DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

Cover Letter

Dear Editor,

We submit the paper titled: "*Credit Distribution through Data Provenance in Relational Scientific Databases*" as a regular paper to Information Systems.

This paper explores the problem of Data Credit Distribution, a process by which a numerical value, called credit, is distributed to the elements of the database responsible for the production of data being cited by a research entity. In particular, the paper introduces two methodologies to perform Credit Distribution in curated relational databases, based on two different forms of data provenance: why-provenance and how-provenance.
We describe the proposed methodologies in detail, and we ran extensive evaluations on the real-world database GtoPdb/IUPHAR, using both real queries extracted from papers published in the British Journal of Pharmacology and synthetic ones. These experiments aim to show how credit allows us to reward cited data and its corresponding authors in a way different than traditional citations, considering aspects that are otherwise neglected, and how different distribution strategies behave.

We share the source code, all system settings, and the methodology used to obtain the citations from papers downloaded from Google Scholar in a publicly available repository.

The content of this paper is for a significant part original since it expands the work begun in our previous article "Data credit distribution: A new method to estimate databases impact," published in the journal of Informetrics (2020), where we first introduced the idea of credit distribution and a first method to solve it by using lineage as data provenance. Minor overlapping, although not literal, are found in the abstract, introduction, related work, description of the use case, running example, and in the methodology sections, as we report the definitions of lineage and the strategy based on it. The proposed methodologies and the experiments are entirely new and original. As such, the overlapping is lower than 20% in terms of contents.

We think this paper is a good fit for Information Systems given its innovative and multidisciplinary nature mixing database and data citation techniques to develop new techniques to better reward data authors and curators.

Best regards,

Dennis Dosso, Susan B. Davidson and Gianmaria Silvello

dennis.dosso@unipd.it
susan@seas.upenn.edu
gianmaria.silvello@unipd.it