L'analyse relationnelle pour la fouille de grandes bases de données.

Hamid Benhadda*, François Marcotorchino*

*160, boulevard de Valmy – BP 82
92704 Colombes Cedex
{ hamid.benhadda,jeanfrancois.marcotorchino }@fr.thalesgroup.com

Résumé. Dans cet article nous montrerons, brièvement, les possibilités offertes par la théorie de l'analyse relationnelle, initiée dans les années 1980 à IBM-Corp. Nous nous concentrerons sur les avancées théoriques et méthodologiques obtenues grâce à cette théorie pour fusionner l'information et pour traiter et analyser de grandes quantités de données qu'elles soient de type structuré ou non structuré. Nous aborderons brièvement la théorie de la similarité régularisée, théorie basée sur l'analyse relationnelle et la généralisant mais plus récente. Nous montrerons aussi des formules de transfert permettant d'exprimer des problèmes combinatoires bien connus sous forme de fonctions économiques linéaires appropriées pour différents type de problématique (tels que des problèmes de classification automatique ou des problèmes d'association,). Ceci en plus de la complexité linéaire O(N) de l'algorithmique sous jacente qui permet à cette approche d'être tout à fait convenable pour différentes applications réelles.

1 Introduction

De nos jours plus encore qu'à d'autres époques, les progrès techniques et scientifiques d'une part et les faibles coûts de stockage d'autre part poussent l'homme à rassembler et conserver des quantités de plus en plus grandes de données. Cette accumulation de données, est amplement justifiée par le fait qu'à notre époque, la possession et l'exploitation du maximum d'information confèrent à ceux qui les maîtrisent un avantage concurrentiel majeur.

Il devient donc nécessaire d'avoir à sa disposition des outils d'analyse et d'exploitation de ces données, afin d'en extraire une information à valeur ajoutée qui pourra être utilisée par la suite pour faire de nouveaux progrès dans les domaines particuliers relatifs à ces données.

Ceci ne pourra se réaliser que si les outils concernés respectent la structure des données qu'on leur confie. En particulier, ces outils doivent permettre de manipuler, de combiner et de structurer les variables et attributs d'analyse, en les considérant comme des entités propres et séparées et non comme un magma global qu'on considèrera comme un tout ou en adaptant les données aux méthodes préexistantes, en violant leur nature pour satisfaire les exigences de ces méthodes.

Afin de respecter les exigences qui viennent d'être citées, nous allons parler, dans cet article, d'une part, de l'analyse relationnelle et de ses extensions et d'autre part de la similarité régularisée, théorie qui a été co-développée par les auteurs et qui généralise la théorie de l'analyse relationnelle.