KU LEUVEN

Knowledge Discovery on Spreadsheet Data using Semantic Information

Sarah Binta Alam Shoilee

Thesis submitted for the degree of Master of Science in Artificial Intelligence, option Big Data Analytics

Thesis supervisor:
Prof. Dr. Luc De Raedt

Assessors:
Assoc. Prof. Dr. Marc Denecker
Dr. Ir. Samuel Kolb

Mentor:
Dr. Ir. Samuel Kolb

Academic year 2019 – 2020

Abstract

For data processing and maintenance, a spreadsheet is the most widely used tools amongst corporations. Performing in-depth research on spreadsheets remains critical to enabling knowledge discovery as well as improving user efficiency and supporting error-free data processing. Further research in this field is essential, considering the sheer expanse and the importance of tabular data in the data community, despite the number of existing works in the area. This paper proposes an algorithm for tabular constraint learning using semantic header information, focusing on enhancing existing tabular constraint learner (TaCLe). The proposed solution will rediscover lost formulas on a poorly filed spreadsheet while conceptually analysing the header. The proposal of the header interpretation assists to recapture formulas faster with more accuracy than the previous solution, overall improving the computational performance.

Furthermore, this research outlines the usability of header information come along with tabular data, often ignored due to its complex structure. The study further proposes a formula reconstruction application, empirically validating each step of the design decision. This technique will not only be capable of formula rediscovery but also for formula suggestion, error-correction, and auto-completion in a spreadsheet environment when integrated with a right interaction model.