Modélisation de la propagation de l'information sur le Web : de l'extraction des données à la simulation

François Nel[*,**], Marie-Jeanne Lesot[*]
Philippe Capet[**] Thomas Delavallade[**]

[*]LIP6 - Université Pierre et Marie Curie-Paris6, UMR7606
4 place Jussieu 75252 Paris cedex 05
{francois.nel, marie-jeanne.lesot}@lip6.fr,
[**] Thales Land and Joint Systems
160, boulevard de Valmy - BP 82 - 92704 Colombes Cedex
{francois.nel, philippe.capet, thomas.delavallade}@fr.thalesgroup.com

Résumé. Nous proposons un modèle de la propagation de l'information dans un réseau, en détaillant toutes les étapes de sa réalisation et de son utilisation dans un cadre de simulation. A partir de données réelles extraites du Web, nous identifions parmi les sources des catégories de comportements de publication distincts. Nous proposons ensuite une extension d'un modèle de diffusion de l'information existant, afin d'augmenter son pouvoir d'expression, en particulier pour reproduire ces comportements de publication, puis nous le validons sur un exemple de simulation.

1 Introduction

L'étude des phénomènes informationnels passe par la modélisation des mécanismes de propagation de l'information. Dans le cadre d'applications telles que le suivi de rumeurs ou la détection de buzz, les modèles utilisés pour simuler les dynamiques informationnelles doivent être en mesure de reproduire différents comportements de publication des sites.

Dans cet article, nous décrivons un modèle théorique de propagation de l'information possédant cette propriété. Dans un premier temps (Section 2), nous étudions un réseau réel extrait du Web pour catégoriser les comportements de publication des sources. Nous présentons ensuite dans la section 3 le modèle proposé, défini comme une extension du modèle ZC (Goetz et al., 2009), basée sur l'introduction de paramètres complémentaires. Dans la section 4, nous validons ce modèle en montrant comment les paramètres peuvent être déterminés pour générer un réseau ayant les mêmes caractéristiques que le réseau réel.

2 Identification des comportements de publication

Extraction d'un réseau de sources Les données réelles utilisées dans l'étude sont extraites du Web par une méthode de crawling dont l'objectif est de collecter tous les articles publiés par un ensemble de sources sélectionnées par l'utilisateur, et d'en extraire les liens entre sources

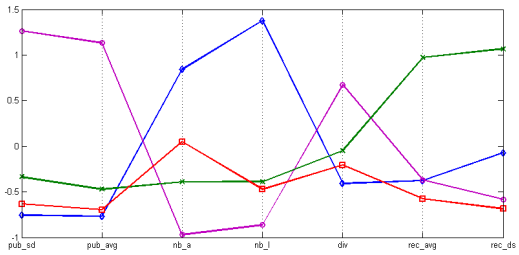

FIG. 1 – *Visualisation des centroïdes en coordonnées parallèles pour les données réelles*

définis par les liens hypertextes (Nel et al., 2009). Afin d'obtenir un réseau cohérent et complet, l'extension du crawling n'est pas récursive sur les liens extraits, mais contrôlée par l'utilisateur qui décide des sources à ajouter. L'extraction des articles est basée sur des heuristiques qui identifient le corps du texte et de n'extraient que les liens pertinents : cette étape supprime les bandeaux publicitaires, les commentaires des utilisateurs et d'autres informations non pertinentes.

Attributs considérés Nous utilisons sept attributs pour caractériser les habitudes de publication et notamment la façon dont un site sélectionne ses propres sources d'information : (1) la moyenne (notée pub_avg) et (2) l'écart type des *intervalles de publication* (pub_sd), (3) le *nombre d'articles publiés* (nb_a) et de (4) *liens* (nb_l), (5) la *diversité* des sources citées (div), définie comme le rapport entre le nombre de sources différentes et le nombre total de liens cités, (6) la moyenne (rec_avg) et (7) l'écart type (rec_ds) du *caractère récent* d'un article cité, défini comme la différence entre la date de publication de l'article considéré et la date de publication de l'article cité.

