Exploration de traces à l'aide de fouille de données

Peggy Cellier[1], Mireille Ducassé[1], Sébastien Ferré[2]

[1] IRISA, UMR 6074 CNRS, INSA de Rennes
[2] IRISA, UMR 6074 CNRS, Université de Rennes 1
Campus de Beaulieu 35042 Rennes Cedex, France
Prenom.Nom@irisa.fr http://www.irisa.fr/LIS

Il existe plusieurs sortes de traces, par exemple les traces d'exécution, les traces d'interaction, les fichiers de log. Une trace peut être vue comme un ensemble d'événements caractérisant le comportement d'un objet ou d'un individu pendant une tâche. Par exemple, une trace d'exécution d'un programme collecte des informations sur le comportement du programme (lignes exécutées, valeurs des variables, etc.). Une fois les traces collectées il est important de pouvoir les exploiter afin, par exemple, de caractériser et d'expliquer des comportements anormaux ou de définir des profils de tâches.

La fouille de données permet d'extraire automatiquement de l'information "pertinente" dans des masses de données. Elle est utilisée dans de nombreux domaines comme le marketing, la bioinformatique. Il existe plusieurs techniques de fouille comme l'extraction de règles d'association (Agrawal *et al.*, 1993) ou de motifs séquentiels (Agrawal & Srikant, 1995).

Dans nos travaux nous proposons d'extraire de l'information des traces en utilisant la fouille de données afin d'aider un utilisateur à comprendre le comportement de l'objet/individu dont proviennent les traces (Cellier *et al.*, 2011). En effet, la fouille de données permet de faire émerger de la connaissance des données, en mettant en évidence des régularités et des tendances appelées *motifs*. Toutefois, le problème de ce genre d'approches est que trop de motifs sont générés rendant l'exploration à la main impraticable. Pour réduire le nombre de motifs, certaines méthodes proposent des représentations

condensées (Pasquier *et al.*, 1999; Plantevit & Crémilleux, 2009) ou l'utilisation de contraintes (Pei *et al.*, 2001). Cependant, le nombre de motifs reste important. D'autres approches proposent d'ordonner totalement les motifs suivant une mesure (ex. : confiance, lift) et de sélectionner les k meilleurs pour les montrer à un utilisateur. Ce genre d'approche ne prend pas en compte la connaissance de l'utilisateur ni les dépendances qui peuvent exister entre les motifs. Nous proposons d'utiliser l'analyse de concepts logique (Ferré & Ridoux, 2004) pour explorer les motifs extraits à partir de la fouille de données en exploitant l'ordre partiel qui existe entre les motifs. Un avantage de cette approche est qu'aucun élagage des motifs n'est fait à priori, ce qui permet de conserver toute la connaissance extraite en la structurant dans un ordre partiel. L'utilisateur peut ensuite naviguer dans l'espace des motifs en se servant de ses connaissances. Un autre avantage est que l'utilisateur n'a pas à regarder tous les motifs, en effet au cours de son exploration des sous-ensembles de motifs non pertinents peuvent facilement être élagués du fait de l'organisation des motifs dans un ordre partiel. L'approche a notamment été utilisée pour fouiller les traces d'exécution des programmes dans le but de localiser des fautes.

Références

AGRAWAL R., IMIELINSKI T. & SWAMI A. N. (1993). Mining association rules between sets of items in large databases. In P. BUNEMAN & S. JAJODIA, Eds., *Int. Conf. on Management of Data* : ACM Press.

AGRAWAL R. & SRIKANT R. (1995). Mining sequential patterns. In *Int. Conf. on Data Engineering* : IEEE.

CELLIER P., FERRÉ S., DUCASSÉ M. & CHARNOIS T. (2011). Partial orders and logical concept analysis to explore patterns extracted by data mining. In *Int. Conf. on Conceptual Structures*, LNCS : Springer.

FERRÉ S. & RIDOUX O. (2004). An introduction to logical information systems. *Information Processing & Management*, **40**(3), 383–419.

PASQUIER N., BASTIDE Y., TAOUIL R. & LAKHAL L. (1999). Discovering frequent closed itemsets for association rules. In *Int. Conf. on Database Theory*, p. 398–416 : Springer-Verlag.

PEI J., HAN J. & LAKSHMANAN L. V. S. (2001). Mining frequent itemsets with convertible constraints. In *Int. Conf. on Data Engineering* : IEE.

PLANTEVIT M. & CRÉMILLEUX B. (2009). Condensed representation of sequential patterns according to frequency-based measures. In *Int. Symp. on Advances in Intelligent Data Analysis*, LNCS(5772) : Springer.