Exploration des Données du Défi EGC 2016 à l'aide d'un Système d'Information Logique

Peggy Cellier*, Sébastien Ferré**, Annie Foret**, Olivier Ridoux**

* INSA Rennes, IRISA — ** Université Rennes 1, IRISA — {prenom.nom}@irisa.fr

Résumé. Nous présentons dans cet article les méthodes employées et les résultats obtenus en réponse au Défi EGC 2016. Notre approche repose d'une part sur des chaînes automatiques de traitements linguistiques en français et en anglais utilisant le plus possible des ressources et outils publics et d'autre part sur un environnement d'exploration des données basé sur les systèmes d'information logiques ; ces systèmes exploitent une généralisation des treillis de concepts formels appliquée aux données attribut-valeur ou au web sémantique.

1 Introduction

Le Défi EGC 2016 propose d'exploiter un fichier texte contenant des descriptifs de publications EGC. Nous avons choisi d'exploiter ces données en utilisant les méthodes de l'analyse de concepts logiques (Ferré et Ridoux, 2004) via les outils Camelis et Sparklis. Ces méthodes étant de nature purement symbolique il faut commencer par extraire des données les traits symboliques que nous voulons exploiter de façon à s'affranchir des variations linguistiques dans la donnée : flexions, synonymie, voire même langue. Dans une première partie nous décrivons la chaîne de traitement utilisée pour le nettoyage et l'enrichissement du jeu de données fourni (section 2). Dans une seconde partie nous décrivons comment les données enrichies peuvent être explorées à l'aide de systèmes d'information logiques (section 3).

2 Nettoyage et enrichissement du jeu de données

2.1 Nettoyage et traitements linguistiques

Jeu de données : Articles RNTI. Le jeu de données fourni pour le Défi [1] est un fichier texte contenant 1937 lignes et où chaque ligne représente un article de recherche publié entre 2004 et 2015. Pour chaque article, 8 champs peuvent être renseignés : *series*, *booktitle*, *year*, *title*, *abstract*, *authors*, *pdf1page* et *pdfarticle*.

Filtrage et nettoyage du jeu de données. Dans un premier temps le fichier de données a été converti d'un encodage Windows à l'encodage UTF8 et les articles ont été filtrés grâce au champ *booktitle* pour ne conserver que les 1103 articles publiés à la conférence EGC entre

1. `RNTI_articles_export.txt` à l'adresse `http://editions-rnti.fr/?m=articles_export`