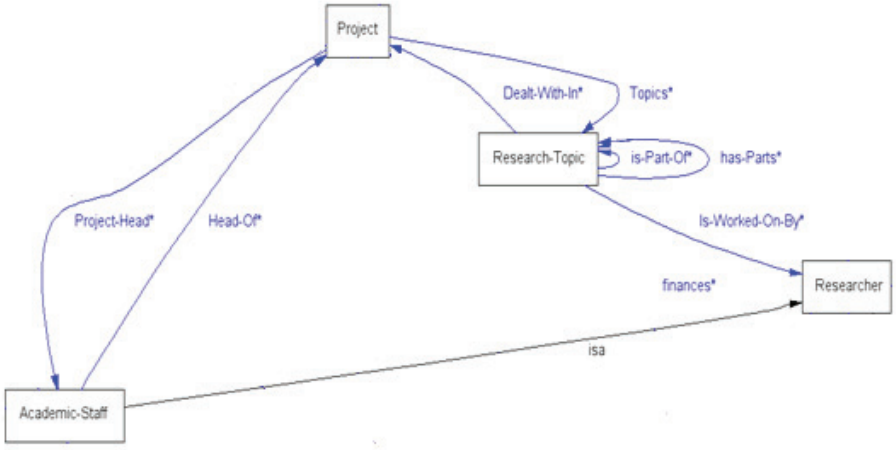


Figure 2 - Relations between some of the principal classes of the ontology of Science.

Source: FREITAS, F. 2001. Ontology of Science. http://protege.stanford.edu/plugins/ontologyOfScience/ontology_of_science.htm.

After describing these requisites - integrated manipulation of information over the Web involving retrieval, classification and extraction of information from the Web and use of ontologies for restricting the domain to be treated - in the next sections we will describe the study conducted by the authors in this research area. Two systems with different architectures MASTER-Web (FREITAS et al., 2003) e o AGATHE (ESPINASSE et al., 2007) will be described in detail and their differences and similarities as well as some case studies confirming the portability of these solutions will be presented. In the end of the article we present some investigations related to these topics and compare the quality of the two proposals.

The MASTER-Web system

The MaSTER-Web system (*Multi-Agent System for Text Extraction, Classification and Retrieval over the Web*) (FREITAS et al., 2003) consists of cognitive multi-agent systems for solving the problem of integrated extraction of entities pertaining to classes that integrate a group of pages (cluster). This system showed good information retrieval, extraction and information results and allows cooperation between the agents for performing these tasks. This approach is knowledge-based and presents different types of reuse, seeing that the agents are sharing the same structure in terms of code, search services and mechanisms as well as a good part of the knowledge they can avail of (ontologies and production rules), facilitating and accelerating the construction of new agents (PALMEIRA et al., 2006).

This multi-agent system manipulates the information referring to a set of classes about the same group such as, for example, the scientific group including classes like scientific articles, events, researchers etc. The architecture as such aims at retrieving, classifying and extracting information from pages belonging to the classes of one group, and the main motivation for employing multi-agent systems is taking advantage of the relations between these classes. A general view of this architecture is given in Figure 3.

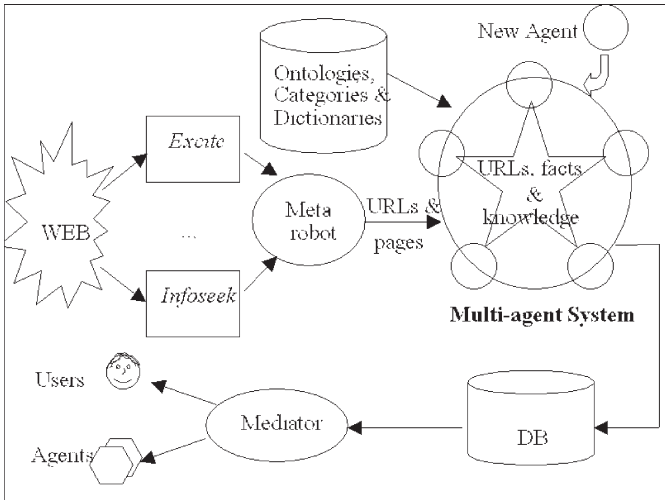


Figure 3 - Architecture of multi-agent systems for integrated mahipulation of information of groups of classes.

Source: FREITAS, F.; BITTENCOURT, G. An Ontology-based Architecture for Cooperative Information Agents. Proceedings of the Internacional Joint Conference on Artificial Intelligence – IJCAI'2003. Acapulco, México, 2003.

