Le défi Fouille de Textes : Quels paradigmes pour la reconnaissance automatique d'auteurs ?

Violaine Prince et Yves Kodratoff

*LIRMM-CNRS et Université Montpellier 2
161 rue Ada, 34395 Montpellier cedex 5
prince@lirmm.fr
http://www.lirmm.fr/ prince
** LRI-CNRS Université Paris Sud
kodratoff@lri.fr

Résumé. Les campagnes d'évaluation en traitement automatique du langage naturel et en informatique documentaire sont devenues un passage obligé pour la reconnaissance des différentes techniques employées. Le Défi Fouille de Texte a pour objectif de permettre aux chercheurs du monde francophone de confronter leurs travaux avec un problème, plus que de primer une équipe, une méthode, ou un outil. Dans cet article nous évoquons les diverses problématiques de la fouille de textes, à savoir la recherche d'information, l'extraction ou l'enrichissement de connaissances, la classification/catégorisation de documents, la segmentation de textes, le profilage. La reconnaissance d'auteur, objet de ce premier défi, est une tâche complexe et composite qui nécessite de traiter simultanément de la segmentation, de la catégorisation et du profilage. L'idée générale est que la mise en place des défis est un outil de cartographie des diverses avancées en fouille de textes, et également un instrument scientifique de compréhension de problèmes de nature complexe.

1 Introduction

Les campagnes d'évaluation en traitement automatique du langage naturel et en informatique documentaire sont devenues un passage obligé pour la reconnaissance de la qualité et de l'efficacité des différentes techniques employées dans ces domaines. Leurs applications, en recherche d'information, extraction de connaissances dans les textes, et fouilles de texte apparaissent naturellement comme le champ privilégié de la mise en compétition des travaux des chercheurs. Au-delà même des techniques, ce sont réellement des paradigmes scientifiques qui s'affrontent.

Si pendant plusieurs décennies les méthodes à base logique, issues de l'Intelligence Artificielle pour laquelle le langage naturel a toujours représenté le défi ultime (n'est-il pas l'élément prégnant du test de Turing ?), ont eu le vent en poupe, les campagnes d'évaluation, depuis les premiers TREC[1] et MUC[2] semblent redorer fortement le blason des méthodes à fondement statistique ou probabiliste.

[1] Text Retrieval Evaluation Conference
[2] Message Understanding Conference