Construction d'ontologie à partir de corpus de textes

Rokia Bendaoud *, Yannick Toussaint *
Amedeo Napoli *

* LORIA - Campus scientifique BP 239
54506 Vandoeuvre-Lès-Nancy, CEDEX.
{bendaoud,napoli,yannick}@loria.fr

Résumé. Cet article présente une méthode semi-automatique de construction d'ontologie à partir de corpus de textes sur un domaine spécifique. Cette méthode repose en premier lieu sur un analyseur syntaxique partiel et robuste des textes, et en second lieu, sur l'utilisation de l'analyse formelle de concepts "FCA" pour la construction de classes d'objets en un treillis de Galois. La construction de l'ontologie, c'est à dire d'une hiérarchie de concepts et d'instances, est réalisée par une transformation formelle de la structure du treillis. Cette méthode s'applique dans le domaine de l'astronomie.

1 Introduction

Une ontologie est une structure formelle dans laquelle les concepts d'un domaine et les relations entre ces concepts sont définis (Gruber (1993)). Notre ontologie porte sur l'astronomie : dans leurs articles scientifiques, les astronomes identifient manuellement les caractéristiques des objets célestes, afin de les associer ensuite à une catégorie (galaxie, étoile, ...). Les catégories sont pré-définies et l'astronome détermine la classe correspondant le mieux à l'objet étudié. Cette classification a permis de catégoriser 3.751.128 objets célestes. Pourtant, il reste encore des milliards d'objets à classifier et à caractériser de la manière la plus exhaustive possible. L'utilisation des articles scientifiques, très facilement accessibles sous format électronique, permettent de répondre à ces attentes.

Nous proposons une méthode semi-automatique de construction d'une ontologie sur le domaine de l'astronomie. Les concepts de l'ontologie sont des classes dont les instances sont les objets célestes. Les propriétés de chaque classe sont partagées par toutes ses instances. Ces propriétés sont extraites automatiquement des textes par un analyseur syntaxique partiel et robuste "Enju" de Miyao et Tsujii (2005). Objets et propriétés sont classés dans un treillis de Galois selon l'analyse formelle des concepts : FCA présentée dans Ganter (1999). Le résultat de cette méthode est fourni aux astronomes afin d'étiqueter chaque classe d'après les propriétés partagées par les instances de la classe.

Notre méthode présente plusieurs avantages :

– elle peut être appliquée quelque soit le corpus de textes et le domaine spécifique sur lequel elle est utilisée,
– elle est formalisée par la FCA,
– elle est rapide comparée à une ontologie construite manuellement,
– et elle permet d'enrichir l'ontologie résultante par la mise à jour du corpus de textes.