Une mesure de proximité et une méthode de regroupement pour l'aide à l'acquisition d'ontologies spécialisées

Guillaume Cleuziou , Sylvie Billot , Stanislas Lew ,
Lionel Martin , Christel Vrain

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans
Rue Léonard de Vinci - 45067 ORLEANS Cedex 2
prénom.nom@univ-orleans.fr

Résumé. Cet article traite du regroupement d'unités textuelles dans une perspective d'aide à l'élaboration d'ontologies spécialisées. Le travail présenté s'inscrit dans le cadre du projet BIOTIM. Nous nous concentrons ici sur l'une des étapes de construction semi-automatique d'une ontologie qui consiste à structurer un ensemble d'unités textuelles caractéristiques en classes susceptibles de représenter les concepts du domaine. L'approche que nous proposons s'appuie sur la définition d'une nouvelle mesure non-symétrique permettant d'évaluer la proximité entre lemmes, en utilisant leurs contextes d'apparition dans les documents. En complément de cette mesure, nous présentons un algorithme de classification non-supervisée adapté à la problématique et aux données traitées. Les premières expérimentations présentées sur les données botaniques laissent percevoir des résultats pertinents pouvant être utilisés pour assister l'expert dans la détermination et la structuration des concepts du domaine.

1 Introduction

L'exploitation de données textuelles issues de fonds scientifiques est un objectif de recherche ambitieux dans le domaine de la gestion et de l'acquisition des connaissances. Une des premières étapes pour la mise en place d'un système d'information est la construction d'une ontologie du domaine. Dans cette étude, nous abordons le problème de construction d'une ontologie spécialisée avec une approche mixte (ou semi-automatique). Pour cela, nous nous intéressons à l'étape d'extraction automatique de classes terminologiques susceptibles d'être ensuite validées comme concepts puis structurées par un expert du domaine, l'embryon d'ontologie résultant devant par la suite être enrichi automatiquement.

La tâche de regroupement de mots peut être envisagée de différentes manières (selon l'application visée, les connaissances disponibles sur le domaine ou les traitements possibles). Les études proposées dans ce domaine s'intéressent généralement à l'une des deux étapes suivantes : la définition d'une mesure de proximité entre mots et/ou la proposition d'une méthode de regroupement efficace.

Il existe de nombreuses mesures destinées à évaluer la proximité sémantique entre des mots. On peut classer ces mesures en trois grandes catégories : statistiques, syntaxiques ou