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

Gaël Poux-Médard, Julien Velcin, Sabine Loudcher Rabaseda

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

continu. Cependant, de récentes études montrent que cette méthode trouve ses limites dans les cas où le contenu textuel est peu informatif (textes courts, vocabulaires similaires entre thématiques, etc.) (Yin et al., 2018).

Notre travail vise donc à dépasser ces limites en développant le processus Powered Dirichlet Hawkes (ou Powered Dirichlet-Hawkes process, PDHP). Nous montrons qu'il existe d'autres contextes dans lesquels DHP échoue à rendre compte du processus à l'oeuvre, là où PDHP présente de bonnes performances. Typiquement, DHP échoue dans les cas où les dates de publication sont peu informatives. Nous montrons également qu'il existe des cas où l'information textuelle est décorrélée des dynamiques de publication, ce que DHP ne considère pas. Par exemple, un article publié par un journal influent aura bien plus de résonance qu'un article strictement identique publié par un journal moins populaire ; l'influence de cet article ne dépend pas uniquement de son contenu, mais également de sa date. Nous répondons à toutes ces limitations avec le PDHP et démontrons sont efficacité sur plusieurs jeux de données (jusqu'à +0.3 NMI par rapport à DHP). Notre méthode permet également de distinguer les clusters *textuels* des clusters *temporels* lorsque le texte et la dynamique ne sont pas parfaitement corrélés.

Ce travail s'articule comme suit : nous brossons d'abord un court état de l'art des méthodes de clustering temporel et des alternatives aux processus de Dirichlet. Nous formulons ensuite notre *a priori* mathématiquement, puis l'associons à un modèle de langue simple. Enfin, nous effectuons des expériences sur des jeux de données synthétiques et réels, et résumons les résultats obtenus, avant de conclure en ouvrant sur de possibles extensions du PDHP. Explicitement, listons nos contributions :

— Explication des limites de DHP : quand le contenu textuel ou temporel est peu informatif. En outre le modèle suppose une corrélation parfaite entre ces deux variables.
— Dérivation du processus Powered Dirichlet-Hawkes (PDHP), qui généralise DHP et le processus Uniforme (UP).
— Démonstration par une étude systématique sur plusieurs jeux de données que PDHP fournit de meilleurs résultats que DHP et UP.
— Démonstration que PDHP permet d'ajuster l'importance donnée à l'information textuelle ou temporelle dans de la création des clusters de documents.

2 Contexte

Pourquoi considérer le temps — La date de publication d'un document fournit une information utile aux problèmes de clustering. Souvent, l'apparition d'un nouveau document a été conditionnée par l'existence des documents précédents. Par exemple, dans la publication scientifique : le présent article est construit grâce à l'existence des références qu'il cite, et qui traitent d'un sujet proche sinon similaire. En 2012, il a été montré que l'ordre d'apparition (ou la date) de tweets modifie notre réaction aux tweets suivants (Myers et Leskovec, 2012). En outre, la dimension temporelle n'importe que pour certains groupes bien définis d'information (Poux-Médard et al., 2021a) et ne perdure pas dans le temps (Cao et Sun, 2019). Il est donc nécessaire d'allier la notion de cluster à la notion de dynamique pour parvenir à une description correcte des processus de publication en ligne.

Clustering temporel de documents textuels — Plusieurs modèles ont été proposés pour modéliser une évolution dynamique de clusters textuels (Blei et Lafferty, 2006; Wang et McCallum, 2006; Ahmed et Xing, 2008; Blei et Frazier, 2010). Cependant, la plupart d'entre

eux fonctionnent en fixant le nombre de clusters à inférer au préalable, ou approximent les dynamiques temporelles via une fonction d'échantillonnage. ce qui ne permet pas de modéliser des dynamiques temporelles. En 2015, N. Du *& al* ont proposé le DHP pour considérer des comptes non-entiers dans un espace de temps continu (en se passant d'échantillonnage temporel). L'idée clé est de remplacer les comptes du processus de Dirichlet standard par des fonctions d'intensité provenant des processus de Hawkes, dépendant du temps, résultant en un processus de Dirichlet-Hawkes (DHP).

