Knowledge Channels*

Bringing the Knowledge on the Web to Software Agents

J.L. Arjona, R. Corchuelo, D. Ruiz, J. Peña, and M. Toro

Escuela Técnica Superior de Ingeniería Informática de la Universidad de Sevilla
Departamento de Lenguajes y Sistemas Informáticos
Avda. de la Reina Mercedes, s/n, Sevilla (SPAIN)
{arjona}@lsi.us.es

Abstract. In this paper, we present a new framework to extract knowledge from today's non-semantic web. It associates semantics with the information extracted, which improves agent interoperability; it can also deal with changes to the structure of a web page, which improves adaptability; furthermore, it achieves to delegate the knowledge extraction procedure to specialist agents, easing software development and promoting software reuse and maintainability.

Keywords: *knowledge extraction, wrappers, and ontologies.*

1 Introduction

In recent years, the web has consolidated as one of the most important knowledge repositories. Furthermore, the technology has evolved to a point in which sophisticated new generation web agents proliferate. A major challenge for them has become sifting through an unwieldy amount of data to extract meaningful information. This process is difficult because of the following reasons: first, the information on the web is mostly available in human-readable forms that lack formalised semantics that would help agents use it [1]; second, the information sources are likely to change their structure, which usually has an impact on their presentation but not on their semantics [2, 11, 14].

Our proposal provides agent developers with a framework in which they can have access to semantically-meaningful data that resides on heterogeneous, user-friendly web pages. It relies on using a number of agents [16] that we call *knowledge channels*, or KCs for short. KC's allow to separate the extraction of knowledge from the logic of an agent, and they are able to react to knowledge inquiries (reactivity) from other agents (social ability), and act in the background (autonomy) to maintain a local knowledge base (KB) with knowledge extracted from a web site (proactivity). In order to allow for semantic interoperability, the knowledge they manage references a number of concepts in a given application domain that are described by means of ontologies [4].

* The work reported in this article was supported by the Spanish Inter-ministerial Commission on Science and Technology under grants TIC2000-1106-C02-01 and FIT-150100-2001-78

2 Related work

Several authors have worked on techniques for extracting information from today's non-semantic web, and inductive wrappers are amongst the most popular ones [3, 9, 10, 12]. They are components that use automated learning techniques to extract information from similar pages automatically. Although induction wrappers are suited to extract information from the web, they do not associate semantics with the data extracted, this being their major drawback.

The Web-KB (World Wide Knowledge Base) [6] project at CMU aims to develop a probabilistic, symbolic knowledge base that mirrors the content of the web. It uses several machine learning algorithms for this task. Adding these algorithms to logic that a web agent encapsulates, can produce tangled code and does not achieve a clear separation of concerns. Furthermore, the information that resides in a web page may change unexpectedly; this changes may invalidate the knowledge stored in the KB.

Our solution builds on the best of current inductive wrappers, and extends them with techniques that allow us to deal with web knowledge. Using inductive wrappers allows us take advantage of all the work developed in this arena, as boosted techniques or verification algorithms [10, 13] that detect if there are changes in the layout of a web page that invalidate the wrapper.

3 Knowledge Channels

Figure 1 sketches the architecture of our proposal. KCs are core agents responsible for managing local knowledge base (KB). Knowledge is extracted from a web site using semantic wrappers. Before storing the knowledge in the KB, it is verified using semantic verification to check the existing relations amongst the different concepts that give semantics to the information extracted. Thus, a KC can answer inquiries from other agents that need some knowledge to accomplish its goals.

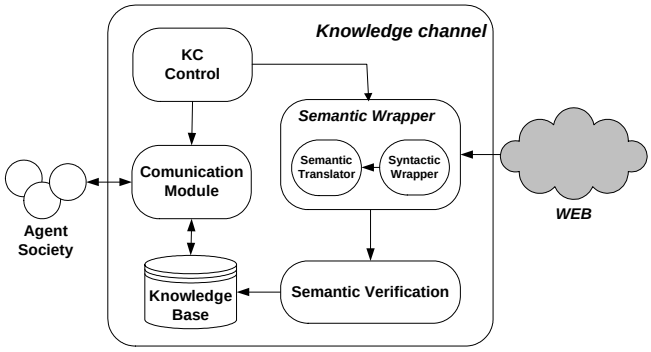

Fig. 1. A Knowledge channel

There are many formalisms to dealt with knowledge, namely: semantic networks, frames systems, logic, decision trees, and so on. KCs use the DARPA Agent Markup Language (DAML+OIL) [15], and FaCT [8] provides[1] reasoning services on the TBox (assertions on concepts) and the ABox (assertions on individuals) that represent the extracted knowledge.