The agents represented by circles in this figure have the expertise to recognize and extract data from pages supposed to belong to the class of pages processed by the agent (the agent PPR, for example, processes pages of scientific publications while other agents of the same multi-agent system are responsible for processing other classes of the domain "science"). The agents recognize a page of their class principally when they recognize the existence of attributes referring to this class (se example below). The agents cooperate by exchanging messages containing recognition rules and facts (knowledge of the agents), besides exchanging suggestions of pages

The users can benefit from access to information extracted by a special agent called mediator (WIEDERHOLD, 1995). This mediator is able to help in the query by providing a more simple view of the database, allowing the user to formulate complex queries involving a variety of databases. For example, a user could make the query already described in section 1, searching for articles in the field of "neural networks" published in Asian events after 2000. It is noteworthy that the currently available search mechanisms such as GOOGLE (GOOGLE, 2008) do not have means for conducting a query of this kind.

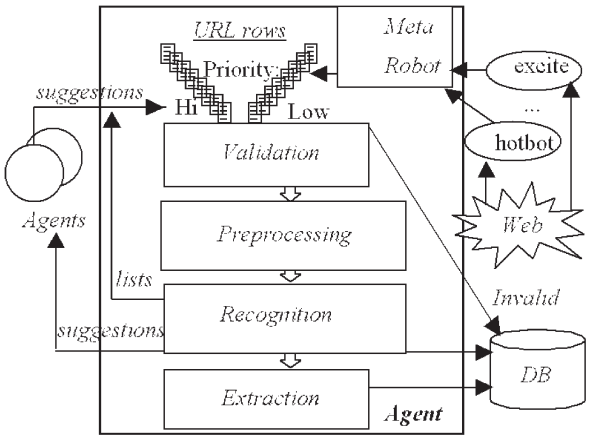
When an agent is added to the system, it registers and introduces itself by sending to all of the other agents a set of rules to be used by them on the recognition of pages likely to belong to its associated page class. The other agents update their recognition rules and send, in turn, their own recognition rules to the new agent. When a link or page fires any other agent's recognition rule, the agent sends the link or page to that agent. In the next section the functioning of an agent is described in detail.

Functioning and structure of an agent in the MASTER-Web

Each agent, in detail represented in Figure 4, recognizes, filters and classifies pages that correspond to entities of the page class it is processing (for example

pages of researchers, calls for papers) also extracting their attributes (for example research fields and institution of the researcher). Each agent utilizes a meta-robot that can be connected to multiple search engines like Altavista or Excite for instance. The meta-robot queries the search engines with terms that assure recall for that agent's page class (e.g., 'Call for papers' and 'Call for participation' for the CFP agent) (PALMEIRA et al., 2006).

Figure 4 - An agent and the four successive steps it performs in the processing.



Source: FREITAS, F.; BITTENCOURT, G. An Ontology-based Architecture for Cooperative Information Agents. Proceedings of the International Joint Conference on Artificial Intelligence - IJCAI'2003. Acapulco, México, 2003.

As can be observed in Figure 4, each agent performs four consecutive steps in the processing of each page (PALMEIRA et al., 2006): validation, pre-processing, recognition and extraction.

Validation rules out inaccessible or repeated Web pages and pages in formats, which the agent is unable to process. The Preprocessing step identifies the content, the title, the links, key words and their frequency among others, using techniques of information retrieval and, if necessary, natural language techniques. These data are passed to agent's inference engine. In the steps classification and extraction of attributes, the system discovers if the page fits into the treated domain, recognizes of which class the treated page is instance and extracts attributes, which will compose the instance of the class. These steps are performed through computational processes known as automatic reasoning or logic inference. In the next sub-section we will explain how these processes are implemented in the MASTER-Web.

The reasoning of the agents

Each of these classes has a set of attributes that have to be extracted or identified and whose presence can indicate if a page fits into a class or not. This process involves a combination of cases, rules and ontologies that are better explained using an example.

A very common case for the agent of scientific articles is that a page is recognized when containing in the beginning the abstract and the attributes First Name. Org-Name (name of the organization, to which one or more of the authors is affiliated and Location-Place (country or American state where the organization is seated). This case is described below:

```
([ppr_00356] of Case
  (Description "aff,1st,loc")
  (Concepts-in-the-Beginning [abstract])
  (Slots-in-the-Beginning
          [First-Name]
          [Org-name]
          [Location-Place]))
```

This case should be associated with a class to be recognized through a Class-Recognizer. As all articles published in scientific journals or in the proceedings of scientific events normally follow this pattern we associate this case to the class Part-Publication because articles are always part of a divisible publication (e.g. a book, or the proceedings of an event). This association is shown below:

```
([ppr_00528] of Class-Recognizer
  (Cases
          [ppr_00536]
[ppr_00356])
  (Class [Part-Publication]))
```

The relation is completed by rules reusable through instantiation, many of them common to various agents as illustrated in the next code.

```
Rule r_900_slots_hi_funct
  Having a STORED page instance and
having a case that has as list of attributes in the beginning of the page,
a list of concepts in the beginning of the page,
and a Class-Recongizer of an abstract case with a list of test cases and

  If
the specified case is in the list of cases
If some of the concepts contained in the beginning of the page are in the list and
if the attributes of the case are in the list of attributes found
Then
the page passes to be RECOGNIZED as fitting into the same class of the specified case.
```

Thus, as the greater part of recognition rules refer to cases they can be included in all agents and are completely reusable by agents of other cases, a fact that facilitates the construction of a new agent. In fact, a new agent only needs to define new cases of recognition, classification and extraction.

