World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:08, 2015

On Exploring Search Heuristics for improving the efficiency in Web Information Extraction

Authors : Patricia Jiménez, Rafael Corchuelo

Abstract : Nowadays the World Wide Web is the most popular source of information that relies on billions of on-line documents. Web mining is used to crawl through these documents, collect the information of interest and process it by applying data mining tools in order to use the gathered information in the best interest of a business, what enables companies to promote theirs. Unfortunately, it is not easy to extract the information a web site provides automatically when it lacks an API that allows to transform the user-friendly data provided in web documents into a structured format that is machine-readable. Rule-based information extractors are the tools intended to extract the information of interest automatically and offer it in a structured format that allow mining tools to process it. However, the performance of an information extractor strongly depends on the search heuristic employed since bad choices regarding how to learn a rule may easily result in loss of effectiveness and/or efficiency. Improving search heuristics regarding efficiency is of uttermost importance in the field of Web Information Extraction since typical datasets are very large. In this paper, we employ an information extractor based on a classical top-down algorithm that uses the so-called Information Gain heuristic introduced by Quinlan and Cameron-Jones. Unfortunately, the Information Gain relies on some well-known problems so we analyse an intuitive alternative, Termini, that is clearly more efficient; we also analyse other proposals in the literature and conclude that none of them outperforms the previous alternative.

Keywords : information extraction, search heuristics, semi-structured documents, web mining.
Conference Title : ICADMA 2015 : International Conference on Advanced Data Mining and Applications
Conference Location : Barcelona, Spain
Conference Dates : August 17-18, 2015