Abstract

Alaa M. Sobhy

Automatic Structuring and Classification of Researchers' Theses

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 theses documents by extracting some fields including title, author, year, abstract, table of contents, conclusion and references. This research, as an initial goal, explores a way to structure a semi-structured thesis documents. The first step would be to Select and extract thesis documents common features, and then create an inverted file to represent each thesis document along with its extracted features, and then add a response column to prepare the inverted file to be an input to the machine learning systems. This research uses artificial intelligence algorithms such as decision trees, the decision trees were used in 4 different ways (Simple, Medium, Complex and using KNIME), and compares the results with Support Vector Machine and Neural Networks. Using the output of the machine learning system decision trees, important thesis blocks are extracted, and then preprocessed in order to prepare thesis documents to be classified. The preprocessing includes treating sectors similarly, except for the references, they went through another extraction process where each reference where identified whether it is a paper, book a URL and then defining and extracting the parts of each reference, according to the detection that was done. All of the extracted parts, including preprocessed references, were cleansed and converted to lowercases for classification preparation. Classifying thesis documents will help researchers access the data more easily the classification was done using ACM Computing Classification, and comparing the classification results with WordNet categories. Furthermore, a small search engine was implemented to search within the classified documents, the user s the needed keyword for the topic they want to look for, then they can choose to retrieve either the metadata of the thesis such as author, year and supervisor, the all of the contents of the thesis. The medium decision tree model scored an overall accuracy of 99.2% better than the other machine learning techniques and other types of decision trees. The ACM Classification model scored an accuracy of 93% compared to a supervised model. This research introduces the structuring of theses documents, which is considered a new unexplored dataset, it also, when completed, presents a large platform for researchers and bring them closer through their work. Researchers will be able to solve their problems and gain a deeper knowledge through this model.