Classification non supervisée des données L'identification des comportements de publication est alors basée sur une méthode de *stacked clustering* (Kuncheva et Vetrov, 2006) à la fois robuste et ne nécessitant aucune connaissance a priori sur les résultats attendus. Elle applique d'abord les k-moyennes avec différentes valeurs de k pour, dans un premier temps, choisir le nombre de clusters, puis pour définir un codage des données sous la forme d'une matrice de similarité qui dépend du nombre de partitions dans lesquelles les points sont affectés à un même cluster. La partition finale est obtenue par application du clustering hiérarchique à cette matrice.

Résultats La base utilisée pour l'identification expérimentale des comportements de publication contient 190000 articles et 140000 liens pour 110 sites Web d'information généralistes, crawlés quotidiennement entre février et novembre 2009. L'étape de clustering produit 4 clusters dont les centroïdes sont représentés en coordonnées parallèles sur la figure 1 qui indique pour chaque cluster la valeur moyenne de chaque attribut, normalisé pour avoir une moyenne de 0 et un écart-type de 1. On peut établir la caractérisation suivante.

Le 1er cluster (losanges bleus) est petit et caractérisé par un intervalle de publication faible et un grand nombre d'articles. Le grand nombre de liens et la faible mesure de diversité sont

cohérents avec le fait que ce cluster est principalement composé de blogs spécialisés (techcrunch, pcworld). Le 2ème cluster (ronds violets) est petit et caractérisé par des intervalles de publication très élevés, un nombre de liens faible et la mesure de diversité maximale. Celle-ci est cependant peu représentative car ces sites ne publient que rarement et n'utilisent que peu de liens. Ces sites semblent être encore plus spécialisés (kassandre, laquadrature) ce qui peut expliquer leur faible activité en termes de publication.

Les 3ème (croix vertes) et 4ème (carrés rouges) clusters ont des tailles supérieures (27% et 56% des données initiales respectivement) et ont globalement les mêmes caractéristiques (faibles intervalles de publication et nombre moyen de liens) ; ils se différencient sur le caractère récent des articles cités. Une étude détaillée de leur composition montre que le 3ème contient des sites probablement plus familiers des outils spécifiques de plateformes de publication (blogs ou sites de publication collaborative), comme rue89 ou googleblog, alors que le 4ème cluster contient principalement des versions Web de journaux papier (timesonline, le monde, el pais). Cette interprétation est confirmée par la très haute mesure de réactivité du 3ème cluster et le fait qu'il utilise plus de liens hypertextes par article que le 4ème.

3 Formalisation de la propagation de l'information

La propagation de l'information est liée au thème de la contagion (Dodds et Watts, 2004) et a fait l'objet de diverses modélisations (Gruhl et Liben-Nowell, 2004; Java et al., 2006). Nous proposons une extension du modèle ZC introduit par Goetz et al. (2009) pour simuler de façon intuitive et adaptable la publication sur les blogs. L'extension proposée est conçue pour reproduire les caractéristiques de chacune des 4 catégories identifiées, par l'introduction de paramètres additionnels. Nous décrivons les étapes du processus de publication, ainsi que leur modélisation, illustrées sur la figure 2 qui montre également les paramètres proposés.

Publication d'un article La première étape du processus de publication définit si à un instant t, un site Web publie un article. Le modèle ZC utilise pour cela un critère de franchissement du zéro sur une marche aléatoire. Nous introduisons 2 paramètres pour distinguer des habitudes de publication différentes : le *pas* de temps permet de définir la fréquence de changement de la valeur de la marche aléatoire, il est utilisé pour représenter la disponibilité des personnes qui mettent à jour le site. La *borne* définit les valeurs maximales et minimales que peut prendre la marche aléatoire et donne un contrôle sur les périodes de non publication d'une source : elle permet de simuler la possibilité qu'une source ne publie pas pendant une longue période de temps.

Publication d'un lien Le modèle ZC définit P_l comme la probabilité qu'un article publié contienne un lien. Par extension, nous donnons la possibilité à un article de citer plusieurs liens avec une probabilité décroissante avec le nombre de liens : si $(k-1)$ liens sont cités dans l'article, la probabilité de citer un k-ième lien est P_l/k.

