K-WORDS LAB : un outil d'analyse des mots clés permettant d'explorer les dynamiques d'un domaine scientifique.

Audrey Baneyx*, Philippe Breucker**

*Université Paris-Est, IFRIS, LATTS UMR CNRS 8134, France
baneyx@ifris.org
**INRA Sens, Université Paris-Est, IFRIS, France
breucker@ifris.org

1 Introduction

L'utilisation par les chercheurs des Sciences Humaines et Sociales (SHS) d'outils de fouille de données, d'extraction et d'analyse des connaissances est aujourd'hui une nécessité pour comprendre et étudier les dynamiques de certains domaines scientifiques (Turenne et Barbier, 2004). C'est dans ce contexte que s'inscrit le développement de l'outil K-WORDS LAB au sein de la plateforme CorTexT[1] de l'Institut francilien « Recherche, Innovation et Société ». K-WORDS LAB est un outil construit pour analyser les dynamiques qui témoignent de l'émergence d'un nouveau domaine scientifique. Nous pensons qu'il est possible de suivre les dynamiques qui se forment au sein d'un domaine scientifique en constitution en analysant sa production (projets, articles scientifiques, brevets, ...) (Baneyx et al., 2010).

2 Matériel et méthodologie

K-Words Lab est un outil basé sur la technologie php5/MySql5. L'interface web est écrite en php/HTML/CSS et certaines parties en flash (technologie Flex) et JavaScript. La méthodologie mise en œuvre dans K-WORDS LAB mobilise les apports du traitement automatique de la langue, de l'intelligence artificielle, des statistiques et de la sociologie des sciences. L'outil propose une chaîne de traitement automatique qui fait appel à plusieurs modules et va de l'analyse du corpus à la représentation dynamique et interactive des clusters de concepts qui caractérisent le domaine étudié, ainsi que son évolution.
(1) **Le module d'importation** analyse les différents types de corpus (notices bibliographiques, notices de dépôt de brevets, ...), les stocke et les indexe dans la base de données.
(2) **Le module linguistique** a pour objectif de réduire le nombre d'occurrences de mots clés[2] aux mots clés distincts compris dans le corpus.
(3) **Le module de clustering** implémente un clustering linguistique et un clustering contextuel. Le clustering linguistique regroupe les mots clés ayant des formes orthographiques proches les

[1] http ://www.cortext.fr
[2] Nous avons pour l'instant testé jusqu'à 10 000 000 occurrences de mots clés.

unes des autres (QTClust). Le clustering contextuel implémente des algorithmes recouvrants (notamment hiérarchiques) et non recouvrants et regroupe les mots clés qui apparaissent régulièrement ensemble.
(4) **Le module statistique** calcule, sur la base des mots clés distincts, un ensemble de statistiques classiques : taux de croissance, écart-type, nombre d'occurrences par année, ...
Ces mesures permettent de définir une typologie en distinguant quatre profils de mots clés : émergent, générique, persistant et non persistant. L'interface permet à l'utilisateur de fournir sa propre définition pour chaque profil de mots clés.
(5) **Le module de visualisation** implémente quatre interfaces pour permettre à l'utilisateur (a) d'interroger la base de données et de filtrer sa recherche en fonction de données statistiques, (b) de visualiser un mot clé, son évolution sur la période et de naviguer parmi ses plus proches voisins, (c) d'interroger la base de données en fonction de la typologie des mots clés définie précédemment, enfin (d) de visualiser et naviguer de manière dynamique et interactive dans les clusters de mots clés qu'ils soient linguistiques ou contextuels.

3 Conclusion

Les nombreuses interactions entre concepteurs spécialistes de l'ingénierie des connaissances et utilisateurs sociologues ont permis de développer un outil qui profite des techniques de l'informatique, du traitement automatique de la langue et du regard que porte la sociologie des sciences sur le traitement de l'information. A l'utilisation, K-WORDS LAB est un outil robuste, capable de traiter un volume important de données là où d'autres outils (Calliope, Alceste et ReseauLu) connaissent des difficultés.

Références

Baneyx, A., A. Delemarle, P. Breucker, L. Villard, B. Kahane, et P. Larédo (2010). Exploiting keywords life-cycle to analyse the dynamics of an emerging field : an experiment in nanosciences with the k-words lab. In *Proceedings of the European Network of Indicator Designers Conference (ENID) - A paraitre*.

Turenne, N. et M. Barbier (2004). Beluga : un outil pour l'analyse dynamique des connaissances de la littéraure scientifique d'un domaine. première application au cas des maladies à prions. *Revue des Nouvelles Technologies de l'Information E-2*, 423–428.

Summary

K-WORDS LAB is a tool designed to analyse the dynamics of an emerging field. At the end of the automatic processing chain, the interface provides the user a two-level topology of the field: (1) keywords classification, with a differentiation between persistent and non-persistent, emergent and generic keywords, (2) dynamical exploration between the major concepts of the field and their evolution over time.