Sujet de projet 2016 – 2017

Fouille de données textuelles pour l'enrichissement de thésaurus

Contact : A. Widlöcher (antoine.widlocher@unicaen.fr)
Encadrement : B. Crémilleux & E. Giguet & F. Loew-Turbout & A. Widlöcher

Contexte et présentation générale

Un thésaurus est une ressource terminologique permettant de représenter et d'articuler entre eux les concepts d'un domaine de connaissance. Il est notamment utilisable lors de l'indexation des documents pour établir un lien entre leur contenu (textuel, graphique...) et les concepts qui y sont manipulés. Les thésaurus peuvent être encodés de différentes manières, les technologies du web sémantique (RDF, SKOS...) jouant toutefois en la matière un rôle prépondérant.

Quelle que soit la méthode de représentation retenue, la constitution d'un thésaurus repose pour sa part dans la plupart des cas sur une expertise humaine devant garantir la fiabilité de la ressource. L'objectif de ce projet est d'offrir à l'expert un environnement logiciel pouvant simplifier le processus de composition d'un thésaurus.

Pour les besoins de l'expérimentation, les solutions proposées s'appliqueront plus particulièrement aux données textuelles issues de l'*Atlas Politique*[1] et devront assister l'expert dans la composition d'un thésaurus consacré à la terminologie politique et électorale.

Méthodologie

La solution proposée devra reposer sur la double articulation entre un thésaurus d'une part et un processus de fouille de données textuelles d'autre part :

1. un premier thésaurus minimaliste, porteur de termes connus et d'articulations connues entre ces termes sera utilisé pour l'amorçage du système ;
2. ces termes et articulations connues seront projetés sur le corpus textuel, afin de localiser, dans le texte, des emplacements où les termes apparaissent et où leur articulation est probablement décrite ;
3. des procédés de fouille de données textuelles seront appliqués à ces emplacements dans le but d'identifier des motifs caractéristiques des articulations conceptuelles recherchées ;
4. les motifs identifiés seront ensuite appliqués au reste du corpus, dans le but de découvrir des relations de même nature entre de nouveaux termes ;
5. ces termes, leurs relations supposées et les extraits qui ont conduit à leur extraction seront présentés à l'expert, pour validation ;
6. les termes et relations validés alimenteront le thésaurus ;
7. la procédure décrite sera à nouveau appliquée, pour enrichissement incrémental du thésaurus.

Plan de travail

Pour mener à bien ce travail, il sera donc nécessaire :

1. d'étudier les formats conventionnels de représentation de thésaurus, et notamment SKOS, car il sera nécessaire d'exploiter les informations présentes dans le thésaurus en cours de constitution et d'enrichir celui-ci ;
2. de proposer à l'utilisateur un environnement de visualisation de l'état courant du thésaurus ;
3. de définir une méthode de reprérage, dans le texte, de passages où occurrent des termes décrits et mis en relation dans le thésaurus ;
4. de lancer sur ces passages des outils d'extraction de motifs, outils déjà disponibles (l'étudiant n'aura donc pas à implémenter lui-même des algorithmes de fouille) ;
5. de définir une méthode permettant, étant donnés des motifs extraits à l'étape précédente, de repérer, au sein du corpus, des occurrences de ces motifs, afin d'identifier de nouveaux termes et de nouvelles relations supposés pertinents pour le thésaurus ;
6. d'offrir à l'utilisateur un outil de consultation de ces termes et relations candidats, outil lui permettant de valider les éléments à ajouter au thésaurus.

1. *Atlas politique - Évolution politique de l'Ouest de la France*, Université de Caen Basse-Normandie, SAIC-CERTIC, UMR CNRS 6590 ESO, accessible à l'adresse http://atlas-politique.certic.unicaen.fr.