In general the recognition is made firstly utilizing an abstract class that cannot have instances - as Part-Publication in the Article agent or Live Scientific

Event or Publication Event in the CFP agent. Based on the recognition of the abstract class, classification rules including cases try to classify the page among its subclasses. If this does not occur, the page will be classified into a generic class like Part of Generic Publication in the Articles' or Live Generic Scientific Events' or Generic Publication Events' agents.

A case study

In its fist version the MASTER-Web was developed utilizing the inference engine JESS for the reasoning tasks and the Protégé ontology editor for the specification and manipulation of ontologies (PROTÉGÉ, 2008). The system was initially tested in the scientific domain (FREITAS et al., 2003), utilizing the ontology of Science (described in (FREITAS, 2001). For the production of this ontology, the ontology of the $(KA)^2$ project (*Knowledge Annotation Initiative of the Knowledge Acquisition Community*) (BENJAMINS et al. 1998) was reused and improved, mainly as refers to granularity. The main improvement was the addition of classes aimed at reorganizing the ontology from the viewpoint of classes with common characteristics. As a practical example we can mention the class Scientific Event that was divided into two subclasses, Live Scientific Event (with the subclasses conference and workshop, among others) and Scientific Publication Event (with the subclasses Journal and Periodical). With this, the recognizing capacity is increased and granularity and coherence are added to the ontology (FREITAS et al., 2003). This way, entities of the cluster (domain knowledge) can be identified with adequate granularity, representing the classification even with subtle differences between entities (ESPINASSE et al. 2007).

Three tests for classification of pages were made with each agent (FREITAS et al., 2003). The first two tests utilized the *corpora* of pages retrieved from queries at search engines (like Google and Altavista). The first *corpus* was employed in the process known as knowledge acquisition, consisting in the definition of cases and rules that will be helpful in the classification and extraction processes, running the system interactively until reaching an acceptable performance. The second *corpus* was used for a blind test. The third test collected candidate pages directly from the Web. Thus, these two tests evaluated the performance of the agents in face of new *corpora* not searched during the elaboration of the rules and cases.

Two agents of the scientific group were constructed. The agent CFP processes Call for Papers candidate pages for scientific events such as conferences and journals, classifying them into eight classes (the four beforementioned and Live Generic Event, Generic Publication Event, Journal Special Edition, Periodical Special Edition). The second agent processes candidate pages of scientific articles and documents, classifying them into articles from workshops, conferences, journals and periodicals, book chapter and generic articles, besides theses, dissertations, technical and project reports.

Results

The presented results refer to two sets of tests. The fist includes a total of four tests aimed at evaluating the system as refers to classification. Table 1 displays the performance of the CFP and Article agents. In this table the results are categorized according to the kind of *corpus* in the columns and type of classification in the lines.

The results seem quite promising: the greater part of results reached more than 90%. The rare cases of failure in the case of the CFP agent were due to long pages, generally referring to an issue or a community (ex: *Open Source*).

Table 1 - Performance of CFP and article agents

	"Call for papers" agent				Article agents		
	Acquisition corpus	Test corpus	Test in the web	Web using lists	Acquisition corpus	Test corpus	Test in the web
Recognition	97.1	93.9	96.1	96.3	93.1	82.7	87.8
Content classification	94.9	93.3	92.9	91.7	97.0	93.0	81.4
Processed pages	244	147	129	188	190	150	184

Source: FREITAS, F.; BITTENCOURT, G. An Ontology-based Architecture for Cooperative Information Agents. Proceedings of the International Joint Conference on Artificial Intelligence - IJCAI'2003. Acapulco, México, 2003.

Similarly, the agent PPR, which processed scientific articles, was not able to recognize articles with few attributes, attributes of difficult identification (for example, affiliation with an unknown company) or with attributes in the end of the article. This however occurs in similar systems like *CiteSeer* as well (BOLLACKER et al., 1998).

In the next section we will see the evolutions of the proposed architecture in the sense of providing it with two relevant requisites: portability to other domains and scalability, i.e. the possibility to handle much more pages.

