20 ans de découverte de motifs : une étude bibliographique quantitative

Arnaud Giacometti, Dominique Haoyuan Li, Arnaud Soulet

Université François Rabelais Tours, LI
3 place Jean Jaurès
F-41029 Blois France
prenom.nom@univ-tours.fr

Résumé. Depuis deux décennies, la découverte de motifs a été l'un des champs de recherche les plus actifs de l'exploration de données. Cet article en établit une étude bibliographique quantitative en nous appuyant sur 1030 publications issues de 5 conférences internationales majeures : KDD, PKDD, PAKDD, ICDM et SDM. Nous avons d'abord mesuré depuis 2005 un sévère ralentissement de l'activité de recherche dédiée à la découverte de motifs. Puis, nous avons quantifié les principales contributions en terme de langages, de contraintes et de représentations condensées de sorte à comprendre ce ralentissement et à esquisser les directions actuelles.

1 Introduction

En 1993, Rakesh Agrawal, Tomasz Imielinski et Arun N. Swami publiaient l'un des articles fondateurs de la découverte de motifs "Mining Association Rules between Sets of Items in Large Databases" dans les actes de "the ACM SIGMOD International Conference on Management of Data" en introduisant le problème de l'extraction des règles d'association intéressantes. Formellement, ce problème consiste à énumérer toutes les règles du type $X \rightarrow I$ où X est un ensemble d'items et I un item non présent dans X telles que les probabilités $P(X \rightarrow I)$ et $P(X \rightarrow I|X)$ soient suffisamment élevées. Moins qu'une finalité nouvelle, Agrawal et al. (1993) ont surtout substitué à la traditionnelle recherche heuristique, la recherche complète et consistante. En effet, le problème de la découverte de règles de classification (où I est seulement une valeur de classe) était déjà une thématique de recherche active dans le domaine de l'intelligence artificielle mais les algorithmes n'étaient pas exhaustifs (Breiman et al., 1984),(Quinlan, 1986),(Piatetsky-Shapiro, 1991). Ces notions de complétude et consistance très importantes en bases de données justifient la publication de Agrawal et al. (1993) dans ACM SIGMOD. De même, la généralisation proposée l'année suivante par Agrawal et Srikant (1994) (où la conclusion de la règle est désormais un ensemble d'items) est publiée dans VLDB [1].

1. D'ailleurs, Rakesh Agrawal revendique son affiliation au domaine des bases de données dans Winslett (2003) :"I'm a database person, so my view of data mining has been that it is essentially a richer form of querying. We want to be able to ask richer questions than we could conveniently ask earlier."