Finding fragments of orders and total orders from 0-1 data

Heikki Mannila

HIIT Basic Research Unit, University of Helsinki, Department of Computer Science
and Helsinki University of Technology, Laboratory of Computer and Information Science
Heikki.Mannila@cs.helsinki.fi

High-dimensional collections of 0-1 data occur in many applications. The attributes in such data sets are typically considered to be unordered. However, in many cases there is a natural total or partial order underlying the variables of the data set. Examples of variables for which such orders exist include terms in documents and paleontological sites in fossil data collections. We describe methods for finding fragments of total orders from such data, based on finding frequently occurring patterns. We also discuss techniques for finding good total orderings (seriation) based on spectral ordering and MCMC methods.

Résumé

On s'intéresse aux collections de données 0-1 de haute dimension que l'on rencontre dans de nombreuses applications. Bien que les attributs soient dans de tels ensembles de données typiquement considérés comme non ordonnés, un ordre total ou partiel sous-tend souvent les variables. Par exemple, il existe de tels ordres entre les termes utilisés dans un ensemble de documents, ou les sites paléontologiques dans les collections de données de fossiles. Nous décrivons des méthodes, fondées sur la recherche de motifs fréquents, qui permettent de retrouver des fragments d'ordre total à partir de telles données. Nous discutons également des techniques fondées sur l'ordre spectral et les modèles MCMC qui permettent de trouver de bons ordres totaux (sériations).