Evolution

One of the major advantages of a declarative architecture is its genericity; just by changing the involved ontologies and knowledge bases, a deep change in the behavior of the system can be implemented. The following experiment proves that this requisite can be applied to architecture.

Portability across domains: MASTER-Web in the domain of Artificial Intelligence

This test proves that the portability of domains in knowledge-based systems to tasks related to information retrieval is possible without abrupt alterations in the system.

The alterations made for this test were the following:

• The ontology of scientific events was removed and an ontology of the area of Artificial Intelligence (AI) and some of its sub-areas were added;

• The representation of HTML pages in the system was extended to a representation of scientific articles in order to facilitate the identification of the typical sections occurring in scientific articles such as introduction, related works, conclusion, abstract and others. Recognition of these sections was included during preprocessing, a fact that was extremely helpful during the recognition and classification of articles in the field of AI;

• The sections were divided into two classes: relevant and irrelevant sections. The relevant sections refer specifically to the principal topics of the article, and are generally more concise and as refers to theoretical aspects (e.g. abstract, key words and conclusion). The sections irrelevant in terms of recognition and classification are less specific, presenting contents of a more general nature and from different areas. Citations, related and future works are among the sections considered irrelevant in terms of classification (PALMEIRA et al., 2006).

The recognition and classification experiments of AI articles utilized a quite heterogeneous *corpus* gathered from the Web, with scientific articles from different fields of knowledge but at the same time heterogeneous as refers to the division of the sections. For this test, we started from the premise that the system should show robustness in face of this lack of standardization of article sections (PALMEIRA et al., 2006).

The experimental basis was composed by html documents obtained in different sites of universities and homepages of professors and scientific journals. The experiments classified the articles based on the analysis of the sections of the text considered relevant and utilizing an AI ontology built specifically for this purpose. Separately, ontologies of sub-areas and applications were built, that are later grouped for composing the AI ontology. The scope of this ontology involves in a first moment the fields Search, Knowledge Representation, Neural Networks and Machine Learning. Ontologies of other sub-areas of AI can be incorporated, extending the existing ontology (PALMEIRA et al., 2006).

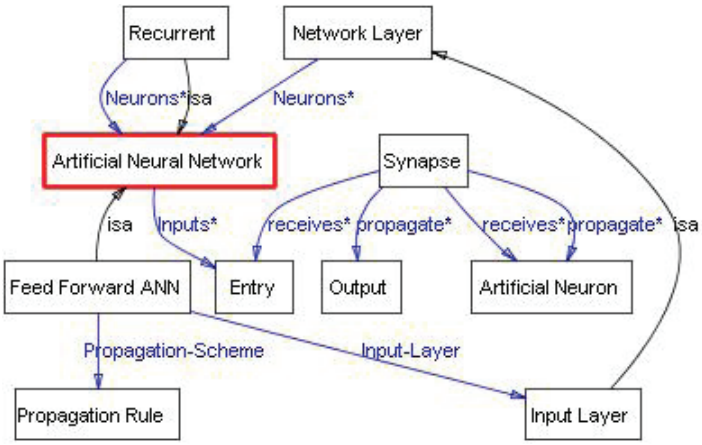


Figure 5 - Some classes and their relations in the domain Neural Networks of the Artificial Intelligence ontology.

Source: PALMEIRA, E. and FREITAS, F. Detailed Ontologies and text classification: a promising union, Proceedings of the Workshop on Building Applications with Ontologies for the Semantic Web, Encontro Português de Inteligência Artificial, 2006.

Figure 5 highlights some relations between the principal classes defined in the ontology Neural Networks. The figure presents two types of Neural Networks, *Recurrent* and *Feed Forward ANN*, as well as some relevant concepts such as Synapse, Neuron and Layer. Different relations are available between these concepts.

The ontologies of the branches Search, Knowledge Representation and Machine Learning were developed on a level of detail similar to that of the Neural Networks. The articles of the field of AI not dealing with areas making part of the ontology are not classified within this domain (PALMEIRA et al., 2006). Another relevant observation is that an article can involve more than one area, for example Machine Learning and Search (PALMEIRA et al., 2006). As shown in Table 2, the results obtained showed promising percentages in the classification of AI articles

Table 2 - Percentages of successful classification of articles according to field of knowledge

Recognition	Corrects	False positive	False negative	Corrects (%)
Artificial Neural Network	48	1	2	94.1
Knowledge acquisition	17	0	1	94.4
Knowledge Engineering	3	0	0	100.0
Knowledge representation formalisms	56	9	1	84.8
Machine learning	51	2	6	86.4
Ontology	19	0	0	100.0
Search	38	1	1	95.0
Other domains	228	7	11	92.7

Source: PALMEIRA, E. and FREITAS, F. Detailed Ontologies and text classification: a promising union, Proceedings of the Workshop on Building Applications with Ontologies for the Semantic Web, Encontro Português de Inteligência Artificial, 2006

Once the portability between domains has been proven, the next requisite for the architecture was scalability. This requisite was approached from a completely new perspective of the system in terms of functionalities and topology of the agents that required confection of a new system, the Agent GATHEring system, AGATHE (ESPINASSE et al., 2007), that will be detailed in the next section.

