Une approche combinée pour l'enrichissement d'ontologie à partir de textes et de données du LOD

Céline Alec*, Chantal Reynaud-Delaître*, Brigitte Safar*

*LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay, F-91405
prenom.nom@lri.fr

Résumé. Cet article porte sur l'étiquetage automatique de documents décrivant des produits, avec des concepts très spécifiques traduisant des besoins précis d'utilisateurs. La particularité du contexte est qu'il se confronte à une triple difficulté : 1) les concepts utilisés pour l'étiquetage n'ont pas de réalisations terminologiques directes dans les documents, 2) leurs définitions formelles ne sont pas connues au départ, 3) toutes les informations nécessaires ne sont pas forcément présentes dans les documents mêmes. Pour résoudre ce problème, nous proposons un processus d'annotation en deux étapes, guidé par une ontologie. La première consiste à peupler l'ontologie avec les données extraites des documents, complétées par d'autres issues de ressources externes. La deuxième est une étape de raisonnement sur les données extraites qui recouvre soit une phase d'apprentissage de définitions de concepts, soit une phase d'application des définitions apprises. L'approche SAUPODOC est ainsi une approche originale d'enrichissement d'ontologie qui exploite les fondements du Web sémantique, en combinant les apports du LOD et d'outils d'analyse de texte, d'apprentissage automatique et de raisonnement. L'évaluation, sur deux domaines d'application, donne des résultats de qualité et démontre l'intérêt de l'approche.

1 Introduction

Ce travail se situe dans le cadre d'un partenariat entre le LRI et la startup Wepingo [1], qui développe des applications en ligne proposant des produits à des internautes. Pour faciliter la conception de systèmes flexibles, adaptables à différentes catégories de produits et à différents points de vue sur ces produits, notre objectif est de concevoir une approche permettant d'étiqueter automatiquement des documents décrivant des produits, avec des concepts très spécifiques traduisant des besoins précis des utilisateurs. La particularité de notre approche est qu'elle se confronte à une triple difficulté : 1) les concepts utilisés pour l'étiquetage n'ont pas de réalisations terminologiques directes dans les documents, 2) les définitions formelles de ces concepts ne sont pas connues au départ même si le concepteur du système sait à partir de quelles informations elles pourraient être construites, 3) toutes les informations nécessaires ne sont pas forcément présentes dans les documents mêmes.

1. http://www.wepingo.com/fr-fr/