Approche innovante pour la recherche et l'extraction coopérative et dynamique d'informations sur Internet

Xavier Denis*,**, Gaële Simon*
Nicolas Chanchevrier**

*UFR Sciences & Techniques, 25 Rue Philippe Lebon, 76600 Le Havre
xavier.denis@operamail.com, gaele.simon@iut.univ-lehavre.fr
http ://www-lih.univ-lehavre.fr
**EADS S&DE, Parc d'Affaires des Portes BP613, 27106 Val De Reuil Cedex
nicolas.chanchevrier@tiscali.fr
http ://www.eads.com

Résumé. Il existe de nombreuses techniques qui permettent de classifier des documents textuels en fonction du centre d'intérêt d'un utilisateur (kNN, SVM, ...). Malheureusement, l'intégration de ces méthodes dans des plate-formes de textmining est souvent très statique et ne permet pas facilement d'affiner les traitements et/ou résultats au cours du temps. Le but de cet article est de présenter une plate-forme de webmining dans laquelle les données hétérogènes sont représentées uniformément selon un formalisme XML/TEI et où l'utilisateur peut interagir sur les processus de récupération et d'analyse de ces données. Pour cela, les modules de traitements sont représentés par des agents fonctionnant sur la plate-forme MadKit et l'apprentissage se fait sur une méthode dérivée de VSM[1] et TFIDF utilisant un principe de listes noires pondérées permettant la reconnaissance de documents indésirables. La dynamique de la plate-forme repose principalement sur la possibilité d'ajouter à la volée des agents de traitement et de pouvoir modifier l'ordre et les paramètres d'analyse des documents.

1 Introduction

La richesse des informations disponibles en ligne [Woodruff *et al.*, 1996] et leur diversité de contenu a ouvert la porte à un besoin grandissant ces dernières années : la recherche, l'analyse et la distribution d'informations. Ces trois grandes étapes définissent ce qui est communément appelé un processus de veille [Goujon, 2000]. Des logiciels existent déjà pour répondre à un certain nombre de ces besoins mais leur conception généralement fermée (API obscure et non documentée, outils commerciaux, ...) limite leur utilisabilité.

L'approche qui a été retenue ici est basée sur un système multi-agents MadKit [Gutknecht et Ferber, 2000] où chaque agent, relié ou non à un autre (principe de flux d'agents décrit au paragraphe 2.4), participe à chaque étape du processus de veille. Bien qu'assez simple en apparence, ce principe de flux implique une grande souplesse

[1] Vector Space Model