Description of the AGATHE system

The AGATHE system (ESPINASSE et al., 2007) is the result of a reengineering of the topology of MASTER-Web agents (FREITAS et al., 2003). The specific purpose of this system is to define a scalable, adaptable and extensible multi-agent architecture to facilitate intelligent information retrieval in the Internet. The architecture of the AGATHE system includes modularization and distribution of tasks, avoiding function overload that could occur in the base system. This way a better retrieval, extraction and classification performance is possible. This new architecture is based on a specific organization of agents defining different types of agents as described in the next subsection. This version integrates the JADE platform (Java Agent Development Environment TILAB, 2008), the inference engine JESS and the PROTÉGÉ environment.

Some new functionalities were included, for example cooperation between multi-agent systems of different domains such as, for example the domains Academia and Tourism. The pages of scientific events include a variety of aspects related to tourism (traveling, hotel-accommodation, tourist activities and social events, etc.). This allows us to catch a glimpse of the cooperation between agents of different domains: the CFP agent sends pages to agents of the tourism domain so that these can classify and extract information of interest to them. The MASTER-Web in its original conception could not have offered this service.

Architecture

The AGATHE system is composed by 3 interacting subsystems, the Search, Extraction and Front Office subsystems, described below and presented in Figure 6.

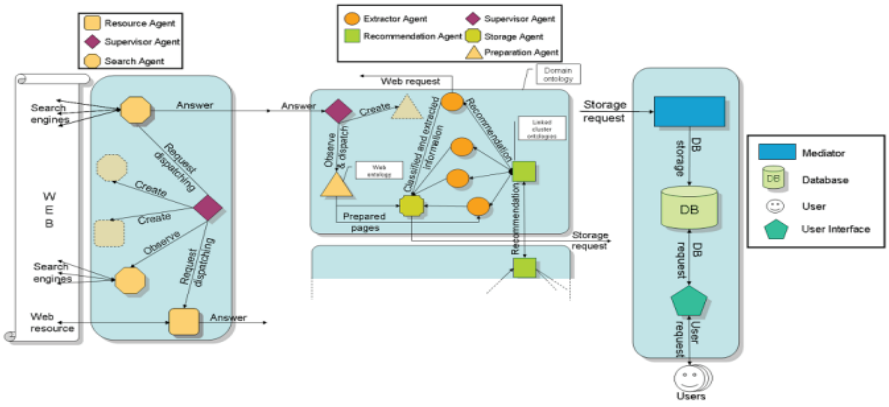


Figure 6 - Architecture of the AGATHE system with extraction subsystem in detail.

Source: ESPINASSE, B., FREITAS, F. and FOURNIER, S. AGATHE: an Agent and Ontology based System for Restricted-Domain Information Gathering on the Web. Proceedings of the International Conference on Research Challenges in Information Sciences (IEEE-RCIS), April 23-26, Ouarzazate, Morocco, 2007.

a) The Search subsystem is in charge of querying external search engines on the Web (such as Google) in order to select and process Web pages for retransmitting them to the Extraction subsystem.

b) The Extraction subsystem is composed of different extraction agents, specialized in the processing of Web pages on a specific field (like that of academic search or that of artificial intelligence, etc.)

c) The Front Office subsystem ensures the organization and storage of information extracted from the processed Web pages and provides a query interface with human or software agents.

All mentioned subsystems are multi-agent systems, composed of information agents utilizing the domain ontology for performing their tasks (ESPINASSE et al., 2007). An example for this is the Extraction subsystem containing different information agents as displayed in Figure 6, differently from the MASTER-Web, where all agents are of the same type and perform all tasks.

Supervisor Agent

Receives the queries' results from the Search subsystem and creates one or more Preparation Agents that will treat these results before transmitting them to the Extraction and Recommendation Agents.

Preparation Agents

These agents receive Web pages from the Search subsystem and perform the validation and pre-processing tasks described earlier. They are created by the Supervisor Agent and deleted by this agent when they are not being used any more.

Extractor Agent

These agents perform the classification and information extraction tasks described before over the Web pages received from Preparation Agents. The results of this treatment (extracted information and classification of Web pages) are transmitted to the Storage Agent.

Recommendation Agent

This agent receives prepared pages from the Preparation Agent and dispatches them to agents from the same domain (internal recommendation) or from other domains (External recommendation).

