Abstract

Information Extraction is a process to find a specific and important data from an unstructured document (natural language document) into a structured document. Information Extraction information is a solution that can change the job posting format from unstructured document or semi-structured document into a structured document. The concept is a way to extract information from job posting based on some field labels, such as company, title or position, city, salary, etc. The method used is boosted wrapper induction method that can handle free text to generate rules that can recognize the existence of fields that should be extracted. Evaluation of system performance using precision, recall and F-Measure. Parameters that affect the performance of the system is the number of boosting iterations that will affect the number of rules generated detector, the value of stating the number of lookahead tokens that will be considered as candidates for the prefix and suffix, and the use of wildcards. From the results obtained can be inferred the existence of a wildcard is very influential to increase system performance. And boosting iterations also tend to increase the performance but were highly dependent on the amount of variation generated rule. And for the lookahead parameter, system performance depends on the number prefix or suffix of the detector is always in pairs.

Keyword: *Information Extraction, Wrapper, Wrapper Induction, AdaBoost, Boosted Wrapper Induction*