Abstract

Alaa M. Sobhy

Automatic Extraction of Main Thesis Documents Fields Using Decision Trees

Abstract — Thesis documents are underestimated even though they hold large sets of useful information –as they include most of the research information–, but since they are harder to obtain, researchers were lead to depend on research papers even though they have a size limitation and lack elaboration. A lot of time and effort are invested in research, so having a linkage among researchers based on their work would somehow facilitate solving the research problem process. A major step to tackle this goal is to structure thesis documents by extracting some fields such as title, author and abstract. This paper presents a way to structure a semi-structured thesis documents using decision trees in 4 different ways (Simple, Medium, Complex and using KNIME), they scored an overall accuracy of 99.2%.