Storage Agent

The Storage Agent is in charge of storing the extracted/classified information in the database of the Front Office Subsystem to be exploited by the users.

The test of the AGATHE architecture was done reusing the same domain of the MASTER-Web system, restricted to scientific events in academic research.

In the first test, on a sample of 310 pages obtained from the Web by a search engine, the AGATHE system correctly classified 280 pages, promising 90,32%.

In the next section, we will discuss and compare some works related to the here presented systems.

Related work

This architecture involves a variety of fields: retrieval of information, extraction, classification, multi-agent systems, natural language processing and ontologies, among others. However, we will concentrate on some proposed solutions similar to ours: *WebKB* (CRAVEN et al., 1998), *CiteSeer* (BOLLACKER et al., 1999), *DEADLINER* (KRUGER et al., 2000) and *Ontoseek* (GUARINO et al., 1999). These systems will be compared with the systems here presented (MASTER-Web and AGATHE) and support tools for semantic annotation will be introduced, among them the KIM platform that consists in the idea of linking the identity of entities to their semantic descriptions, e.g. provide information of metadata to the instances mentioned in the text (POPOV et al., 2004).

WebKB: Learning and ontology-bases classification and extraction

The WebKB system (CRAVEN et al., 1998) learns automatically rules for integrated categorization and extraction of Web pages, employing domain ontology with classes and relations. The Web pages are represented with title, key words, frequencies and links.

The system employs a domain ontology with only four entities: activities (subdivided into projects and courses, persons (subdivided into students, professors, member of the academic body) and departments. The ontology also includes relations such as: course instructors, Project members and advisors, among others.

The WebKB corresponds more or less to the works of three future agents of the MASTER-Web or AGATHE systems – the agents for researchers, projects and organizations. On the other hand, these systems aim at treating the scientific domain based on Web search, while WebKB is processing sites of universities.

The MASTER-Web is ontologically richer for approaching the research area as a whole and with more complex relations, and inference capacity already during classification and extraction. Therefore this system requires greater efforts in the creation of agents for each class of Web pages. The WebKB has the advantage of quick adaptation to new domains and utilizes statistical heuristics of connection patterns between pages and key words (expressions are not processed) while the MASTER-Web is based on key words and expressions associated with concepts contained in links for suggesting them to other agents processing classes of Web pages semantically related to the referred concepts.

The authors of the WebKB evaluate the classification performance only in terms of false positives, reporting percentages ranging between 73% and 38% except for the classes Member of the Staff and Others (rejected). However, when counting the false negatives, the class "others" shows a good performance (93.6%), the class "student" follows with 43% and the other six classes show accuracy less than 27%, reducing the mean accuracy to only about 50%. This leads to the hypothesis that the ontology employed in the WebKB was not

comprehensive enough. The ontology of Science in the MASTER-Web on the other hand possesses classes, such as projects and products, which were not used for two reasons: The agents need these concepts for their tasks and future agents treating these classes can be elaborated. On the other hand, an ontology with too much classes can be difficult to be learned and in this case more agents would be necessary.

The *CiteSeer* e *DEADLINER* systems

These systems are efficient in retrieving, filtering abs extracting information from the Web utilizing statistical and learning methods combined with *a priori* knowledge.

The *CiteSeer* (BOLLACKER et al., 1999) is the information agent most used in the retrieval of scientific publications. The system monitors newsgroups, editors and search mechanisms on the basis of the terms *publications, papers and postscript.* Bibliographic data are extracted from the article and the bibliography, which acts like a list helping to find other articles. The number of times an article is cited in other articles is a measure of its relevance. Databases of authors and academic journals as well as complex techniques are applied for identifying co-references of authors and articles.

The DEADLINER (KRUGER et al., 2000) searches for announcements of conferences, extracting: date of beginning and end and deadlines, committee, affiliation of the members of the committee, name of the event and country. The accuracy of the *DEADLINER* is over 95%, however its definition of event is more restricted: all attributes except country must be present in addition to submission data. The MASTER-Web offers greater flexibility and coverage, accepting announcement of book chapters, journals, periodicals and civil service exams. The requisites are in cases, which are more flexible.

OntoSeek: a Web-based information retrieval system

OntoSeek (GUARINO et al., 1999) is an information retrieval system designed for content-based information retrieval from online yellow pages and product catalogs. It combines an ontology-driven content-matching mechanism with a moderately expressive representation formalism. The system also utilizes terms in natural language in order to obtain more precise resource descriptions besides possessing full terminological query flexibility due to a process of semantic matching between queries and resource descriptions.

This system utilizes the the *WordNet*, a linguistic database formed by *synsets*—terms grouped into semantic equivalence sets, each one assigned to a lexical category (noun, verb, adverb, adjective). Each synset represents a particular sense of an English word and is usually expressed as a unique combination of synonymous words.