Choix d'une source Pour sélectionner la source pointée par un lien, le modèle ZC combine deux stratégies, appelées *exploitation* et *exploration*. Nous proposons de les faire dépendre de paramètres complémentaires, la *visibilité* v_a d'un article a et la *réactivité* r_A d'un site A : la

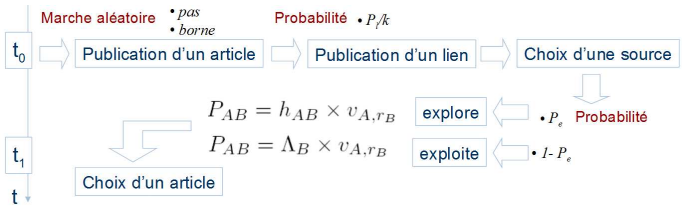

FIG. 2 – *Modèle de propagation de l'information.*

première reflète l'intérêt potentiel qu'un lecteur a pour son contenu informationnel ; la seconde est définie comme l'ancienneté au-delà duquel A se refuse à citer un article. La visibilité d'une source A selon un site B, v_{A,r_B} est alors définie comme la valeur maximale de visibilité des articles de A répondant au critère de réactivité du site B : formellement, en notant $t_{pub}(a)$ la date de publication de a, on a $v_{A,r_B}(t) = \max\{v_a, a \in A \; et \; (t - t_{pub}(a)) \leq r_B\}$

Dans le mode d'exploitation, qui a une probabilité $1 - P_e$, le site fait référence à une source qu'il a déjà citée auparavant. Pour deux sites A et B, on note h_{AB} le nombre de liens cités par A et pointant vers B, qui permet d'identifier les sources d'information favorites de A. La probabilité que A choisisse B est alors : $P_{AB} = h_{AB} \times v_{B,r_A} / \sum_C h_{AC} \times v_{C,r_A}$

Dans le mode d'exploration, qui a une probabilité P_e, le site fait référence à une source qu'il n'a jamais citée. Nous proposons d'utiliser Λ_A l'*influence absolue*, ou la popularité, de A, interprétée comme la probabilité qu'un internaute trouve le site en errant aléatoirement sur le Web. La probabilité que A cite B est alors $P_{AB} = \Lambda_B \times v_{B,r_A} / \sum_C h_{AC} \times v_{C,r_A}$

Choix d'un article La dernière étape consiste à choisir précisément l'article a cité. La probabilité de choisir a est proportionnelle à sa visibilité v_a ; si l'article a déjà été choisi dans l'article à publier ou s'il ne convient pas à la contrainte de réactivité du site, cette probabilité est nulle. À cette étape, le processus de génération d'un article est alors complet et une valeur de visibilité v est attribuée aléatoirement au nouvel article.

4 Validation du modèle de simulation

Cette section présente les résultats d'une simulation effectuée afin de valider le modèle de propagation proposé et ses paramètres : l'objectif est, à partir d'un paramétrage de simulation choisi manuellement, de générer un réseau de sources dont les comportements soient similaires à ceux identifiés à la section 2. Pour cela, on applique le processus d'identification de comportements aux données générées par le modèle et on compare la partition obtenue, ainsi que sa caractérisation, à celles que l'on cherche à reproduire (représentées sur la figure 1).

Paramétrage de la simulation La simulation est paramétrée de façon à reproduire les 4 types de comportement identifiés à la section 2, en définissant 4 classes distinctes, dont la proportion est identique à celle identifiée dans le corpus réel. Ainsi les 100 sites générés sont répartis en groupes de 8, 9, 28 et 55. Chaque classe est associée à une plage de valeurs des paramètres de simulation , selon leur sémantique définie à la section 3. Ainsi le paramètre P_l qui

	cluster 1	cluster 2	cluster 3	cluster 4
borne	1-5	5-10	3-6	1-5
pas	1-5	5-10	1-5	1-5
P_l	0,8-1	0,2-0,5	0,4-0,6	0,4-0,6
P_e	0,4-0,6	0,6-0,8	0,4-0,5	0,4-0,5
r	150-400	0-200	350-500	0-200

