Extraction de concepts guidée par le contexte

Lobna Karoui, Marie-Aude Aufaure

Ecole Supérieure d'Electricité
Plateau de Moulon 3 rue Joliot Curie
91192 Gif-sur-Yvette cedex, France
{Lobna.karoui, Marie-Aude.Aufaure}@supelec.fr
http://www.supelec.fr/ecole/si/pers.html

Résumé. Les ontologies constituent la brique supportant les échanges et le partage des informations en étendant l'interopérabilité syntaxique du web en une interopérabilité sémantique. Dans cet article, nous présentons une méthode d'extraction de concepts ontologiques utilisant un algorithme de clustering non supervisé et guidé par le contexte à partir de pages Web. Notre méthode est basée sur une approche unifiée intégrant des dimensions complémentaires pour l'acquisition de connaissances conceptuelles. En particulier, nous exploitons les caractéristiques structurelles des documents HTML afin de localiser et de définir un contexte approprié pour chaque terme en respectant ses différentes positions dans le corpus. Notre définition contextuelle permet de sélectionner les co-occurrents sémantiquement proches et de définir une mesure de pondération appropriée pour chaque couple de termes. Notre méthode se base sur une évaluation interactive et incrémentale de la qualité des clusters par l'utilisateur. Nous l'avons expérimentée sur un corpus du domaine portant sur le tourisme. Les premiers résultats obtenus montrent bien que la prise en compte du contexte des termes guidant le clustering améliore considérablement la pertinence des concepts extraits

1 Introduction

Les ontologies constituent la brique supportant les échanges, le partage et la recherche d'information en étendant l'interopérabilité syntaxique du web en une interopérabilité sémantique. Elles permettent de représenter un ensemble de concepts formellement définis, acceptés par une communauté d'utilisateurs. Selon les domaines et les besoins applicatifs, les ontologies seront plus ou moins riches, allant de simples métadonnées, à des taxonomies jusqu'à de véritables bases de connaissances. Elles servent de squelette de structuration sémantique des données et représentent une valeur ajoutée pour leur manipulation, leur traitement et leur interrogation. La réalisation du web sémantique dépend de la construction des ontologies et de leur déploiement. Le problème majeur concerne les coûts de cette construction pour un grand nombre d'ontologies de domaines et d'applications. La classification automatique est un angle d'attaque permettant de déterminer des regroupements des données en classes et d'extraire des concepts du domaine et leurs relations ; ce qui constitue un point crucial pour cette tâche lourde et complexe que représente la construction d'ontologies

Dans cet article, nous présentons une approche unifiée d'extraction de connaissances à partir de documents HTML en vue de construire une ontologie du domaine. En effet, l'abondance et l'importance de pages HTML comme une riche source d'information sont un fait