In general, each word is associated to more than one synset and more than one lexical category. Thus, for sense disambiguation for a given word, the OntoSeek works interactively with the lexical interface of the WordNet, allowing for selecting the appropriate synset and category. Various kinds of semantic relations are maintained among synsets, which are fundamental for the disambiguation process. If we, for example, want to search for cars with radios, the descriptions that should be selected are only those, in which the concepts "radio" and "car" appear in connection with the relation "part of", thus eliminating the stores that sell radios and cars, for example.

KIM: Knowledge and Information Management

The KIM platform (POPOV et al., 2004) provides infrastructure for knowledge and information management, automatic semantic annotation, indexation and document retrieval services based on semantic restrictions and finally query and modification of the ontologies and of the knowledge base.

The platform combines information extraction based on another mature text engineering platform, GATE (General Architecture for Text Engineering), which is a comprehensive platform for natural language processing and extraction of information, developed by the University of Sheffield, United Kingdom. This platform has been constantly developed since 1995 and is used in a variety of research projects (MAYNARD et al., 2000).

One of the strong points of the KIM platform is the automatic annotation of Named-Entities (NE) – real-world entities referenced by their name, such as: *Person, Organization, Company, Localities, etc.*- with references of classes and instances pointing to a semantic repository. Through these entities a knowledge base with vast coverage of entities of the real world is maintained, used and continuously enriched, facilitating the interpretation of names.

This platform utilizes two knowledge repositories for performing its tasks: KIM Ontology (KIMO) and a knowledge base.

More specifically, KIMO is an upper- ontology consisting of about 250 classes and 100 properties and relations. It starts with some basic philosophic distinctions between types of entities, such as: *Objects* – real-world entities like localities and agents; *Events* – defining events and situations, and *Abstractions*, that are neither objects nor events.

It is also noteworthy that the KIM KB is pre-populated with entity descriptions (more than 80.000) and relations between these entities that allow for enough clues for the information extraction process to perform well on inter-domain Web content.

Discussion

The facts here presented reveal a promising tool, which in all phases of its evolution distinguished itself positively from other tools with similar purposes such as WebKB (CRAVEN et al., 1998), Citeseer (BOLLACKER

et al., 1999), DeadLiner (KRUGER et al., 2000), KIM (POPOV et al., 2004), and Ontoseek (GUARINO et al., 1999).

The great and innovating Idea of the MASTER-Web/AGATHE systems resides in their reusable architecture and portability between domains, like evidenced in the tests using the academic and artificial intelligence domains.

In similar systems the knowledge is hidden inside the algorithms, neither allowing for sharing the knowledge nor for specifying contexts, in which they could be useful. These approaches require the creation of a new system without great possibilities of reuse for processing a new class (FREITAS et al., 2003).

Final remarks and future works

The problem related to the manipulation of information in the Internet or in great digital libraries poses a great challenge and demands for solutions facilitating effective access to the huge amount of available information for the user. This work was based on the premise that knowledge-based systems represent a more flexible and promising alternative than the traditional procedure-based approaches. The advent of the ontologies in particular allowed structuring the knowledge necessary for performing the proposed tasks. In the elaboration of the two systems here presented, we sought to improve the scalability of our solution progressively and made experiments demonstrating the feasibility of exploiting new domains utilizing the same solution.

The present work and its prototypes can still be extended in different directions. Retrieval, extraction and classification of information should involve the use of technologies able to learn to recognize information patterns for an accurate retrieval of the necessary information. Basically, we intent to employ these techniques for two reasons: they accelerate the knowledge acquisition necessary for performing the tasks (e.g. extraction of exchange rates, learning the retrieval and extraction patterns from the Web pages) and also for improving the performance of the tasks.

The ability to recognize linguistic patterns present in or proximate to information items to be extracted allows a good extraction performance even in the case of unstructured texts like in the Web.

The integration of linguistic tools, knowledge bases and upper ontologies and/or specific domains points to an expressive improvement in the performance of information manipulation tasks, particularly as refers to extraction and classification. There is an enormous range of resources available for the English language. Among the most popular resources we can mention the linguistic ontology Wordnet (MILLER, 1995) and the tools and resources of the GATE architecture.

We still intend to adapt our solutions to the knowledge representation languages used in the semantic Web, such as the Ontology Web Language (OWL) (HERMAN, 2007) and the *Semantic Web Rule Language* (HORROCKS et al., 2006).

Acknowledgements

To *Scopus Tecnologia* for supporting this research.

Bibliographic references

BENJAMINS, V.R. et al. Knowledge Management through Ontologies. In: INTERNATIONAL CONFERENCE ON PRACTICAL ASPECTS OF KNOWLEDGE MANAGEMENT, 2., 1998, Basel. **Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management**, Basel: Switzerland, 1998, p.1-5.

