Le Processus Powered Dirichlet-Hawkes comme A Priori Flexible pour Clustering Temporel de Textes

Gaël Poux-Médard*, Julien Velcin*, Sabine Loudcher*

*Université de Lyon, Lyon 2, ERIC UR 3083
5 avenue Pierre Mendès France, F69676 Bron Cedex, France
http://eric.univ-lyon2.fr

Résumé. Le contenu textuel d'un document et sa date de publication sont corrélés. Par exemple, une publication scientifique est influencée par les précédents articles cités dans ladite publication. Utiliser cette corrélation permet d'améliorer la compréhension de grands corpus textuel datés. Cependant, cette tâche peut se compliquer lorsque les textes considérés sont courts ou possèdent des vocabulaires similaires. De plus, la corrélation entre texte et date est rarement parfaite. Nous développons une méthode répondant à ces limites, permettant de créer des clusters de documents en fonction de leur contenu et de leur date : le processus Powered Dirichlet-Hawkes (PDHP). Nous montrons que PDHP présente de meilleures performances que les modèles état de l'art (qu'il généralise) lorsque l'information textuelle ou temporelle est peu informative. Le PDHP se libère également de l'hypothèse d'une corrélation parfaite entre texte et date des documents. Enfin, nous illustrons une possible application sur des données réelles, provenant de Reddit.

1 Introduction

Les contenus numériques sont générés à une vitesse sans précédent. Chaque minute, environ 500 000 commentaires sont postés sur Facebook, 400h de vidéo sont mises en ligne sur Youtube, et 500 000 messages sont publiés sur Twitter. Une approche possible pour comprendre cette masse d'informations est de regrouper ces publication en groupes (clusters) thématiques. Grouper des publications similaires aiderait par exemple à automatiser la détection non-supervisée de thématiques ou à générer des résumés de nouvelles journalières.

De récents travaux ont montré que considérer la date de publication augmente les performances des algorithmes de clustering (Du et al., 2012). La plupart de ces modèles fonctionnent par échantillonnage : une observation récente aura plus de chances d'être utilisée dans l'apprentissage qu'une observation éloignée dans le temps (Ahmed et Xing, 2008; Blei et Lafferty, 2006; Yin et al., 2018). Cependant, cela implique que la fonction d'échantillonnage temporel transcrive bien la réalité des dynamiques à l'oeuvre, ce qui n'est pas évident. En outre, ces modèles se basent sur un *a priori* de Dirichlet (DP) pour créer les clusters. Hors, il a été montré que DP n'est pas assez flexible pour décrire des processus en temps continu. Dans (Du et al., 2015), les auteurs dérivent le processus de Dirichlet-Hawkes (DHP), et l'utilisent comme *a priori* bayésien pour grouper des documents en considérant leur date de publication en temps