INFORMATION SCIENCE PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com
IGR5198

This paper appears in the publication,
Handbook of Research on Global Information Technology Management in the Digital Economy
edited by Mahesh S. Raisinghani © 2008, IGI Global

Chapter XXII
Natural Language Processing Agents and Document Clustering in Knowledge Management:
The Semantic Web Case

Steve Legrand
University of Jyväskylä, Finland

JRG Pulido
University of Colima, Mexico

ABSTRACT

While HTML provides the Web with a standard format for information presentation, XML has been made a standard for information structuring on the Web. The mission of the Semantic Web now is to provide meaning to the Web. Apart from building on the existing Web technologies, we need other tools from other areas of science to do that. This chapter shows how natural language processing methods and technologies, together with ontologies and a neural algorithm, can be used to help in the task of adding meaning to the Web, thus making the Web a better platform for knowledge management in general.

INTRODUCTION

At the end of 2005, the number of Internet users worldwide passed the 1 billion mark and is increasing steadily (de Argaez, 2006). The number of Web pages indexed, by the best search engines, is estimated to be 6 to 8 billion (Patterson, 2005), although both Yahoo and Google have, at times, claimed figures two or three times higher. No one seems to know the exact figure. The total

number of Web pages, whether indexed or not, is estimated to be even higher, over 200 billion by Patterson (2005).

The Web in its current form provides a vast source of information accessible to computers but understandable only by humans. The research community has been looking at ways of making this information understandable to computers as well. In the Semantic Web (SW), different information structuring formats are used to make the information available for automatic machine processing. Ontologies and artificial learning techniques can help in this task of making the Web's full potential available. By now there are a great many ontology repositories with manually created ontologies, which can be used in various SW tasks, but due to the task-specificity, changing nature, and concomitant variety of ontologies that are required for these tasks, these repositories can only partially satisfy the demand.

Two basic avenues for solving the problem of knowledge acquisition bottleneck and construction and integration of various ontologies can be distinguished: ontology learning and ontology integration. These are not usually separate issues: the former plays the primary and the latter secondary role in this chapter. The need for semi-automatic ontology engineering and learning techniques in both approaches is dictated by time and money constraints. Ontologies can be learned from many different sources including free text, tagged documents, and databases, and by many different methods. To select a suitable ontology, it is useful to have some guidelines and comparative knowledge about the existing ontology learning applications. Instead of doing this comparative analysis here, we refer the reader to one of the state-of-art descriptions of ontology learning (Shamsfard & Barforoush, 2003) and merely present one of the many ontology learning applications, Text-To-Onto (Maedche & Staab, 2001), to exemplify the use of the comparison framework.

One thing that is often mentioned and just as often forgotten when we talk about the Semantic Web: people working on the SW-related research agree that manual document structuring (XML-related technologies) and semantic tagging (technologies such as RDF and OWL, which are based on XML) are too complex and time consuming to ever become a popular pastime for your average Web user. One cannot force people to use something which they know very little of and which may be hard to learn. For this reason alone, we need to automate the Semantic Web and thus make the machines there understand each other and rid the human being from the boring details. This is, of course, easier said than done—we should never forget that human beings use natural language, and that most of the Web documents are only partially structured and written in natural language. Apart from creating standards based on XML, Web services, and so forth, we also need automated methods to understand and manipulate documents that are not tagged, that use natural language. The time is ripe to unify the research on natural language processing and on the Semantic Web, and reap the benefits that this can provide.

We believe that we will be seeing, in the near future, a gradual inclusion of NLP subsystems into the service of SW agents to form the backbone for new knowledge discovery systems. Many of the various SW technologies based on ontology-related annotations and utilizing various markup language standards have already established themselves as path setters for the future and can justify their place in the grand scheme of the Semantic Web. To discover new knowledge, however, a specialized agent network, capable of cooperation and able to process natural language, is needed to process raw text and refine it, through annotations, and make it digestible for other more generalized agents. This chapter, rather than trying to impose an artificial framework for such an agent network, instead reflects on the current developments in the area, pointing out some converging trends, highlighting the possibilities that new developments in the Semantic Web can offer, and drafting a possible information flow within the NLP agent framework gradually taking shape. The chapter emphasizes the fact that many of the natural language processing methods and technologies are readily adaptable for agent use, and what is needed, apart from new technological developments, is even deeper cooperation between the artificial intelligence com-