PIITHIE

Programme Technologies Logicielles 2006

PIITHIE
Plagiat et Impact de l'Information Textuelle recHerchée dans un contexte InterlinguE

Février 2007 – Mars 2009

Chef de file Sinequa

Partenaires Advestigo, LIA, LINA, Syllabs (prestataire)

Résumé Le projet **PIITHIE** s´inscrit dans un mouvement de plus en plus important de maîtrise de l´information diffusée. Il vise premièrement la ***détection de plagiats de textes***. Les techniques de traitement automatique des langues (TAL), devraient permettre d'améliorer les performances et d'accroître le potentiel de recherche des outils d'Advestigo et de Sinequa. Le deuxième objectif concerne le ***suivi d´impact*** : les diffuseurs d'information sont très intéressés par la possibilité d´évaluer l´impact de leur production. Aujourd´hui cette évaluation est faite par une étude manuelle alors que des méthodes automatiques sont possibles. Les traitements nécessaires à ces deux applications sont de même nature ; ils demandent seulement un paramétrage différent selon que l´on cherche une copie illégale de l´information ou une utilisation parfaitement légale et dont le contenu peut être très divergent. Les principaux défis scientifiques de ce projet concernent

1. la capacité à évaluer la proximité de deux contenus textuels en tenant compte des différents phénomènes de réécriture
2. l'extraction de termes suffisamment représentatifs d'un document pour pouvoir retrouver des documents similaires sur Internet en posant des requêtes à un moteur classique
3. la détection de citations dont il faut tenir compte pour l'évaluation d'impact et qui perturbent la détection de plagiat.

D'autres applications sont possibles à partir des éléments développés dans le cadre de ce projet :

- Un module de *recherche de documents similaires* sur Internet en utilisant des moteurs de recherche classiques. Si la plupart des moteurs de recherche sur Internet proposent actuellement une option « rechercher des documents similaires » lorsqu'on regarde un document trouvé par le moteur, il n'existe pas encore (à notre connaissance) de système permettant de retrouver des documents

similaires à partir d'un document se trouvant sur notre disque dur. Un tel module entrerait naturellement dans un système de veille.

- Un module de *détection de citations*, très utile pour l'analyse d'un texte.
- Un module d'*extraction d'empreinte* qui permettra d'améliorer les performances du moteur de recherche de Sinequa (rapidité de recherche de documents similaires, extraction de mots clés, aide à la navigation, visualisation de l'information importante).