TAB. 1 – *Plages de valeurs des paramètres de simulation.*

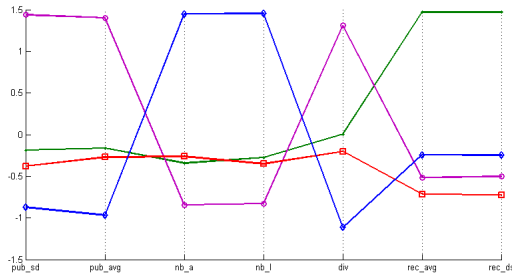

FIG. 3 – *Visualisation des centroïdes en coordonnées parallèles pour les données simulées.*

influence directement le nombre de liens publiés, nb_l, est associé à une plage de valeurs élevées pour simuler les sites de la première classe. Le tableau 1 résume les valeurs de paramètres choisies pour la simulation.

Validation des comportements de publication On peut tout d'abord comparer la partition obtenue par stacked clustering sur les données simulées à la partition attendue d'après l'appartenance fixée aux quatre classes. La valeur de l'indice de Rand ajusté est de 0.63, ce qui est satisfaisant compte tenu du fait de l'attribution manuelle des paramètres. Le tableau 2 donne la matrice de confusion entre ces deux partitions et permet de mieux apprécier les clusters les plus modifiés : ainsi on remarque que les clusters de petite taille (clusters 1 et 2) se sont agrandis, en particulier le cluster 1 est très modifié. Les plus gros clusters se retrouvent presque inchangés.

La figure 3 montre les centroïdes de chaque cluster pour les données simulées projetées sur les attributs utilisés pour caractériser les données réelles. Cette visualisation est à comparer avec celle obtenue sur les données réelles (figure 1). On observe que les valeurs des variables observées sont semblables et que l'on retrouve les comportements de publication dans les caractéristiques globales de chaque cluster.

5 Conclusions et perspectives

Nous avons proposé un modèle de propagation de l'information capable de générer un réseau de citations entre sources en reproduisant avec flexibilité des comportements de pu-

classe obtenue →		1	2	3	4
	1	3	0	1	4
classe	2	0	8	1	0
attendue	3	0	1	27	0
	4	8	3	0	44

TAB. 2 – *Matrice de confusion.*

blication observés sur des données réelles. La validation préliminaire effectuée par le biais d'une simulation paramétrée manuellement donne des résultats satisfaisants : on retrouve les 4 catégories de comportements identifiées expérimentalement sur des données réelles. Des travaux en cours visent à apprendre automatiquement la correspondance entre les paramètres du modèle de propagation et les valeurs des attributs caractérisant les comportements.

Une autre perspective vise à appliquer le modèle de propagation proposé à la problématique générale de détection de phénomènes d'amplification ou de rumeurs. En effet, il utilise des paramètres dont la sémantique est cohérente et intuitive, ce qui facilite la définition de mesures permettant de formaliser ces phénomènes. Il devient possible de les générer artificiellement par simulation et d'étiqueter des données. Les corpus résultants constituent des bases de tests sur lesquelles des algorithmes de détection peuvent être évalués et validés.

Références

Dodds, P. S. et D. J. Watts (2004). Universal behavior in a generalized model of contagion. *Physical Review Letters 92*(21).

Goetz, M., J. Leskovec, M. Mcglohon, et C. Faloutsos (2009). Modeling blog dynamics. In *International Conference on Weblogs and Social Media*.

Gruhl, D. et D. Liben-Nowell (2004). Information diffusion through blogspace. In *Proceedings of the International Conference on World Wide Web*, pp. 491–501. ACM Press.

Java, A., P. Kolari, T. Finin, et T. Oates (2006). Modeling the spread of influence on the blogosphere. In *Proceedings of the International Conference on World Wide Web*.

Kuncheva, L. I. et D. P. Vetrov (2006). Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE TPAMI 28*, 1798–1808.

Nel, F., A. Carré, P. Capet, et T. Delavallade (2009). Detecting anomalies in open source information diffusion. In *IST087 NATO Symp. on Information management and Exploitation*.

Summary

We propose a model of information propagation, describing every steps leading to its design and its use in simulation. From real data extracted from the Web, we identify distinct publishing behaviours inherent to each source. We then extend an existing model of information diffusion, so as to increase its expressive power and in particular to reproduce these publication habits. We then validate the proposed model on a simulation example.