Abstract

Personal names are among one of the most frequently searched items in web search engines. Extracting information in the form of attributes and values for a particular person enables us to uniquely identify that person on the web. For example, although namesakes share the same name they usually have different date of births or affiliations. Given a set of documents retrieved for a particular person, we propose two stage approach to extract values for a set of attributes for that person. In the first stage we mark all potential attribute strings in a given text. The second stage then attempts to select the attribute values relevant to a person name. We use a named entity recognition tool to mark all occurrences of named entities in a given document. We then use a rule-based tagger to identify the variants of the given person name. Next, we employ a combination of rules and pre-compiled attribute value candidate lists to extract values for a given set of attributes. The candidate value lists are manually created using resources available on the web such as Wikipedia. Finally we select the attribute values by using Support Vector Machine (SVM). Features we use on learning SVM are specific expression, distance from target name, distance from other name, and n-gram. The proposed method is evaluated on the test data collection created for the attribute extraction subtask at the second Web People Search Evaluation Workshop (WePS2). According to the results in the evaluation, the proposed method improved f-measure by 62.3 persent from the top system among the 15 participating systems.