BOLLACKER, S. et al. CiteSeer: an autonomous web agent for automatic retrieval and identification of interesting publications. In: INTERNATIONAL ACM CONFERENCE ON AUTONOMOUS AGENTS, 2., 1998, **Proceedings of the 2nd International ACM Conference on Autonomous Agents**, USA, 1998.

ESPINASSE, B. et al. AGATHE: an agent and ontology based system for restricted-domain information gathering on the web. In: INTERNATIONAL CONFERENCE ON RESEARCH CHALLENGES IN INFORMATION SCIENCES (IEEE-RCIS), April 23-26, **Proceedings of the International Conference on Research Challenges in Information Sciences (IEEE-RCIS), April 23-26**, Ouarzazate, Morocco, 2007.

FREITAS, F. et al. An ontology-based architecture for cooperative information agents. In: INTERNACIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE – IJCAI, 2003, Acapulco. **Proceedings of the Internacional Joint Conference on Artificial Intelligence – IJCAI'2003**, Acapulco, México.

FREITAS, F. **Ontology of science.** Available at: <http://protege.stanford.edu/plugins/ontologyOfScience/ontology_of_science.htm>. Accessed: 2001.

FREITAS, F. Sistemas multiagentes cognitivos para a recuperação, classificação e extração integradas de informação da web. 2002. Tese (Doutorado em Engenharia Elétrica) - Universidade Federal de Santa Catarina, Florianópolis.

FREITAS, F.L.G. Ontologias e a web semântica. In: CONGRESO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003. Campinas: SBC, 2003.

GOOGLE, Google Search Engine. Available at: <http://www.google.com>. Accessed: 2008

GUARINO, N. et al. OntoSeek: content-based access to the web. **IEEE Intelligent Systems**, v.14, n.3, p.70-80, May-Jun. 1999.

HERMAN, I. Web ontology language (OWL). Available at: <www.w3.org/2004/OWL/>. Accessed: 2007.

HORROCKS, I. et al. SWRL: a semantic web rule language combining owl and ruleml: w3c member submission. Available at: <http://www.w3.org/Submission/2004/SUBM -SWRL-20040521/>. Accessed: April 2006.

MAYNARD, D. et al. **A survey of uses of gate**. Technical report cs–00–06. Department of Computer Science, University of Sheffield, 2000.

MILLER, G.A. Wordnet: a lexical database for English. **Communications of the ACM**, v.2, n.11, p.39-41, Nov. 1995.

PALMEIRA, E. et al. Detailed ontologies and text classification: a promising union. In: WORKSHOP ON BUILDING APPLICATIONS WITH ONTOLOGIES FOR THE SEMANTIC WEB, 1., 2006. ENCONTRO PORTUGUÊS DE INTELIGÊNCIA ARTIFICIAL, 12., 2006, Porto, **Proceedings of the Workshop on Building Applications with Ontologies for the Semantic Web**, Lisboa: APPIA, 2006.

POPOV, B. et al. KIM: a semantic platform for information extraction and retrieval. **Journal of Natural Language Engineering**, v.10, n.3-4, p.375-392, Sep. 2004.

PROTÉGÉ. Available at: < http://Protégé.stanford.edu>. Accessed: 2008.

RIBEIRO-NETO, B. et al. **Modern information retrieval**. Addison Wesley: ACM Press, 1999.

TILAB, JADE tutorial. Available at: < http://jade.tilab.com>. Accessed: 2008.

WIEDERHOLD, G. Mediation in information systems: research directions in software engineering. **ACM Computing Surveys,** v.27, n.2, p.265-267, June 1995.

About the authors

Fred Freitas

The author holds a PhD (2002) in Electric Engineering from the Department of Automation and Systems of the Federal University of Santa Catarina, Brazil. Since 2005 and up to the present he is an associate professor in the undergraduate courses in Computer Science and Engineering of the Informatics Center of the Federal University of Pernambuco and member of the graduation body in Computer Science. His fields of interest are ontologies, artificial intelligence, mediators, multi-agent systems, intelligent agents, and information retrieval and classification.

Luciano Cabral

The author is currently concluding his Masters degree in Computer Science in the Informatics Center of the Federal University of Pernambuco (2008) and has a bachelor's degree in Information Systems/Software Engineering by the Integrated Faculty of Recife (2006). Currently he is a virtual tutor of the Federal Rural University of Pernambuco and an Analyst Trainee at *Scorpus Tecnologia*. He has experience in the field of Computer Science with emphasis on Artificial Intelligence. His main areas of interest are ontologies, text mining, KDT, machine learning, extraction/classification of knowledge and information.