Methods and Techniques for Information Extraction by Text Segmentation

Altigran Soares da Silva and Eli Cortez

Instituto de Computação
Universidade Federal do Amazonas
Manaus, Brazil
{alti,eccv}@icomp.ufam.edu.br

Abstract. The growing use of text files for information exchange, such as HTML pages, XML documents, e-mail, blogs posts, tweets, RSS and SMS messages, has brought many problems related to how properly exploit the information implicitly contained therein. In particular, problems related to Information Extraction from such sources have motivated many studies in various scientific communities in areas such as Databases, Data Mining, Information Retrieval and Artificial Intelligence. In this tutorial, we present recent methods and techniques in the literature to address the problem of Information Extraction by Text Segmentation (IETS), which consists in extracting values of interest organized into semi-structured records (e.g., postal addresses, bibliographic citations, classified ads, etc.), implicitly present in textual sources. We will discuss the most recent major approaches proposed in the literature, with particular emphasis on probabilistic methods.