Caractérisation et extraction de biclusters de valeurs similaires avec l'analyse de concepts triadiques

Mehdi Kaytoue*, Sergei O. Kuznetsov**,
Amedeo Napoli***, Juraj Macko****, et Wagner Meira Jr.*

*Universidade Federal de Minas Gerais – Belo Horizonte – Brésil
kaytoue@dcc.ufmg.br (contact principal)
**Higher School of Economics – Moscou – Russie
***INRIA Nancy Grand Est/LORIA – Nancy – France
***Palacky University – Olomouc – République Tchèque

Résumé

Le biclustering de données numériques est devenu depuis le début des années 2000 une tâche importante d'analyse de données, particulièrement pour l'étude de données biologiques d'expression de gènes. Un bicluster représente une association forte entre un ensemble d'objets et un ensemble d'attributs dans une table de données numériques. Les biclusters de valeurs similaires peuvent être vus comme des sous-tables maximales de valeurs proches. Seules quelques méthodes se sont penchées sur une extraction complète (i.e. non heuristique), exacte et non redondante de tels motifs, qui reste toujours un problème difficile, tandis qu'aucun cadre théorique fort ne permet leur caractérisation. Dans le présent article, nous introduisons des liens importants avec l'analyse formelle de concepts. Plus particulièrement, nous montrons de manière originale que l'analyse de concepts triadiques (TCA) propose un cadre mathématique intéressant et puissant pour le biclustering de données numériques. De cette manière, les algorithmes existants de la TCA, qui s'appliquent habituellement à des données binaires, peuvent être utilisés (directement ou après quelques modifications) après un prétraitement des données pour l'extraction désirée.

Summary

Biclustering numerical data became a popular data-mining task in the beginning of 2000's, especially for analysing gene expression data. A bicluster reflects a strong association between a subset of objects and a subset of attributes in a numerical object/attribute data-table. So called biclusters of similar values can be thought as maximal sub-tables with close values. Only few methods address a complete, correct and non redundant enumeration of such patterns, which is a well-known intractable problem, while no formal framework exists. In this paper, we introduce important links between biclustering and formal concept analysis. More specifically, we originally show that Triadic Concept Analysis (TCA), provides a nice mathematical framework for biclustering with a better algorithmic scalability over existing methods.