

# 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.

## 1 Introduction

Le biclustering de données numériques a connu un essor considérable depuis le début des années 2000, apportant des réponses aux nouveaux challenges soulevés par l'analyse de données biologiques, et plus particulièrement l'analyse de données d'expression de gènes. A partir d'une table de données objets/attributs, le but du biclustering est de caractériser des sous-tables d'intérêt selon les valeurs qu'elles comportent (voir par exemple l'état de l'art de Madeira et Oliveira (2004)). Ainsi, un bicluster correspond à une paire, composée d'un sous-ensemble d'objets et d'un sous ensemble d'attributs. Une telle paire se représente comme un rectangle dans la table de données, modulo des permutations de colonnes et de lignes. La table 1 est un exemple de telles données numériques : chaque ligne représente un objet ; chaque colonne un attribut ; chaque case la valeur prise par l'objet en ligne pour l'attribut en colonne. Cette table illustre également le bicluster  $(\{g_1, g_2, g_3\}, \{m_1, m_2, m_3\})$  représenté par un rectangle grisé.