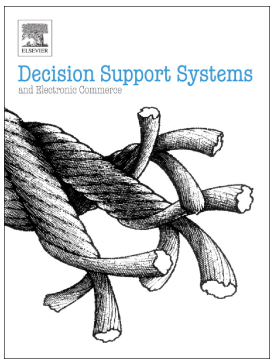

Assessing data quality – A probability-based metric for semantic consistency

Bernd Heinrich, Mathias Klier, Alexander Schiller, Gerit Wagner

PII: S0167-9236(18)30059-9
DOI: doi:10.1016/j.dss.2018.03.011
Reference: DECSUP 12946

To appear in: *Decision Support Systems*

Received date: 7 October 2017
Revised date: 28 March 2018
Accepted date: 28 March 2018

Please cite this article as: Bernd Heinrich, Mathias Klier, Alexander Schiller, Gerit Wagner , Assessing data quality – A probability-based metric for semantic consistency. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Decsup(2017), doi:10.1016/j.dss.2018.03.011

Assessing Data Quality – A Probability-based Metric for Semantic Consistency

Bernd Heinrich[a], Mathias Klier[b,*], Alexander Schiller[c], Gerit Wagner[d]

a Department of Management Information Systems, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany, Bernd.Heinrich@ur.de

b Institute of Technology and Process Management, University of Ulm, Helmholtzstr. 22, 89081 Ulm, Germany, Mathias.Klier@uni-ulm.de

c Department of Management Information Systems, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany, Alexander.Schiller@ur.de

d Department of Management Information Systems, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany, Gerit.Wagner@ur.de

* *Corresponding author. Tel.: +49 731 503-2312*

Abstract:

We present a probability-based metric for semantic consistency using a set of uncertain rules. As opposed to existing metrics for semantic consistency, our metric allows to consider rules that are expected to be fulfilled with specific probabilities. The resulting metric values represent the probability that the assessed dataset is free of internal contradictions with regard to the uncertain rules and thus have a clear interpretation. The theoretical basis for determining the metric values are statistical tests and the concept of the p-value, allowing the interpretation of the metric value as a probability. We demonstrate the practical applicability and effectiveness of the metric in a real-world setting by analyzing a customer dataset of an insurance company. Here, the metric was applied to identify semantic consistency problems in the data and to support decision-making, for instance, when offering individual products to customers.

Keywords: data quality, data quality assessment, data quality metric, data consistency