Haruspex, outil de gestion de connaissances non strucutrées

Matthieu Quantin[1,2], Benjamin Hervy[3], Jean-Louis Kerouanton[2] et Florent Laroche[1]

[1]IRCCyN - École Centrale de Nantes | prenom.nom@irccyn.ec-nantes.fr
[2]CFV - Univeristé de Nantes | prenom.nom@univ-nantes.fr
[3]MSH Ange-Guépin - Pays de la loire | prenom.nom@univ-nantes.fr

Mots-clés : *graphe ; indexation ; distance ; sémantique ; TAL ; base de données ; corpus ; non-structure ; document ; outil ; logiciel ; mot-clé*

L'objet de cette communication est de proposer un outil pour l'analyse et l'exploitation de corpus de documents non-structurés ou faiblement structurés.
Aujourd'hui la création de corpus de données numérique (ouverts ou privés) est un phénomène massif. De plus en plus de données sont scannées, photographiées, retranscrites, etc pour être analysées. Les jeux de données numériques (que l'on se crée souvent sois-même) constituent la matière exclusive, quotidienne du chercheur. Ce phénomène demande à être accompagné par une évolution des outils d'analyse : données physiques et données numérique ont des potentiels d'analyse différents. Or le chercheur en SHS est souvent démuni face aux sources non structurées qu'il collecte : articles, scan d'archives, documents OCR, images et métadonnées. La mise en place d'une base de données se résume souvent (au mieux) à un "tableau excel". Les domaines du bigdata et du data-mining sont cantonnés à des projets de très grande envergure, pour des données déjà structurées, avec une équipe de soutien logistique conséquente. Un "gap" s'établit entre le chercheur en histoire, en archéologie, en sociologie et les "humanité numériques".

L'outil proposé, intitulé Haruspex, vise à réduire ce gap. Il traite des données texte (et images éventuellement) en français ou en anglais, pour produire une base de données orientée graphe, requêtable, contenant les documents liés entre-eux (distance sémantique). En entrée, divers formats (pdf, txt, odt, latex...) sont pris en charge, le processus se déroule ensuite en 4 étapes :

1. Gestion de corpus : création ou récupération d'éventuelles métadonnées (dates, lieux, étiquetage) pour les documents ; concaténation, découpage, regroupements, exclusion, ...
2. Indexation sémantique de ce corpus : extraction de mots clés (génériques mais aussi très spécifiques), puis classification de ces mot-clés en catégories (si possible).
3. Modération des résultats précédents par l'utilisateur.
4. Calcul de la "distance sémantique" entre documents à partir de l'indexation modéré.

Les premiers essais de ce logiciel dans divers domaines : patrimoine industriel, histoire de la chimie au XX^{e} siècle, histoire du travail dans les colonies et analyse des publication scientifiques, sont concluant aux yeux des chercheurs du domaine concerné. Haruspex est fonctionnel et son développement est rapide et dynamique, son interface graphique est très en retard. Enfin une perspective forte réside dans la création de documentation multi-accès. En effet, la structure des données permet de documenter des éléments publiés en proposant des liens vers d'autres items proches (sémantiquement voire géographiquement ou temporellement par exemple). Une forte contrainte de proximité conviendrait plutôt au grand public curieux, une ouverture sur des documents plus éloignés conviendrait à des objectifs de recherche.