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