Alternatives aux processus de Dirichlet — DHP (et ses variantes) utilisent comme base un processus de Dirichlet (DP) classique. Cependant, DP est un choix arbitraire (Welling, 2006). (Wallach et al., 2010) a proposé le processus Uniforme (UP), puis (Poux-Médard et al., 2021b) l'a généralisé avec le processus Powered Dirichlet (PDP) ; les variantes UP et PDP fonctionnent mieux que DP pour plusieurs jeux de données. Comme PDP généralise UP et DP, nous proposons PDHP comme généralisation de DHP et UP. Comme nous le verrons, cela nous permettra de pallier les limites de DHP lorsqu'il s'agit de textes courts (Yin et al., 2018) ou de dynamiques décorrélées du contenu textuel.

3 PDHP : Powered Dirichlet-Hawkes Process

3.1 PDP : Powered Dirichlet Process

Dans un processus de Dirichlet, une nouvelle observation aura une chance *a priori* d'appartenir à chaque cluster existant proportionnelle au nombre d'observations qui y appartiennent – "les riches deviennent plus riches". Le processus plus général décrit dans (Poux-Médard et al., 2021b) propose de modifier cette dépendance linéaire par une forme permettant une dépendance en puissance de r. Dans la formulation proposée, pour $r = 1$ le processus est identique à un processus de Dirichlet (DP), et pour $r = 0$ il est identique au processus uniforme (UP). Mathématiquement, si C_i est le cluster choisi par la $i^{\text{ème}}$ observation, $\vec{C^-}$ l'historique d'allocation des $i-1$ observations précédentes, N_k la population du cluster k parmi K clusters non-vides, et α_0 un paramètre, on peut écrire le Powered Dirichlet Process (PDP) :

$$\text{PDP}(C_i = c | r, \vec{C^-}, \alpha_0) = \begin{cases} \frac{N_c^r}{\alpha_0 + \sum_k^K N_k^r} \text{ si c = 1, 2, ..., C} \\ \frac{\alpha_0}{\alpha_0 + \sum_k^K N_k^r} \text{ si c = C+1} \end{cases} \quad (1)$$

3.2 Processus de Hawkes

Le processus de Hawkes est une classe de processus du point (*point process)* définie comme auto-stimulées ; la probabilité d'un nouvel événement dépend de la réalisation d'événements précédents. Un tel processus est entièrement caractérisé par sa fonction d'intensité $\lambda(t)$, reliée à la probabilité qu'un événement survienne entre t et $t + \Delta t$ par $\lambda(t) \overset{\Delta t \to 0}{=} \frac{P(t_{event} \in [t; t+\Delta t])}{\Delta t}$. Dans notre modèle, chaque cluster se voit associer une telle fonction d'intensité $\lambda_c(t|\mathcal{H}_{<t,c})$, où $\mathcal{H}_{<t,c}$ est l'historique des événements survenus avant t dans le cluster c. Hypothèse est faite que les événements au sein d'un même cluster provoqueront les futurs événements appartenant

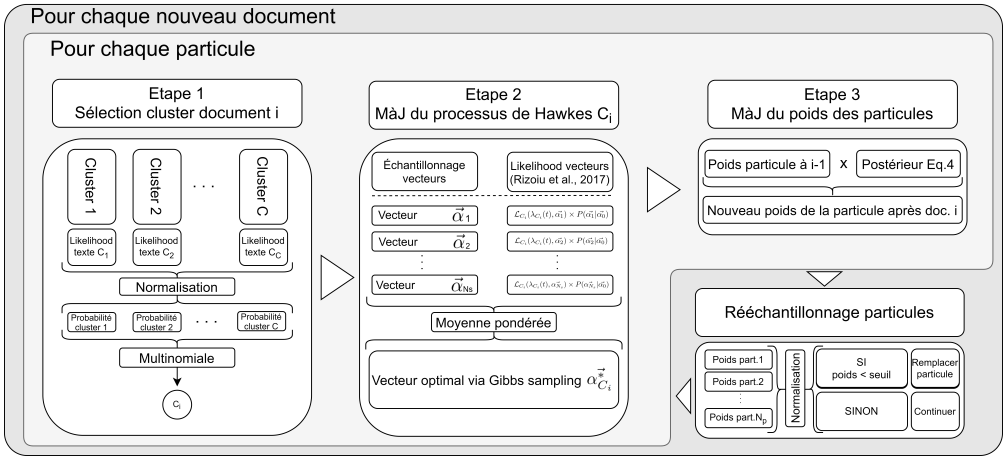

FIG. 1 – ***Fonctionnement de l'algorithme*** *— Pour chaque nouveau document, on exécute les étapes 1 (échantillonnage du cluster), 2 (mise à jour de la dynamique du cluster) et 3 (mise à jour de la plausibilité des particules) pour chaque particule, puis on remplace les particules trop peu plausibles par d'autres plus plausibles.*

