Information Propagation on the Web: Data Extraction, Modeling and Simulation

F. Nel[*+] and **M.-J. Lesot**[*]
[*] LIP6 - UPMC
4 place Jussieu
75252 Paris cedex 05, France

P. Capet[+] and **T. Delavallade**[+]
[+] Thales Land and Joint Systems
160, boulevard de Valmy
92704 Colombes Cedex, France

Abstract

This paper proposes a model of information propagation mechanisms on the Web, describing all steps of its design and use in simulation. First the characteristics of a real network are studied, in particular in terms of citation policies: from a network extracted from the Web by a crawling tool, distinct publishing behaviours are identified and characterised. The Zero Crossing model for information diffusion is then extended to increase its expressive power and allow it to reproduce this variety of behaviours. Experimental results based on a simulation validate the proposed extension.

1 Introduction

The study of information diffusion on the Web is based on models of the information propagation mechanisms. Many are transpositions of biological models of disease propagation (Gruhl et al. 2004; Adar and Adamic 2005; Kempe, Kleinberg, and Tardos 2005; Java et al. 2006; Leskovec, Adamic, and Huberman 2006), others propose to imitate the process of information publishing in an intuitive and adaptable way, as the *zero-crossing* (ZC) model, based on random walks (Goetz et al. 2009).

In this paper we propose a general information propagation model that goes beyond blogs, to the Web in general, following an imitation principle based on an extension of the ZC model. It aims at reproducing an observed variety of publishing behaviours, e.g. taking into account the differences between blogs and journal websites. Publishing behaviours are defined both in terms of general characteristics, e.g. the global publication amount, and more specifically in terms of citation policies, e.g. use and diversity of hyperlinks.

To that aim, as described in Section 2, we first identify and characterise four distinct and contrasted publication and citation patterns, from the study of a source network extracted from the Web by a crawling tool. Section 3 then presents the proposed model, defined as an extension of the ZC model: it describes additional parameters and their semantics, proposed to allow it to reproduce this variety of behaviours. Section 4 lastly presents the simulation performed to validate the proposed model through the comparison between the characteristics of the generated and the real networks.

2 Identification of Publishing Behaviours

2.1 Data Acquisition: Source Network Extraction

The publication behaviours we are interested in are defined in terms of general publication habits, e.g. the global publication amount, and in terms of citation policies derived from hyperlinks relations between the considered sources. Therefore the studied corpus is made of a source network, whose edges represent hyperlink citations.

This network is extracted from real data, using the user-controlled Web crawling strategy and data cleaning method proposed in (Nel et al. 2009). It consists in gathering all articles published by sources automatically selected under a user control and identifying the relevant hyperlinks they contain, after the removal of commercial links, those being part of the website internal browsing structure, as well as those in the article comments.

We apply this methodology to 110 generalist information websites, including Web versions of traditional daily newspapers, specialised blogs, blog platforms and collaborative publishing tools (the list can be read from the leaves of the dendogram shown on Figure 1). From a daily crawling performed between February and November 2009, we collected a database containing 190,000 articles and 140,000 links.

2.2 Data Processing: Clustering Step

To characterise the publication habits of the considered sources, we use the following 7 descriptors: we take into account their publication amount, measured by their *number of published articles* (denoted nb_a) and the average and standard deviation of the *publication intervals* (resp. pub_avg and pub_sd). The particularities of publication on the Web are measured through the *total number of hyperlinks* included in their articles (nb_l). Lastly we consider the diversity and the recentness of their citations: *diversity* (div) is defined as the number of different cited sources divided by the overall number of cited links; the average and standard deviation of the *recentness* of the cited articles (resp. rec_avg and rec_ds) are derived from the intervals between the publication dates of a considered article and the articles it cites.

The identification of the publication behaviours is then obtained by clustering these data applying the robust stacked clustering method (Kuncheva and Vetrov 2006).

