La qualité des données comme condition à la qualité des connaissances : un état de l'art

Laure Berti-Équille

IRISA, Campus Universitaire de Beaulieu
35042 Rennes, France
Laure.Berti-Equille@irisa.fr
http://www.irisa.fr

Résumé. Les travaux actuels sur l'extraction de connaissances à partir des données (ECD) se focalisent sur la recherche de règles intéressantes dont on souhaite pouvoir qualifier l'intérêt ou le caractère exceptionnel, mais dont la validité dépend bien évidemment de celle des données. En amont du processus d'ECD, il semble donc essentiel d'évaluer la qualité des données stockées dans les bases et entrepôts de données afin de : (1) proposer aux utilisateurs une expertise critique de la qualité du contenu d'un système, (2) orienter l'extraction des connaissances en fonction d'un profil ciblé d'utilisateurs et de décideurs, (3) permettre à ceux-ci de relativiser la confiance qu'ils pourraient accorder aux données et aux règles extraites, et leur permettre ainsi de mieux en adapter leur usage, (4) assurer enfin la validité et l'intérêt des connaissances extraites à partir des données. Cet article fait une synthèse de l'état de l'art dans le domaine de la qualité des données en présentant, dans un premier temps, les causes de la non-qualité des données, puis en décrivant un panorama des travaux sur la qualité des données, travaux pertinents dès lors que l'on s'intéresse à modéliser, mesurer et à améliorer la qualité des connaissances "élaborées" à partir des données. Enfin, l'article propose d'exploiter les méta-données décrivant la qualité des données dans le processus d'ECD.
Mots-clés. Qualité des données, méta-données

1. Introduction

Avec la multiplication des sources d'informations disponibles et l'accroissement des volumes de données potentiellement accessibles, l'extraction de connaissances à partir des données a pris une place de premier plan tant au niveau académique qu'au sein des entreprises. En effet, la mise en évidence de liens cachés ou de phénomènes de causalités non-triviales à partir de grandes quantités de données permettra d'aider les décideurs dans leurs choix. Cependant, l'identification, à partir d'une grande collection de données, de motifs valides, nouveaux, potentiellement utiles et compréhensibles dépend de manière très critique de la qualité des données (généralement intégrées) qu'utilisent les algorithmes de fouille de données [CM95] [HG+95] [CDL+97] [Vas00].

Si l'analyse des données peut être réalisée sur des données inexactes, incomplètes, ambiguës et de qualité médiocre, on peut s'interroger sur le sens à donner aux résultats de ces analyses et remettre en cause, à juste titre, la qualité des connaissances ainsi "élaborées".