Défi EGC 2016 : Analyse par Motifs Fréquents et Topic Modeling

Julien Aligon*, Fabrice Guillet*, Julien Blanchard*, Fabien Picarougne*

*Université de Nantes
Laboratoire Informatique de Nantes Atlantique (LINA)
Equipe Duke
prenom.nom@univ-nantes.fr

Résumé. Dans le domaine de l'analyse de textes, l'extraction de motifs est une technique très populaire pour mettre en évidence des relations fréquentes entre les mots. De même, les techniques de topic modeling ont largement fait leurs preuves lorsqu'il s'agit de classer automatiquement des ensembles de textes partageant des thématiques similaires. Ainsi, ce papier a pour ambition de montrer l'intérêt de l'utilisation conjointe de ces deux techniques afin de mettre en évidence, sous la forme d'un graphe biparti, des mots partageant des thématiques similaires mais aussi leurs relations fréquentes, intra et inter thématiques. Les données du Défi EGC 2016 permettent de valider l'intérêt de l'approche, tout en montrant l'évolution des thématiques et des mots clés parmi les papiers de la conférence EGC sur ces onze dernières années.

1 Introduction

La fouille de données (Han et Kamber (2000)) est devenue un domaine incontournable pour l'analyse de grands volumes de données, auxquels nous sommes désormais confrontés au quotidien (notamment dans le contexte du web). Le principe général de la fouille est d'extraire de l'information pertinente dans l'objectif de caractériser des phénomènes que l'on suppose présents dans les données à traiter. Dans le domaine de l'analyse de textes, l'extraction de motifs fréquents est une technique très populaire pour mettre en évidence des relations fréquentes entre les mots à analyser. De même, les techniques de topic modeling ont largement fait leurs preuves lorsqu'il s'agit de classer automatiquement des ensembles de textes partageant des thématiques similaires. Ainsi, ce papier a pour ambition de montrer la complémentarité et l'intérêt de l'utilisation conjointe de ces deux techniques, afin de produire des visualisations mettant en relation temporellement puis spatialement les thématiques. La visualisation temporelle est basée sur un diagramme de Sankey et permet d'étudier l'influence des thématiques entre elles sur des périodes de temps séquentielles. L'analyse spatiale est construite sur des thématiques et des associations de mots extraits sur un aggloméra de périodes temporelles et prend la forme d'un graphe biparti reliant ces thématiques aux mots. La relation mot-thématique est construite sur une notion de similarité mais aussi sur leurs relations fréquentes, intra et inter thématiques.

Ces deux méthodes sont appliquées aux données fournies par le Défi EGC 2016, et limitées aux résumés des papiers de la conférence EGC sur ces dix dernières années. Le choix de ne