Automatic semantic knowledge extraction from electronic forms

Journal Article

A/Prof Timothy French

Authors: Haolin Wu, Tim French, Wei Liu, Melinda Hodkiewicz
2022-05-23

Publication

Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability

CiteScore

3.200

Safety, Risk, Reliability and Quality 61 out of 165
Scopus provides the journal's CiteScore, which is calculated as the total citations made in the CiteScore year and the prior three years to content published in the CiteScore year and the prior three years, divided by the total number of items published by the journal in the CiteScore year and prior three years.

Source-normalized Impact per Paper (SNIP)

1.022

SNIP calculates a journal's average citations from the SNIP year to items published in the prior three years, and divides the average by the citation potential in the journal's subject area to account for variability between subject categories.

SCImago Journal Rank (SJR)

0.518

Safety, Risk, Reliability and Quality 81 out of 1080
The SJR weighs incoming citations according to the prestige of the publications they come from. Total, weighted citations made in the SCImago year to content published in a journal in the prior three years are divided by the total number of items published by the journal in the prior three years.

Quality Indicators

Peer Reviewed

Relevance to the Centre

Electronic tabular forms are an intuitive way for organisations to collect, present and store structured information for human readers. Forms use features such as fonts, colours and cell positioning to help readers navigate and find information. Millions of forms, typically in Portable Document Format (PDF), are generated by businesses as part of routine operations. Unlike human readers, machines are not able to directly 'understand' the implicit cues contained in the fonts, colours and use of boxes without explicit processing. In this paper, a supervised computer vision model is proposed to decompose the PDF form document into nested microtables. The cells within these microtables are then processed using a customisable rule bank for meaningful table content and semantic relationship extraction. The process is demonstrated on an industry dataset of 37 maintenance procedure documents containing 373 pages and 1016 unique microtables. A web application EMU (Extracting Machine Understandable Semantics from Forms) demonstrates how data captured in tables with different dimensions in procedural forms can be automatically extracted and stored in JavaScript Object Notation (JSON). Identifying and extracting nested tables is a critical fundamental step for future applications to support machine-automated search and extraction of data at scale for both maintenance and other procedural documentation.

DOI: 10.1177/1748006X221098272