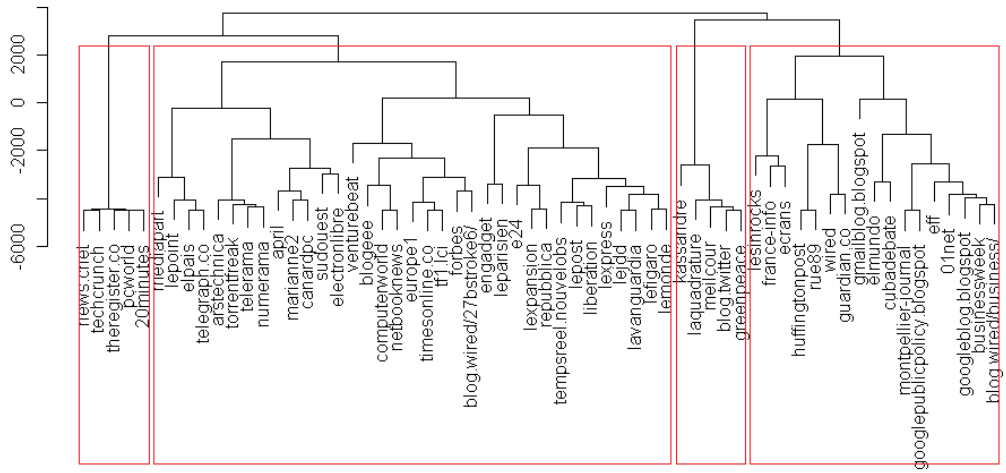

Figure 1: Dendogram of the hierarchical clustering with a cut at four clusters, obtained from the corpus described in Section 2.1.

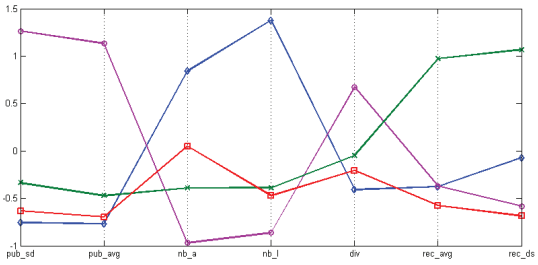

Figure 2: Identified publication behaviours: cluster centroids in parallel coordinates.

2.3 Results: Obtained Publication Behaviours

An optimal number of 4 clusters is obtained for the corpus described in Section 2.1, and for an independent corpus built in the context of a competitive intelligence analysis in the defence area. This suggests that the value of 4 distinct publication behaviours is not specific to the considered corpus.

Figure 1 shows the dendogram and its partition in 4 clusters. To characterise the corresponding publication behaviours, Figure 2 shows the cluster centroids in parallel coordinates, scaled for each descriptor to have mean 0 and standard deviation 1.

The first cluster (blue diamonds) contains 8% of the sources, characterised by a very high number of links, a high number of articles, correlated to a low publication interval, and a low diversity measure. These characteristics are consistent with the fact that it is mainly composed of specialised blogs (techcrunch, pcword).

The second cluster (violet circles) contains 8% of the sources too, whose behaviour is characterised by a high publication interval and a low number of links. They are even more specialised than the first cluster (kassandre, laquadrature), which can explain their low publishing activity.

The third cluster (green crosses) has a middle size (27% of the sources) and is mainly characterised by its high reactivity, shown by the high *recentness*; it shares its properties of low publication interval and medium number of links with the 4th cluster. It contains websites probably familiar with Web specific tools like blog platform or collaborative publishing (rue89, googleblog). This explains its high reactivity measure and the fact that even with a smaller publication frequency, it uses more hyperlinks than the 4th cluster.

The fourth cluster (red squares) that groups more than half the sources (56 %) has a rather high number of articles and a low reactivity. It mainly contains Web versions of existing traditional newspapers (e.g. timesonline, lemonde).

As a summary, the publication behaviours are mainly distinguished according to the publication frequency, the ability to exploit Web publication particularities and the source diversity. They can be characterised as i) very active specialised blogs with low publication interval and high number of links, ii) very specialised blogs with low publishing activity, iii) general websites familiar with collaborative publishing and iv) Web version of existing newspapers.

3 Proposed Formalisation of the Information Propagation Process

This section proposes a formalisation of information propagation capable of reproducing the previous variety of publication behaviours. To that aim, we extend the zero-crossing, ZC, model (Goetz et al. 2009) introduced to generate a blogosphere by imitating the process of information publishing in blogs. The proposed enrichment is based on additional parameters to increase its flexibility and its expressive power. We recall the initial definition of each step of the ZC model, and present the proposed parameters together with their semantics and expected effects, as summarised on Figure 3.