à ce même cluster. Explicitement :

$$\lambda_c(t|\mathcal{H}_{<t,c}) = \sum_{t_{i,c}\in\mathcal{H}_{<t,c}} \vec{\alpha_c}^T \cdot \vec{\kappa}(t_{i,c}) \tag{2}$$

où $\vec{\alpha_c}$ est un vecteur de coefficients que l'on cherchera à inférer, et $\vec{\kappa}(t)$ un vecteur de fonctions prédéfinies (*kernel functions*) dépendant du temps. La vraisemblance associée avec le processus de Hawkes de chaque cluster est détaillée dans la littérature associée (Rizoiu et al., 2018).

3.3 PDHP : Powered Dirichlet Hawkes Process

Suivant le raisonnement effectué dans (Du et al., 2015), nous substituons les comptes du processus Powered Dirichlet par les fonctions d'intensité des processus de Hawkes, ce qui permet d'avoir une dépendance non-linéaire (r) avec les fonctions d'intensité associées à chaque cluster. Nous aboutissons à la forme suivante du PDHP :

$$P(C_i = c|t_i, r, \lambda_0, \mathcal{H}_{<t_i,c}) = \begin{cases} \frac{\lambda_c^r(t_i)}{\lambda_0+\sum_{c'}\lambda_{c'}^r(t_i)} & \text{si c}\leq\text{C} \\ \frac{\lambda_0}{\lambda_0+\sum_{c'}\lambda_{c'}^r(t_i)} & \text{si c=C+1} \end{cases} \tag{3}$$

Nous rappelons que le PDHP a pour vertu d'être utilisé comme *a priori* bayésien, et doit donc être couplé à un autre modèle –en l'occurrence, un modèle de langue. Nous choisissons de considérer un simple modèle Dirichlet-Multinomial pour modéliser le contenu textuel des documents, par soucis de simplicité. Nous pouvons écrire explicitement *a posteriori* bayésien de notre modèle. Soient N_c le nombre total de mots dans le cluster c avant la $i^{\text{ème}}$ observation,

n_i le nombre total de mots dans le document i, $N_{c,v}$ de nombre d'occurrences du mot v dans le cluster c, $n_{i,v}$ le nombre d'occurrences du mot v dans le document i et $\theta_0 = \sum_v \theta_{0,v}$ le paramètre de concentration du modèle Dirichlet-Multinomial. :

$$
\begin{aligned}
P(C_i = c|r, n_i, t_i, N_c, \mathcal{H}_{<t,c}) &\propto \underbrace{P(n_i|C_i = c, N_{<i,c}, \theta_0)}_{\text{Vraisemblance textuelle}} \underbrace{P(C_i = c|t_i, r, \lambda_0, \mathcal{H}_{<t_i,c})}_{\text{A priori temporel}} \\
&= \frac{\Gamma(N_c + \theta_0)}{\Gamma(N_c + n_i + \theta_0)} \prod_v \frac{\Gamma(N_{c,v} + n_{i,v} + \theta_{0,v})}{\Gamma(N_{c,v} + \theta_0)} \times \begin{cases} \frac{\lambda_c^r(t_i)}{\lambda_0 + \sum_{c'} \lambda_c^r(t_i)} & \text{si c = 1, ..., C} \\ \frac{\lambda_0}{\lambda_0 + \sum_{c'} \lambda_{c'}^r(t_i)} & \text{si c = C+1} \end{cases}
\end{aligned} \tag{4}
$$

Abaisser la valeur de r se traduit par une modification de l'importance accordée à la dynamique dans notre modélisation. Une faible valeur de r lissera les différences entre les fonctions d'intensité des clusters (pour $r = 0$, la valeur de l'*a priori* est identique pour tous les clusters – c'est le processus uniforme), si bien que l'allocation à un cluster se fera sur la base du contenu textuel uniquement. Au contraire, une grande valeur de r exacerbe les différences d'intensité temporelle entre les clusters, et donc force l'allocation à se baser sur l'aspect temporel, alors discriminant. C'est ici que se trouve la principale contribution de notre méthode : régler r permet de choisir si les clusters doivent être formés sur la base de l'information textuelle, temporelle, ou d'une mixture des deux.

