Deux méthodologies de classification de règles d'association pour la fouille de textes

Hacène Cherfi, Amedeo Napoli et Yannick Toussaint
LORIA, BP 239, 54506 Vandœuvre-lès-Nancy cedex
{cherfi,napoli,yannick}@loria.fr,
http ://www.loria.fr/equipes/orpailleur/

Résumé. Parmi les inconvénients d'un processus de fouille de données textuelles fondé sur l'extraction de règles d'association figurent le grand nombre de règles extraites et la difficulté d'affecter à une règle un critère de qualité fiable par rapport aux connaissances de l'analyste (*i.e.*, l'expert du domaine). La plupart des approches pour la classification des règles d'association utilisent des méthodes statistiques pour juger de la qualité d'une règle et ne s'appuient pas sur les connaissances du domaine des données disponibles *a priori* pour classer les règles extraites. Dans cet article nous définissons la notion de qualité d'une règle d'association. Nous étudions en premier lieu les mesures statistiques permettant de classer les règles et nous proposons un algorithme combinant ces différentes mesures. Nous introduisons ensuite une nouvelle méthodologie de classification des règles exploitant un modèle de connaissances. Nous expérimentons cette mesure sur un exemple formel puis nous l'évaluons sur des données réelles.

1 Introduction

Les textes, du point de vue de la fouille de données, sont des données complexes qui posent de nouveaux défis. En premier lieu, les textes sont des données peu structurées contrastant avec les données des bases de données pour lesquelles un travail de modélisation est réalisé au préalable. Ils sont rédigés en langue naturelle avec tout ce que cela suppose en termes d'implicite, d'ambiguïté, d'imprécision, etc. Il n'existe pas de représentation standard décrivant l'intégralité du contenu d'un texte et sur laquelle il est possible d'appliquer des méthodes de fouille de données ; les représentations habituelles sont le plus souvent partielles et bruitées. Un second facteur de complexité vient du fait qu'expliciter l'implicite véhiculé par un texte nécessite le recours à un modèle de connaissances du domaine. Pour pouvoir déduire de nouvelles connaissances à partir de textes, il faut dépasser le principe de la co-occurrence des mots-clés dans les textes et pouvoir appliquer des opérations de généralisation ou de spécialisation sur les mots-clés en fonction de connaissances disponibles, et cela afin de pouvoir manipuler, regrouper ou dissocier les textes par exemple.

Dans cet article, nous présentons un processus de *fouille de textes*, qui s'aligne sur le schéma de référence de l'extraction de connaissances dans des bases de données introduit dans [Fayyad *et al.*, 1996]. Ce processus de fouille s'appuie sur une boucle itérative et interactive et place l'expert du domaine, appelé *analyste* par la suite, au centre du processus de fouille. Le processus de fouille de textes vise à construire, par des enrichissements successifs, un modèle du domaine. Réciproquement, à chaque itération,