ABSTRACT

Finding identity using names is something that many of us perform every day. A match is easily found if the name searched for, is entered exactly as it is recorded in the database. However, names are often ambiguous, not unique and can easily misspelled. In particular, when various names exist that refer to the same entity, detecting all variants and consolidating them into a single entity is a significant problem. This problem is known as name matching, record linkage or entity resolution problem. Name matching plays a vital and essential role in many applications, from searching, deduplication, financial, law enforcement, bibliographic, etc.

Most existing name matching methods are developed for English language and so they cover the characteristics of this language. Up to this moment, there is no specific one has been designed and implemented for Indonesian names. The purpose of this thesis is to develop Indonesian name matching dataset as a contribution to academic research and to propose suitable feature set by utilizing combination of context of name strings and its permute-winkler score. Machine learning classification algorithms is taken as the method for performing name matching. Based on the experiments, by using tuned Random Forest algorithm and proposed features, there is an improvement of matching performance by approximately 1.7% and it is able to reduce until 70% misclassification result of the state of the arts methods. This improving performance makes the matching system more effective and reduces the risk of misclassified matches.