We use ACL [7] as a transport language to send messages amongst KC and user software agents in the agent society. The content of the messages describes how an agent wants to interact with another, and it is written in DAML+OIL based on an ontology that defines communication [5].

3.1 Semantic Wrappers

A semantic wrapper takes a web page as input, and returns a set of instances of concepts defined in an ontology that represents the information of interest. It is composed of an inductive wrapper and a semantic translator. In order to extract knowledge from the web, it is necessary to feed the semantic wrapper with the web page that contains the information. The inductive wrapper extracts the structured information from that web page, and the semantic translator assigns then meaning to it by means of an ontology.

Using inductive wrappers allows us to apply syntactic verifiers to the information extracted. They are algorithms that aims at decide if the wrapper works correctly, or on the contrary, is it invalid because of changes in the web page structure. For instance, the algorithm `Rapture` [10] defined by Kushmerick uses statistical features, such as length, number of words, number of special characters etc. to characterize the extracted data. It learns the parameters of normal distributions describing the feature distributions of the extracted data. This information helps to decide if the Wrapper is valid by means of analysing the statistical values of the information extracted.

The semantic translator needs the user to specify a semantic description that maps the information to be extracted with assertions on individuals defined in the TBox to perform this task.

4 Conclusions

The current web is mostly user–oriented. The semantic web shall help extract information with well–defined semantics, regardless of the way it is rendered, but it does not seem it is going to be adopted in the immediate future, which argues for another solution to the problem in the meanwhile.

In this article, we have presented a new framework to knowledge extraction from web sites based on semantic wrappers. It is based on specialised knowledge channels agents that extract information from the web. It improves on other proposals in that it associates semantics with the extracted information, and can also deal with changes because the information is extracted by means of current

[1] A simple translation of DAML+OIL into SHIQ is previously required.

wrappers. Furthermore, our proposal achieves a separation of the knowledge extraction procedure from the base logic that web agents encapsulate, thus easing both development and maintenance.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic Web. *Scientific American*, 284(5):34–43, May 2001.
2. B.E. Brewington and G. Cybenko. Keeping up with the changing web. *Computer*, 33(5):52–58, May 2000.
3. W.W. Cohen and L.S. Jensen. A structured wrapper induction system for extracting information from semi-structured documents. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (IJCAI-2001)*, 2001.
4. O. Corcho and A. Gómez-Pérez. A road map on ontology specification languages. In *Workshop on Applications of Ontologies and Problem solving methods. 14th European Conference on Artificial Intelligence (ECAI'00)*, 2000.
5. S. Cranefield and M. Purvis. Generating ontology-specific content languages. In *Proceedings of Ontologies in Agent Systems Workshop (Agents'01)*, pages 29–35, 2000.
6. Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew K. McCallum, Tom M. Mitchell, Kamal Nigam, and Seán Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118(1/2):69–113, 2000.
7. FIPA. FIPA specifications. Technical Report http://www.fipa.org/specifications, The Foundation for Intelligent Physical Agents, 2000.
8. Ian Horrocks, Ulrike Sattler, and Stephan Tobies. Practical reasoning for expressive description logics. In Harald Ganzinger, David McAllester, and Andrei Voronkov, editors, *Proceedings of the 6th International Conference on Logic for Programming and Automated Reasoning (LPAR'99)*, number 1705, pages 161–180. Springer-Verlag, 1999.
9. C.A. Knoblock, K. Lerman, S. Minton, and I. Muslea. Accurately and reliably extracting data from the web: A machine learning approach. *IEEE Data Engineering Bulletin*, 23(4):33–41, 2000.
10. N. Kushmerick. Wrapper verification. *World Wide Web Journal*, 3(2):79–94, 2000.
11. L. Lim, M. Wang, S. Padmanabhan, J.S. Vitter, and R. Agarwal. Characterizing web document change. *Lecture Notes in Computer Science*, 2118:133–144, 2001.
12. Ling Liu, Calton Pu, and Wei Han. XWRAP: An XML-enabled wrapper construction system for web information sources. In *ICDE*, pages 611–621, 2000.
13. I. Muslea, S. Minton, and C. Knoblock. STALKER: Learning extraction rules for semistructured, web–based information sources. In *Proceedings of the AAAI-98 Workshop on AI and Information Integration*, 1998.
14. B. Starr, M.S. Ackerman, and M. Pazzani. Do-I-Care: Tell me what's changed on the Web. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access Technical Papers*, March 1996.
15. Frank van Harmelen, Peter F. Patel-Schneider, and Ian Horrocks. Reference description of the DAML+OIL (March 2001) ontology markup language. Technical report, W3C, March 2001.
16. M.J. Wooldridge and M.R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.