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

A Method to Evaluate and Compare Web Information Extractors

Authors : Patricia Jiménez, Rafael Corchuelo, Hassan A. Sleiman

Abstract : Web mining is gaining importance at an increasing pace. Currently, there are many complementary research topics under this umbrella. Their common theme is that they all focus on applying knowledge discovery techniques to data that is gathered from the Web. Sometimes, these data are relatively easy to gather, chiefly when it comes from server logs. Unfortunately, there are cases in which the data to be mined is the data that is displayed on a web document. In such cases, it is necessary to apply a pre-processing step to first extract the information of interest from the web documents. Such pre-processing steps are performed using so-called information extractors, which are software components that are typically configured by means of rules that are tailored to extracting the information of interest from a web page and structuring it according to a pre-defined schema. Paramount to getting good mining results is that the technique used to extract the source information is exact, which requires to evaluate and compare the different proposals in the literature from an empirical point of view. According to Google Scholar, about 4 200 papers on information extraction have been published during the last decade. Unfortunately, they were not evaluated within a homogeneous framework, which leads to difficulties to compare them empirically. In this paper, we report on an original information extraction evaluation method. Our contribution is three-fold: a) this is the first attempt to provide an evaluation method for proposals that work on semi-structured documents; the little existing work on this topic focuses on proposals that work on free text, which has little to do with extracting information from semi-structured documents. b) It provides a method that relies on statistically sound tests to support the conclusions drawn; the previous work does not provide clear guidelines or recommend statistically sound tests, but rather a survey that collects many features to take into account as well as related work; c) We provide a novel method to compute the performance measures regarding unsupervised proposals; otherwise they would require the intervention of a user to compute them by using the annotations on the evaluation sets and the information extracted. Our contributions will definitely help researchers in this area make sure that they have advanced the state of the art not only conceptually, but from an empirical point of view; it will also help practitioners make informed decisions on which proposal is the most adequate for a particular problem. This conference is a good forum to discuss on our ideas so that we can spread them to help improve the evaluation of information extraction proposals and gather valuable feedback from other researchers.

Keywords : web information extractors, information extraction evaluation method, Google scholar, web
Conference Title : ICADMA 2015 : International Conference on Advanced Data Mining and Applications
Conference Location : Barcelona, Spain
Conference Dates : August 17-18, 2015