Pour l'inférence, nous utilisons un algorithme à particules similaire à celui de (Du et al., 2015; Mavroforakis et al., 2017; Poux-Médard et al., 2021c). Chaque particule représente une hypothèse d'allocation de documents à des clusters. Les particules représentant des hypothèses trop peu plausibles sont remplacées par d'autres à chaque itération. Nous schématisons son exécution Fig.1. Plus de détails dans Poux-Médard et al. (2021c) et dans l'implémentation [1].

4 Expériences

4.1 Données

Données synthétiques — Nous générons des données en utilisant deux clusters. Chaque cluster a son vocabulaire propre (1000 mots), qu'il partage dans une proportion donnée avec l'autre cluster (*textual overlap*). Pour chaque cluster, nous simulons un processus de Hawkes, et associons à chaque événement 20 mots tirés parmi le vocabulaire de son cluster. Les fonctions d'intensité résultantes se chevauchent sur l'axe temporel dans une certaine proportion ("*Hawkes intensity overlap*"). Pour différentes valeurs d'*overlaps*, nous générons 10 jeux de données différents. Nous évaluons nos performances en considérant la NMI (*normalized mutual information*). Les autres métriques testées donnent des résultats similaires et ne sont donc pas reportées ici. En outre, nous simulons une décorrélation entre la dynamique temporelle et le contenu textuel. Pour des *overlaps* nuls, nous sélectionnons au hasard un certain pourcentage d'événements et leur associons un nouveau vocabulaire tiré d'un cluster choisi aléatoirement. Nous avons alors deux types de groupes : celui qui a servi a générer le contenu textuel (cluster textuel) et celui qui a déterminé la date de publication (cluster temporel).

Données réelles — Nous considérons également un jeu de données provenant de Reddit, qui fera ici office d'illustration notre méthode, constitué de 73.000 titres de posts provenant de 9 subreddits d'information en avril 2019 avec date de publication associée.

1. Données et codes disponibles sur https://github.com/GaelPouxMedard/PDHP

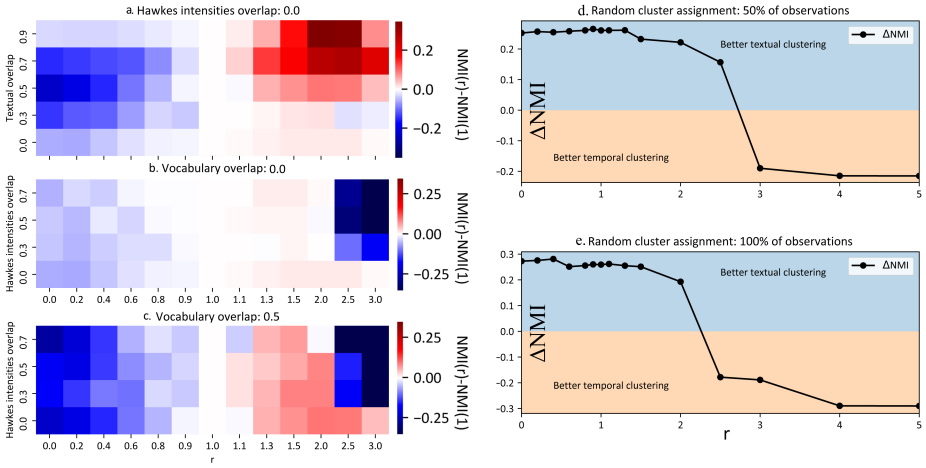

FIG. 2 – ***PDHP permet d'améliorer les performances de l'état de l'art*** *— **(Gauche)** Différence de NMI entre PDHP et DHP Du et al. (2015) pour plusieurs valeurs de r et d'overlap textuel et temporel, moyenné sur 10 jeux de données par configuration.* ***(Droite) Varier*** *r* ***permet de choisir entre un clustering textuel ou temporel*** *— La ligne noire représente le différence entre la NMI temporelle et la NMI textuelle. Pour de petits r, le clustering textuel est meilleur que le temporel ; pour de grands r, l'inverse.*