3.1 Article Publication

To decide whether at time t a website publishes, the ZC model uses a zero crossing criterion on a random walk: for each website, at each time step, the random walk value is in-

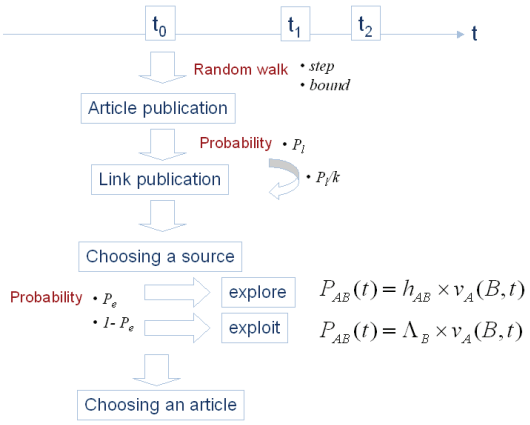

Figure 3: Proposed extended model.

cremented or decremented with probability 0.5. Whenever it equals zero, the website publishes an article.

We introduce a parameter called *step* defining the change frequency of the random walk value: by default, it equals one and the random walk value is updated at each time step. This parameter thus determines the maximum publication frequency.

A second parameter called *bound* controls the maximum and minimum value of the random walk: when the random walk value reaches the maximum (resp. minimum), it is automatically decremented (resp. incremented) at the next step. In the random walks theory, such a constraint, called *reflecting boundary*, is used to represent phenomena on finite domains bounded by barriers (Khantha and Balakrishnan 1985). In our model, it controls the non-publishing period of a source: it simulates the maximum symbolic distraction boundary that a blogger cannot cross, and consequently, the possibility of a source not to publish during a long period of time. For example, it can be applied to differentiate the publishing behaviours of an information website with daily constraints and a more-or-less active blogger.

3.2 Link Publication

When an article is published, the ZC model defines P_l the probability for it to contain a link. We propose to extend this choice through the possibility to cite several articles instead of a single one. We define the associated probability as decreasing with the number of links: for any $k \geq 1$, if $(k-1)$ links are cited, the probability to cite a k-th link is P_l/k.

3.3 Cited Source Choice

Once a link is chosen to be included, the next step is to select the source it points to. The ZC model proposes two strategies: in the *exploitation* mode, selected with probability P_e, the website uses a Web source it has already cited in the past; in the *exploration* mode, selected with probability $1 - P_e$, it cites a Web source it has never used before. In the first case, the choice is made according to the number of past links it has already made to each source and in the second case, according to the total number of in-links of the source.

We introduce the *reactivity* of a website A, r_A, that measures the limit of oldness that A admits to cite an article, and the *visibility of an article* a, v_a, that reflects the potential interest of a reader for its informational content. v_a does not depend on its source, its layout or recentness but evaluates the extent to which a stands out from other articles in terms of the attractiveness of its content. It thus differs from the *topic visibility* (Kiefer, Stein, and Schlieder 2006) that is used to describe topics instead of articles and is based on hit counts of a search engine for the search of the topic term.

The article visibility v_a is then extended at the website level: $v_A(B,t)$, called *source visibility*, represents the visibility, or attractiveness, of B at time t for A, indicating the extent to which B appears important to A. It measures the compatibility between A reactivity and B article recentness: denoting $t_{pub}(a)$ the publication date of a, it is computed as $v_A(B,t) = \max\{v_a, a \in B \text{ and } (t - t_{pub}(a)) \leq r_A\}$.

We then propose to use the source visibility to weight the probabilities for A to cite B: in the exploitation mode, denoting h_{AB} the number of links cited by A and pointing to B, we define $P_{AB}(t) = (h_{AB} \times v_A(B,t))/(\sum_C h_{AC} \times v_A(C,t))$. In the exploration mode, $h_{AB} = 0$ and it is replaced with B *absolute influence*, Λ_B, that represents the probability for a reader to find B when wandering randomly on the Web. The probability for A to choose B is then $P_{AB}(t) = (\Lambda_B \times v_A(B,t))/(\sum_C \Lambda_C \times v_A(C,t))$. This definition is valid if A has never cited B, otherwise $P_{AB}(t) = 0$.

