RAS : Un outil pour l'annotation de documents basée sur les liens de citation

Lylia Abrouk, Danièle Hérin

LIRMM. 161 rue ada, 34392 Montpellier
{abrouk,dh}@lirmm.fr

Résumé. RAS (Reference Annotation System) est un outil d'annotation de documents. Cet outil est le résultat de l'implémentation de notre approche d'annotation basée sur le contexte de citation. L'approche est indépendante du contenu et utilise un regroupement thématique des références construit à partir d'une classification floue non-supervisée. L'outil présenté dans cet article a été expérimentée et évaluée avec la base de documents scientifiques Citeseer.

1 Introduction

RAS[1], Reference Annotation System est un outil semi-automatique d'annotation de documents basé sur le contexte de citation, l'expert du domaine reste décideur de la fiabilité de l'annotation. L'approche d'annotation permet d'annoter un document sans connaissance préalable de son contenu, en se basant sur les références. Cet outil a été réalisé dans le contexte d'un besoin réel, celui d'une communauté souhaitant partager l'information existante et ceci sous certaines contraintes, la plus importante étant celle de l'absence de contenu des documents à partager. Afin de tester les résultats de l'annotation, nous avons utilisé une base avec un nombre important de documents qui s'inter-référencent. L'outil utilise les technologies suivantes :

- Python[2] comme langage de script ;
- la base documentaire Citeseer[3] ;
- L'ontologie dmoz[4] (informatique) ;
- l'algorithme de classification fuzzy C-means Dunn (1973).

2 Fonctionnement et principales fonctionnalités

De manière générale l'outil permet de réaliser une annotation sur un document existant dans la base. L'outil permet de visualiser le résultat de l'annotation sous forme d'une liste de concepts de l'ontologie présentés sous la forme d'une hiérarchie.

Les étapes d'annotation implémentés dans RAS sont les suivantes Abrouk et al. (2006) :

1. Récupérer l'ensemble des documents cités par d dans un ensemble noté Ref_d.

[1] www.lirmm.fr/annotation
[2] http ://www.python.org/
[3] http ://citeseer.ist.psu.edu/
[4] http ://www.dmoz.org/

2. Sélectionner les annotations les plus proches thématiquement. Pour cela nous devons trouver les références les plus proches par un regroupement thématique. En effet, regrouper thématiquement les documents de l'ensemble Ref_d consiste à déterminer les groupements thématiques les plus pertinents et éviter ainsi les références non pertinentes mais présentes dans Ref_d. Pour cela, nous utilisons une mesure de similarité basée sur la méthode de co-citation Small (1973) . Le regroupement s'effectue par l'algorithme fuzzy C-means à partir de cette distance thématique.
3. Importer les annotations des documents cités par d.
4. Sélectionner parmi les annotations importées les plus pertinentes pour les proposer comme annotations du document d. L'outil utilise l'ordre suivant afin d'annoter le nouveau document : *(i.)* l'importance du cluster, *(ii.)* le degré d'appartenance au cluster, *(iii)* le nombre de fois où l'annotation apparaît.

Le résultat dans RAS est illustré dans la figure 1

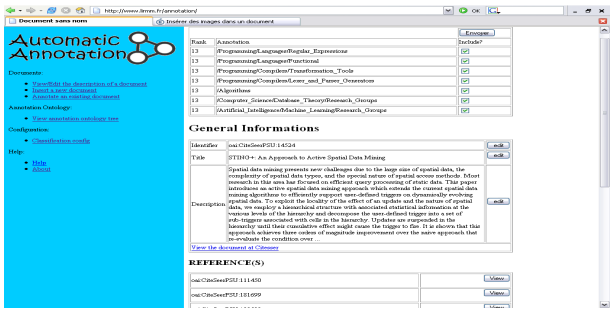

FIG. 1 – *Résultat de l'annotation par RAS*

Références

Abrouk, L., A. Gouaich, et C. Raïssi (2006). Annotation automatique de documents. In *Proceedings of INFORSID 2006*, Hammamet, Tunisie, pp. 483–497.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics 3*, 32–57.

Small, H. (1973). Co-citation in the scientific literature. *Society for Information Science 24*.

Summary

RAS (Reference System Annotation) is a documents annotation tool. This tool is the result of the implementation of our approach of annotation based on the context of citation. The approach is independent of the content and uses a regrouping set of themes of the references builds starting from a not-supervised fuzzy classification. The tool presented in this article was tested and evaluated with the base of scientific documents Citeseer.