4.2 Résultats

Nous reportons les résultats sur les **corpus synthétiques** Fig. 2. Nos expériences mènent aux conclusions suivantes :

— PDHP permet de meilleures performances par rapport à DHP et UP lorsque l'information textuelle est faible (vocabulaires qui se recoupent, textes courts) : gain jusqu'à +0.3 de NMI par rapport à DHP (Fig. 2a). PDHP permet une faible amélioration des résultats lorsque les données temporelles sont informatives, c'est à dire lorsque *Hawkes intensity overlap* est nul (Fig. 2b). Enfin, dans des situations plus réalistes où à la fois le vocabulaire et les intensités se recoupent, PDHP permet de meilleurs résultats : gain jusqu'à +0.2 de NMI par rapport à DHP (Fig. 2c).

— Fig. 2(d,e). Lorsque le contenu textuel est décorrelé de la dynamique de publication (ce qui se rapproche des processus de diffusion d'information réels), varier r permet de modéliser avec de bonnes performances l'une ou l'autre information.

Sur des **données réelles**, notre méthode fournit des résultats satisfaisants. Elle parvient à extraire des séries de publications ayant trait aux événements majeurs d'avril 2019 : bombardements au Sri Lanka, arrestation de Julian Assange, image directe d'un trou noir, incendie de Notre-Dame, entre autres. En nous concentrant sur les bombardements au Sri Lanka Fig.3, notre modèle repère correctement les points forts de la crise les 21, 22 et 23 avril. Lorsque r est grand, le modèle créée des clusters avec des documents suivant un rythme de publication quotidien, sans motif apparent dans le vocabulaire employé, comme attendu. Des images similaires pour d'autres sujets sont disponibles en ligne, avec le code et les données utilisées.

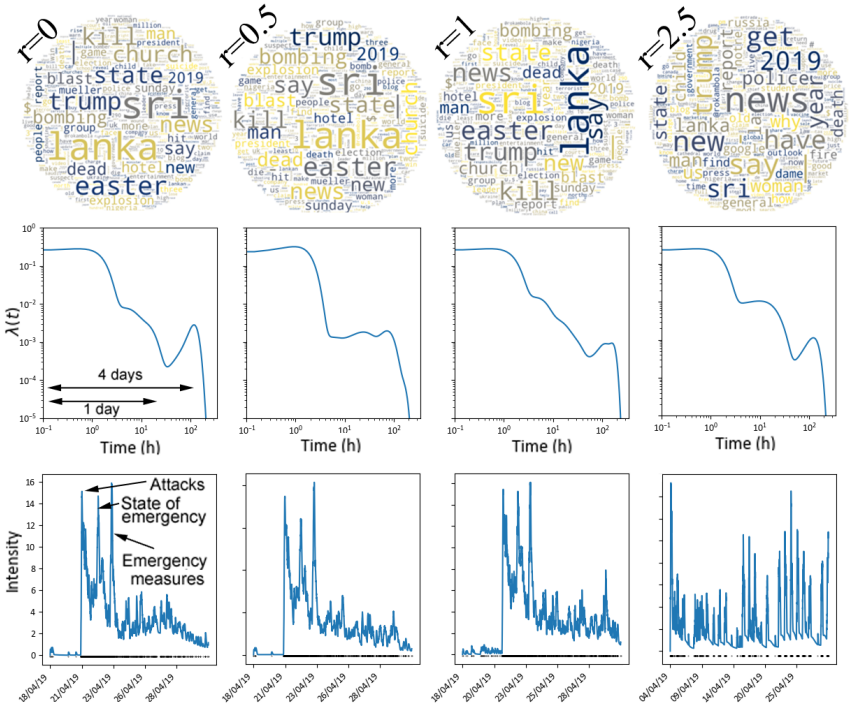

FIG. 3 – ***Un cluster tiré de PDHP appliqué à des titres de journaux : les bombardements au Sri Lanka*** — *Nuages de mots de ce cluster, kernel $\lambda(t)$ (Eq.2), et intensité des clusters de vocabulaire proche des bombardements du Sri Lanka en avril 2019 pour plusieurs r.*