3.4 Cited Article Choice

Finally, the precise article that is cited is chosen. Instead of using the number of in-links and the publication date of the article, we propose, in a similar way, to define the probability to choose an article a as proportional to its visibility v_a. If the article has already been chosen (in the case of multiple citation links) or if it does not fit the reactivity condition of the website, this probability is zero.

At this point, the generation process of an article is complete and a visibility value is randomly assigned to the newly published article, as an integer between 1 and 100.

4 Validation of the Proposed Model

This section discusses the validity of the proposed model, evaluated through its ability to generate a network whose characteristics are similar to that of a real network, i.e. its realism: we show how the "low level" parameters of the model make it possible to reproduce the "high level" descriptors extracted from the real data (see Section 2). We thus implement the proposed model and perform simulations. We then apply to the generated network the analysis described in Section 2 and show that the obtained partition and its characterisation are similar to the ones obtained with the real data.

4.1 Simulation run

The number of sources is set to 100. We allow ourselves to exploit the known proportion of sources of each type: the sites are splitted in groups of 8, 9, 28 and 55. For each source, we must set the parameters of the proposed information propagation model. To that aim, we propose to define 4

	class 1	class 2	class 3	class 4
bound	1-5	5-10	3-6	1-5
step	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

Table 1: Value ranges of the simulation parameters.

classes to model the 4 identified publishing behaviours. We associate each class with a range of possible values for each parameter, as indicated in Table 1. A source is then assigned values randomly drawn from the intervals, and a random absolute influence value.

We define the intervals manually, based on the semantic of each parameter as discussed in Section 3. For example, the probability P_l that an article contains one link influences the descriptor *number of published links*. We thus give it a high value to simulate sites belonging to the first cluster. Similarly P_e obviously influences the diversity descriptor.

The simulation consists in setting a time range, and applying, for each source at each time step, the process proposed in Section 3, determining whether it publishes an article, and if it does, whether it introduces one or several hyperlinks, and if it does, what are the cited sources and articles.

4.2 Comparison of the Real and Simulated Data

We then apply the clustering process presented in Section 2: the validation consists in studying the composition of the obtained partition.

First its consistency with the classes defined to set up the simulation parameters is measured: it is compared to the partition awaited according to the membership to the 4 classes defined in Table 1. The Rand index equals 0.628 which is satisfactory considering the manually setting of the parameters. The confusion matrix (not shown due to space constraint) shows that the biggest clusters are almost unchanged, whereas the small ones become somewhat bigger.

Second, the centroids derived from the real and the simulated data can be compared, from Figures 2 and 4. It can be observed that the values from the generated data are similar to the ones from the real data and that the global characteristics of each cluster meet the publishing behaviours, even if they appear to be more extreme.

5 Conclusion

We proposed an information propagation model presenting the ability to reproduce various publishing behaviours. The first validation carried out through a manually set up simulation shows the relevance of the generation process and choice of low level parameters that make it possible to indirectly build an artificial source network reproducing the publishing behaviours experimentally identified from real data.

The proposed model relies on manually set parameter ranges. Future works aim at applying machine learning algorithms to automatically obtain their ranges according to the values of the attributes describing the behaviours. The

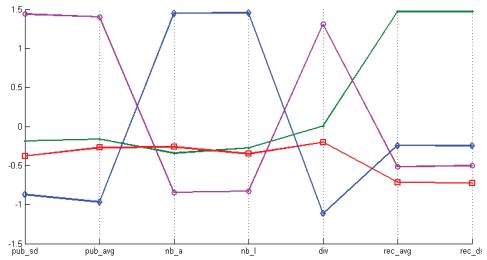

Figure 4: Simulated publication behaviours: cluster centroids in parallel coordinates.

absolute influence parameter could also be updated automatically, e.g. by a PageRank-like algorithm.

Another perspective aims at applying the proposed propagation model to the general issue of information propagation, exploiting the implemented tool to perform realistic simulations with parameters having a coherent and intuitive semantic. It therefore becomes possible to study e.g. amplification phenomena or rumours, artificially initiating or recreating anomalies in the informational dynamics.

References

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *Proc. of the 2005 IEEE/WIC/ACM Int. Conf. on Web Intelligence*, 207–214.

Goetz, M.; Leskovec, J.; Mcglohon, M.; and Faloutsos, C. 2009. Modeling blog dynamics. In *Int. Conf. on Weblogs and Social Media*.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *Proc. of the Int. