L'introduction à la "fouille de texte et de données" et positionnement de l'offre logicielle

Patrice Bellot*[1]

[1]Laboratoire dÍnformatique et Systèmes – Aix Marseille Université : UMR7020, Université de Toulon : UMR7020, CNRS : UMR7020 – France

Résumé

Présentation de la conférence
La fouille de données textuelles informatisée met en jeu un certain nombre de disciplines scientifiques parmi lesquelles la linguistique et les statistiques sont centrales. Au fil des années et selon certains choix guidés par la nature des données manipulées (langues, textes et documents) et des tâches à réaliser mais aussi par des impératifs ergonomiques ou économiques, l'équilibre entre ces disciplines a évolué pour donner lieu à une offre logicielle vaste et variée, plus ou moins interactive ou dépendante de ressources humaines et de données volumineuses. Ce sont ces différents aspects qui seront présentés et qui permettront d'introduire les ateliers en les mettant en perspective avec les enjeux actuels.

Programme détaillé

Introduction sur les domaines scientifiques impliqués dans la fouille de textes

- TAL et fouille de données : En quoi les données textuelles sont particulières (lexique, syntaxe mais aussi diversité langagière, des formats, des entités, des méta-données etc.) et quels sont les types de ressources utiles ou disponibles.
- Des modèles et des tâches (analyse grammaticale, désambiguisation, similarité textuelle, recherche et extraction d'information, classification...) et des collections standard pour évaluer des modèles et des outils
- Les approches automatisées sont associées à différentes manières de travailler les corpus (règles manuelles, apprentissage et bases d'exemples, degrés de supervision humaine, ...) : avantages / inconvénients, risques ...

Panorama méthodologique de l'offre logicielle académique ou commerciale

- Des outils pour l'utilisateur final, des APIs pour le développement, des plateformes d'annotation pour la création de bases d'apprentissage, des outils pour écrire des règles symboliques
- Des outils logiciels plus ou moins interactifs

*Intervenant