

Chapitre 10

Qualité des Données et des Connaissances

Les problèmes de qualité des données stockées dans les bases ou les entrepôts de données s'étendent à tous les domaines gouvernemental, commercial, industriel et scientifique. La découverte de connaissances et la prise de décision à partir de données de qualité médiocre (c'est-à-dire contenant des erreurs, des doublons, des incohérences, des valeurs manquantes, etc.) ont des conséquences directes et significatives pour les entreprises et pour tous leurs utilisateurs. Le thème de la qualité des données et des connaissances est pour cela devenu, depuis ces dernières années, un des sujets d'intérêt émergeant à la fois dans le domaine de la recherche et dans les entreprises.

Toutes les applications dédiées à l'analyse des données (telles que la fouille de données textuelles par exemple) requièrent différentes formes de préparation des données avec de nombreuses techniques de traitement, afin que les données passées en entrée aux algorithmes de fouille se conforment à des distributions relativement « sympathiques », ne contenant pas d'incohérences, de doublons, de valeurs manquantes ou incorrectes. Seulement, entre la réalité des données disponibles et toute la machinerie permettant leur analyse, un assez vaste fossé demeure.

In fine, l'évaluation des résultats issus du processus de traitement des données, est généralement effectuée par un spécialiste (expert, analyste,...). Cette tâche est souvent très lourde, et un moyen de la faciliter consiste à aider le spécialiste en lui fournissant des critères de décision sous la forme de mesures de qualité ou d'intérêt des résultats. Ces mesures de qualité des

connaissances doivent être conçues afin de combiner deux dimensions : une dimension objective liée à la qualité des données, et une dimension subjective liée aux intérêts du spécialiste. L'atelier Qualité des Données et des Connaissances - DKQ 2005 (Data and Knowledge Quality) - associé à EGC'2005 concerne les méthodes, les techniques d'analyse et de nettoyage, les méthodologies, les approches algorithmiques et les métriques de qualité des données et des connaissances permettant de comprendre, d'explorer les données, de détecter et corriger les problèmes de qualité des données et de qualité des connaissances extraites à partir des données.

Comité de programme et d'organisation :

Laure Berti-Équille (IRISA-CNRS Rennes, France)	Fabrice Guillet (IRIN, Université de Nantes France)
Ansaf Salleb (IRISA-CNRS Rennes, France)	Pascale Kuntz, (IRIN, Université de Nantes, France)
Jérôme Azé (LRI, Université de Paris-Sud, France)	Stéphane Lallich (ERIC, Université de Lyon 2, France)
Mokrane Bouzeghoub (PRISM, Université de Versailles, France)	Ludovic Lebart (ENST-CNRS, Paris, France)
Henri Briand (IRIN, Université de Nantes, France)	Philippe Lenca (ENSTbr, Brest, France)
Béatrice Duval (Université d'Angers, France)	Amedeo Napoli (LORIA, Nancy, France)
Johann-Christoph Freytag (Humboldt-Universität zu Berlin, Germany)	Gilbert Ritschard (Université de Genève, Switzerland)
Helena Galhardas (INESC, Lisboa, Portugal)	Monica Scannapieco (Universita' di Roma "La Sapienza", Italy)
Régis Gras (IRIN, Université de Nantes, France)	Dan A. Simovici (University of Massachusetts, Boston, U.S.)
Yves Kodratoff (LRI, Université de Paris-Sud, France)	Einoshin Suzuki (Yokohama National University, Japan)
	Djamel Zighed (ERIC, Université de Lyon 2, France)

Remerciements

Nous remercions tout particulièrement les auteurs et les membres du comité de relecture pour leur contribution au succès de l'atelier DKQ2005.