Validation d'une expertise textuelle par une méthode de classification basée sur l'intensité d'implication

Jérôme David*, Fabrice Guillet*,Vincent Philippé**, Henri Briand*, Régis Gras*

*LINA - Ecole Polytechnique de l'université de Nantes
La Chantrerie - BP 50609 - 44306 Nantes cedex 3
jerome.david,fabrice.guillet,regis.gras,henri.briand@polytech.univ-nantes.fr,
**PerformanSe SAS - Atlanpole - La Fleuriaye - 44470 Carquefou
vincent.philippe@performanse.fr

Résumé. Dans le cadre d'une validation d'expertise textuelle contenue dans un test de compétences comportementales informatisé, nous proposons une méthode visant à extraire des sous-ensembles de termes caractéristiques utilisés pour décrire des caractères psychologiques. Notre approche consiste, après l'extraction de termes, à évaluer les associations possibles entre termes et caractères psychologiques qui structurent le corpus en s'appuyant sur la théorie de l'implication statistique.

1 Introduction

Les documents sous forme de textes représentent des quantités d'information colossales. L'Extraction de Connaissances à partir de Textes (ECT) ou text-mining, vise à extraire des connaissances pertinentes, contenues dans des données textuelles, à l'aide des modèles utilisés en Extraction des Connaissances dans les Données. Parmi les modèles utilisés en ECT, la découverte de règles d'associations entre termes contenus dans les textes est souvent utilisée (Maedche and Staab, 2000; Janetzko et al., 2004; Roche, 2003).

Les règles d'association (Agrawal et al., 1993) sont des tendances implicatives $a \Rightarrow b$ entre attributs booléens caractérisées par deux mesures : le support et la confiance. Parmi les indices alternatifs de qualité proposés dans la litérature (Tan et al., 2004; Guillet, 2004; Lenca et al., 2004), nous nous intéressons à la mesure d'intensité d'implication définie par R. Gras (Gras, 1979; Gras et al., 1996).

Cependant avant d'utiliser les techniques d'ECD, les données linguistiques doivent subir une phase de Traitement Automatique du Langage (TAL), dont le but est d'obtenir à partir d'un texte, la liste des termes qu'il contient. De nombreuses approches sont proposées : approches statistiques (Salem, 1986), approches linguistiques (David and Plante, 1990; Jacquemin, 1997), ou mixtes qui combinent les deux approches précédentes (Smadja, 1993; Daille, 1994).

En nous inscrivant à l'intersection des domaines de la recherche d'information et du text-mining, nous proposons une méthode d'étude et de validation d'une indexation par des profils psychologiques de documents traitant de bilans de compétences comportementales dans le cadre de la théorie de l'implication statistique. L'objectif de notre étude est d'associer à chaque caractère d'un profil psychologique, une classe de termes.