

## *Fouille des données complexes.*

Dirigé par **Omar Boussaïd, Pierre Gançarshi, Florent Massiglia et Brigitte Trousse.**  
Revue des Nouvelles Technologies de l'Information, série E, n°4, Cépadués-Éditions, 2005, 286 p.

L'exploitation des données collectées à partir des différents supports est en passe de devenir un enjeu industriel important. En effet, les objets du monde réel ne se présentent pas tous de façon unimodale : un dossier médical, par exemple, comporte certes des données numériques pouvant être structurées de façon tabulaire comme les résultats d'analyses biochimiques, mais incorpore également des données textuelles comme les comptes-rendus d'observations cliniques, des graphiques tels que les tracés d'électrocardiogramme ou d'électro-encéphalogramme, voire des images complexes fournies par les radiographies, échographies et scanners. De même, les données recélées par les sites web se présentent selon différents modes : tableaux, textes, graphiques, images, voire sons qui ne sont pas forcément redondants du point de vue informationnel. Ainsi dans bien des domaines comme le multimédia, la télédétection, l'imagerie médicale, les systèmes d'information géographique, la bio-informatique, les données traitées pour l'extraction de connaissances regroupent différents formats de codage pour le texte, les graphiques, les images ou les sons voire différentes catégories d'information qu'il s'agisse de descriptions factuelles, de connaissances, voire de véritables ontologies.

Le défi que rente de relever la fouille des données complexes est de pouvoir prendre en compte la totalité des informations disponibles concernant les entités qui constituent les unités taxinomiques élémentaires. Cela suppose d'être capable d'intégrer des informations de nature différente et de pouvoir ainsi les rattacher à une même catégorie sémantique. Sur le plan méthodologique, il s'agit donc de définir une mesure de ressemblance ou de dissemblance entre deux objets dont la description est fournie par un complexe d'informations. Jusqu'à présent, les approches proposées ont surtout été fondées sur la juxtaposition de similarités ou de dissimilarités partielles.

Dans ce contexte, ce numéro 4 de la thématique Extraction et gestion des connaissances dédié à la Fouille de données complexes dresse donc un premier bilan du groupe de travail créé sur l'initiative du professeur Zighed sous l'égide de l'association Extraction et Gestion des Connaissances Certains articles portent sur la qualité, l'hétérogénéité et le caractère multi-sources des données. D'autres s'intéressent à l'image, soit pour l'indexer, soit pour en extraire de l'information, soit pour en pondérer les attributs. Signalons également des contributions portant sur la fouille de documents effectuée soit directement par des méthodes prenant en compte le contexte structurel ou en étudiant l'adéquation des modèles de représentation aux méthodes de catégorisation, soit indirectement en étudiant la sélection de règles issues de la fouille de textes ou en concevant un modèle de cartes topologiques pour données catégorielles, voire en élaborant des méthodes de recherche d'informations inattendues dans des textes. Des contributions plus spécifiques à la gestion des bases de données figurent également dans ce numéro, portant notamment sur la sélection d'index bitmap de jointure.

Parmi les travaux présentés, d'autres contributions se situent au confluent de l'extraction de connaissances et de l'ingénierie des connaissances en manipulant à la fois des données et des connaissances du domaine d'application comme en accidentologie ou dans le domaine de la médecine. La modélisation multivues des connaissances et leur formalisation sous forme de méta-données permet leur intégration au processus d'extraction : certains auteurs tentent alors d'optimiser l'extraction en utilisant des connaissances issues du domaine de l'analyste.