What follows is a summary of the chapter and includes only the Introduction and Bibliography

Analysis and Evaluation of Techniques for the Extraction of Classes in the Ontology Learning Process

Rafael Pedraza-Jimenez, Mari Vallez, Lluís Codina, and Cristòfol Rovira

Department of Communication, Pompeu Fabra University
Campus de la Comunicació, Roc Boronat 138
08018 Barcelona, Spain
{rafael.pedraza,mari.vallez,lluis.codina,
cristofol.rovira}@upf.edu

Abstract. This paper analyzes and evaluates, in the context of Ontology learning, some techniques to identify and extract candidate terms to classes of a taxonomy. Besides, this work points out some inconsistencies that may be occurring in the preprocessing of text corpus, and proposes techniques to obtain good terms candidate to classes of a taxonomy.

Keywords: Semantic Web, Ontology engineering, Ontology learning, Text mining, Language processing.

1 Introduction

In 2001 Berners-Lee and his colleagues made known to the public at large the Semantic Web [1], a short, medium and long term project of the most important agency for the Web standardization: the World Wide Web Consortium (W3C). This proposal implied deep changes that would affect, and, in fact are already affecting, the fields of creation, edition and publication of web pages and sites.

The main goal of this project is to make understandable for machines the Web content [2]. However, three requirements would be necessary to make it possible: a) Web contents must be described: to this end different languages have been created, such as RDF [3], which allows the description of any resource on the Web with metadata. b) The different knowledge domains must be structured and formalized using ontologies [4]. c) Tools to interpret, compare, and merge data on a semantic base are needed: these tools work over ontologies, and they can be built using different languages. The most important of them is OWL [5].

Nevertheless, the formalization of Semantic Web [6], on the one hand describing their resources and on the other hand making ontologies, entails a high cost in time and money. As a result, in 2010 the Semantic Web is not yet a reality [7] and, although many of its technologies are already among us [8], the W3C has recently announced that the entire project can not be achieved in less than 10 years.

To solve the first of these problems several research groups, namely, the one that the authors of this paper belong to, DigiDoc (http://www.upf.edu/digidoc), are working in the development of editors and automatic extractors of metadata (such as DigiDocMeta: http://www.metaeditor.net). Regarding the second issue, in 2001 a new discipline developed, the Ontology Engineering [9], devoted to the study and the design of applications that help to develop, maintain and use these tools semi-automatically.

In this new discipline, the process called "Ontology learning" [10] is very important, which focuses on the generation of tools to import, extract, prune, refine and evaluate the taxonomy of an ontology semi-automatically.

This work is carried out in the Ontology learning field, and focuses on the analysis and evaluation of techniques commonly used to propose terms [11] that constitute the classes of the taxonomy resulting from this process.

This paper is structured as follows: the next section explains the ontology learning process; the following section sets out the main objectives of this research; the third section describes the methodology and tools used in experimentation. Then a discussion concerning the main results of this research is presented. Finally, some conclusions are stated.

Acknowledgments

The development of this research is partially supported by projects CSO 2008-02627 and CSO2009-13713-C05-04 of the Spanish Ministry of Science and Innovation.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
2. Codina, L., Marcos, M.C., Pedraza-Jimenez, R. (coords.): Web semántica y sistemas de información documental. Trea (2009)
3. RDF Working Group. Resource Description Framework (RDF) (2004), http://www.w3.org/rdf
4. Pedraza-Jimenez, R., Codina, L., Rovira, C.: Web semántica y ontologías en el procesamiento de la información documental. El Profesional de la Información 16(6), 569–578 (2007)
5. OWL Working Group. OWL Web Ontology Language (2004), http://www.w3.org/2004/owl
6. Codina, L., Rovira, C.: La Web semántica. In: Tramullas, J., (eds). Tendencias en documentación digital, ch. 1. Trea (2006)
7. Pedraza-Jimenez, R., Codina, L., Rovira, C.: Semantic web adoption: online tools for web evaluation and metadata extraction. In: The 8th International FLINS Conference on Computational Intelligence in Decision and Control, Madrid (2008)
8. Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E., Stephens, S.: The Semantic Web in Action. Scientific American 297(6), 90–97 (2007)
9. Maedche, A., Staab, S.: Ontology learning for the Semantic Web. IEEE Intelligent Sys- tems 16(2), 72–79 (2001)
10. Maedche, A., Staab, S.: Ontology learning. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 173–189. Springer, Heidelberg (2003)
11. Buitelaar, P., Cimiando, P., Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications. Frontiers in Artificial Intelligence and Applications Series, vol. 123. IOS Press, Amsterdam (2005)

12. Gómez-Pérez, A., Manzano-Macho, D.: An overview of methods and tools for ontology learning from texts. The Knowledge Engineering Review 9(3), 187–212 (2005)
13. Vallez, M., Pedraza-Jimenez, R.: Natural Language Processing in Textual Information Retrieval and Related Topics. "Hipertext.net", 5 (2007), http://www.hipertext.net/english/pag1025.htm
14. Hotho, A., Staab, S., Stumme, G.: Explaining text clustering results using semantic structures. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 217–228. Springer, Heidelberg (2003)
15. Beckwith, R., Miller, G.A., Tengi, R.: Design and Implementation of the WordNet Lexical Database and Searching Software. Description of WordNet. Technical report (1993)
16. Miller, G.: WordNet: A lexical database for english. Communications of the ACM 38(11) (1995)
17. Chisholm, E., Kolda, T.: New term weighting formulas for the vector space method in information retrieval. Technical report, Oak Ridge National Laboratory (1999)
18. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)

Publication source reference

[N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part III, LNAI 6098, pp. 488–497, 2010.