Improving Attribute Exploration for the Detection and Correction of Anomalies in an Agroecological Knowledge Base

Nassif Saab[1], Marianne Huchard[1] and Pierre Martin[2]
[1] LIRMM, Université de Montpellier, 161 rue Ada, 34095, Montpellier, France
[2] CIRAD, UPR AIDA, Avenue Agropolis, 34398, Montpellier, France

Corresponding author: nassif.saab@lirmm.fr

Data cleaning is crucial to the knowledge discovery process. Knowledge bases such as Knomana [1] rely on data wrangling to standardise and subsequently centralise information extracted from multiple sources. This makes Knomana prone to anomalies, i.e. to incorrect or incomplete descriptions of plant use, which may cause its users to draw wrong conclusions during knowledge discovery. To detect and correct these anomalies, we propose using Attribute Exploration (AE) [2] to acquire expert knowledge and apply it to identify anomalies and correct or complete the descriptions. It is a process of Formal Concept Analysis, which considers data tables describing binary relationships between objects and attributes. AE relies on the computation of the Duquenne-Guigues basis, a complete, consistent and nonredundant set of implication rules, i.e. regularities of the form "if there is X, then there is always Y" [3]. The expert is asked to validate the generated implications or provide a counterexample when an invalid rule is presented. Tools like ConExp [4] implement AE. With Knomana holding 35 attributes covering over 45,000 descriptions of plant use, the number of computed rules is in the thousands [5]. Therefore, it is consequential to have a pertinent and time-saving order of displaying these rules.

To tackle the problem at hand, this poster presents an improvement of AE. During AE, the computed rules are consecutively shown to the expert in the lectic order, where set A is presented before set B if the smallest differing element belongs to B. According to this definition, the lectic order does not consider the nature of the data it is addressing, and consequently, the implications are not displayed in a meaningful order, i.e. an order that regards the expert's interest in a particular type of data. Thereupon, we propose that experts sort the data prior to exploring the attributes. By providing experts with the means to group attributes into categories and order them by relevance, table columns are rearranged in conformity with the definition of the lectic order for the purpose of generating the most relevant implications first. Applying this change to a single data table allowed to accommodate AE to the interests of the expert. As a next step, we plan to extend this technique to relational data to render it applicable to datasets that employ ternary relationships, as is the case in the agroecological knowledge base Knomana.

Acknowledgements

This work is supported by Montpellier University KIM (Key Initiatives MUSE) DATA & LIFE SCIENCES through an interdisciplinary internship grant.

References

[1] Pierre J. Silvie, Pierre Martin, Marianne Huchard, Priscilla Keip, Alain Gutierrez, and Samira Sarter. Prototyping a knowledge-based system to identify botanical extracts for plant health in sub-saharan africa. *Plants*, 10(5), 2021.

[2] Bernhard Ganter and Sergei Obiedkov. *Conceptual Exploration*. Springer Berlin Heidelberg, 2016.

[3] Alexandre Bazin and Jean-Gabriel Ganascia. Computing the Duquenne–Guigues basis: an algorithm for choosing the order. *International Journal of General Systems*, 45(2):57–85, September 2015.

[4] Serhiy A. Yevtushenko. Conexp, 2022.

[5] Lina Mahrach, Alain Gutierrez, Marianne Huchard, Priscilla Keip, Pascal Marnotte, Pierre Silvie, and Pierre Martin. Combining implications and conceptual analysis to learn from a pesticidal plant knowledge base. In *Graph-Based Representation and Reasoning*, pages 57–72. Springer International Publishing, 2021.