Semantic Information Generation from Classification and Information Extraction

Tércio de Morais Sampaio Silva[1], Frederico Luiz Gonçalves de Freitas[2], Rafael Cobra Teske[1], and Guilherme Bittencourt[1]

[1] Universidade Federal de Santa Catarina - Florianópolis - SC - Brasil
{tercio|cobra|gb}@das.ufsc.br
[2] Universidade Federal de Alagoas - Maceió - AL - Brasil
fred.freitas@mail.tci.ufal.br

1 Introduction

This paper presents MASTERWeb, a multi-agent system for classification and information extraction from Web pages. The multi-agent approach allows that agents, specialized in the different page classes of a cluster, share common information through a cooperation process. The goal of the system is to provide the user with information that is less noisy and more focused in his interests. To represent the domain knowledge, the system uses ontologies and frames [4]. The extraction module explores implicit structures of the page class to extract the information efficiently. It consists of an expert system in which the knowledge is stored using ontologies. MASTERWeb is a cognitive multi-agent system for integrated manipulation of information where each agent has the responsibility for the classification of the page contents inside a knowledge domain [2]. The MASTERWeb system is based on the principle that some page classes may be interrelated, for instance, instances of the page class "scientific events" may contain information or links to "researchers" page class through the attribute "chairman of the event".

The web pages are treated according to two views: content view and functional view. The content view allows to discriminate page classes through particular characteristics, such as keywords and structural similarity. Moreover, the content view allows to collect all pages that potentially belong to the processed class, guaranteeing the covering of the domain. These pages are collected through search engines such as Google, Yahoo and Altavista, using predefined keywords. The functional view allows to discriminate web pages according to their role in the linking between pages and in the presentation and storage of relevant data. The possible roles for a page class are: (i) content-pages; (ii) auxiliary-pages; (iii) content-page lists; (iv) Recommendations (content-pages that belong to another class and may be used in the cooperation process among the agents); and (v) garbage-pages.

There are two kinds of knowledge in the MASTERWeb system: operational knowledge and declarative knowledge. The operational knowledge is represented by production rules that determine how each agent should behave inside the multi-agent society and during the page treatment. On the other hand, ontologies are used to represent the declarative knowledge about the syntactic and semantic structure of the information.

N. Koch, P. Fraternali, and M. Wirsing (Eds.): ICWE 2004, LNCS 3140, pp. 573–574, 2004.

2 Case Study and Results

The proposed multi-agent architecture was used to extract information from pages in the domain of scientific events. The knowledge about the information to be extracted was included as instances of the adequate classes of the science ontology. In the development of ontologies, the Protégé-2000 [5] tool was employed. The inference engine used in the system is Jess (Java Expert System Shell) [3]. The use of the JessTab plug-in [1] allowed the manipulation of the ontologies developed in the Protégé-2000 system by the Jess rules.

The system performance was evaluated by experiments. A corpus composed by 148 web pages about scientific events, such as workshops and conferences was used. The system tried to extract information about: event location, deadline, and subject area. Table 1 shows the results. The system effectiveness is computed as percentage of information correctly extracted.

Table 1. Results

Event Place	Deadline	Subject Areas	Total
74.07%	71.43%	61.4%	67,06%

3 Conclusion and Future Work

We presented MASTERWeb, a multiagent system to classify and extract information from Web pages in a specific domain. Presently, we are developing more extraction rules and making more extensive tests with the system, directly using the Web as page source. We are also studying the application of machine learning techniques in order to improve the knowledge acquisition process.

Future work includes application of the system in different domains, such as the traveling and tourism, e-government and e-education. We also intend to apply the system as a support to build a tool able to automatically create Web pages with semantic tags in XML using the simple HTML versions of the pages as input, in the framework of the Semantic Web.

References

1. Henrik Eriksson. Jesstab plugin for protégé. Dept. of Computer and Information Science, Linköping University. http://www.ida.liu.se/ her/JessTab, 2000.
2. Frederico Freitas and G. Bittencourt. An ontology-based architecture for cooperative information agents. In *Proceedings of International Joint Conferences on Artificial Intelligence 2003 – IJICAI'03*, Alacapuco, Mexico, August 2003.
3. Ernest J. Friedman-Hill. *Jess, The Rule Engine for the Java Platform*. Sandia National Laboratories, Livermore, CA, distributed computing systems edition, September 2003.
4. Marvin Minsky. A framework for representing knowledge. In *Psicology of Computer Vision*, pages 211–281. McGraw-Hill, 1975.
5. N. F. Noy, R. Fergerson, and M. Musen. The knowledge model of protege-2000: Combining interoperability and flexibility, 2000.