5 Conclusion

Nous avons développé le DHP dans une forme plus générale, PDHP, qui permet non seulement de prendre en compte la date de publications de documents dans leur clustering, mais également de choisir l'importance qui lui est donnée par rapport à son contenu textuel. Nombre d'extensions sont possibles pour PDHP. Par exemple, en développer une version considérant des processus de Hawkes *multivariés* permettrait d'étudier comment les thématiques s'influencent les unes les autres. Une application couplant PDHP avec un modèle de langue plus sophistiqué que Dirichlet-Multinomial serait également intéressant. Finalement, une méthode permettant d'inférer la valeur de r en fonction d'une quantité à optimiser (réduire la variance textuelle des clusters, par exemple) permettrait de faire de PDHP, mais également de PDP, des *a priori* objectifs décrivant au mieux le monde réel.

Références

Ahmed, A. et E. Xing (2008). Dynamic non-parametric mixture models and the recurrent chinese restaurant process : with applications to evolutionary clustering. In *SDM08*.

Blei, D. M. et P. Frazier (2010). Distance dependent chinese restaurant processes. ICML'10.

Blei, D. M. et J. D. Lafferty (2006). Dynamic topic models. ICML '06, pp. 113–120. ACM.

Cao, J. et W. Sun (2019). Sequential choice bandits : learning with marketing fatigue. *AAAI-19*.

Du, N., M. Farajtabar, A. Ahmed, A. Smola, et L. Song (2015). Dirichlet-hawkes processes with applications to clustering continuous-time document streams. *KDD*.

Du, N., L. Song, A. Smola, et M. Yuan (2012). Learning networks of heterogeneous influence. *NIPS 4*, 2780–2788.

Mavroforakis, C., I. Valera, et M. Gomez-Rodriguez (2017). Modeling the dynamics of learning activity on the web. WWW '17, pp. 1421–1430.

Myers, S. A. et J. Leskovec (2012). Clash of the contagions : Cooperation and competition in information diffusion. *2012 IEEE 12th International Conference on Data Mining*, 539–548.

Poux-Médard, G., J. Velcin, et S. Loudcher (2021a). Interactions in information spread : quantification and interpretation using stochastic block models. *RecSys'21*.

Poux-Médard, G., J. Velcin, et S. Loudcher (2021b). Powered dirichlet process for controlling the importance of "rich-get-richer" prior assumptions in bayesian clustering. *ArXiv*.

Poux-Médard, G., J. Velcin, et S. Loudcher (2021c). Powered hawkes-dirichlet process : Challenging textual clustering using a flexible temporal prior. *ICDM*.

Rizoiu, M.-A., Y. Lee, S. Mishra, et L. Xie (2018). A tutorial on hawkes processes for events in social media. *Frontiers of Multimedia Research*, 191–218.

Wallach, H., S. Jensen, L. Dicker, et K. Heller (2010). An alternative prior process for nonparametric bayesian clustering. pp. 892–899. JMLR.

Wang, X. et A. McCallum (2006). Topics over time : A non-markov continuous-time model of topical trends. KDD '06, pp. 424–433. Association for Computing Machinery.

Welling, M. (2006). Flexible priors for infinite mixture models. In *Workshop on learning with non-parametric Bayesian methods*.

Yin, J., D. Chao, Z. Liu, W. Zhang, X. Yu, et J. Wang (2018). Model-based clustering of short text streams. KDD '18, pp. 2634–2642. Association for Computing Machinery.

Summary

The textual content of a document and its publication date are intertwined. For example, the publication of a news article on a topic is influenced by previous publications on similar issues, according to underlying temporal dynamics. However, it can be challenging to retrieve meaningful information when textual information conveys little. Furthermore, the textual content of a document is not always correlated to its temporal dynamics. We develop a method to create clusters of textual documents according to both their content and publication time, the Powered Dirichlet-Hawkes process (PDHP). PDHP yields significantly better results than state-of-the-art models when temporal information or textual content is weakly informative. PDHP also alleviates the hypothesis that textual content and temporal dynamics are perfectly correlated. We demonstrate that PDHP generalizes previous work –such as DHP and UP. Finally, we illustrate a possible application using a real-world dataset from Reddit.