Smart Consolidation of Product Information[1]

Maurice van Keulen(1), Dolf Trieschnigg(1,2), Brend Wanders(1)
m.vankeulen@utwente.nl, dtrieschnigg@mydatafactory.com, b.wanders@utwente.nl
(1)University of Twente, Faculty of EEMCS, POBox 217, 7500AE, Enschede
(2)MyDataFactory, Werkhorst 36, 7944 AV Meppel

Maintaining accurate and detailed information is needed in many parts of the production chain: not only do industries have to provide precise information about the produced products, they also have to manage product information involved in the production process, ranging from information about required raw materials to detailed specifications of spare parts used in the production line. Incomplete product information may result for instance in ordering the wrong spare part possibly resulting in a disrupted production line and production loss, or even impacting health and safety. Using inappropriate substances in a food- or pharmaceutical context may result in non-complying end-user products. Gartner estimates that “data quality affects overall labor productivity by as much as a 20%”; for the specific case of product data, these effects might even be larger.

Our goal is to aid support dealing with product data quality issues in a manner that uses human effort and expertise as efficiently as possible by employing advanced machine learning and information retrieval technology. The web provides a wealth of information on products provided in various formats, detail levels, targeted at at a variety of audiences. Semi-automatically locating, extracting and consolidating this information would be a “killer app” for enriching and improving product data quality with a significant impact on production cost and quality.

We present a generic web harvesting infrastructure capable of automatically and robustly harvesting of product data from websites. Easy to use web harvesting solutions exist, for example, import.io. Our technology uses advanced information retrieval and machine learning to achieve more autonomy, less configuration, less website-specificity, and more robustness. The infrastructure Incorporates a probabilistic datalog engine, called JudgeD [1], as a component that adds reasoning power expected to reduce mistakes in harvesting and improve quality of output data. JudgeD is novel in its expressive power for representing and reasoning with dependencies in the uncertainty of data.

During an exploratory pilot, we compared manually cleansed industrial product data to data semi-automatically harvested from 16 websites, ranging from manufacturer information sites to broad e-commerce sites. The majority of desired attribute values could be obtained from the web and could be verified across sources. The pilot clearly shows the cascading effect of uncertain choices and inaccurate data in the processing pipeline. For instance, incorrect seed product identifiers may result in a large number of irrelevant product entities to be retrieved from a web source. Or, using a crude attribute normalization function may result in false positives during entity matching between sources. The pilot demonstrates the need for an architecture which supports feedback on intermediate choices in the processing pipeline.

[1] Wanders, B. and van Keulen, M. and Flokstra, Jan (2016) JudgeD: a probabilistic datalog with dependencies. In: Proceedings of the Workshop on Declarative Learning Based Programming, DeLBP 2016, 13 February 2016, Phoenix, AZ, USA. AAAI Press.

[1] Speaker is Maurice van Keulen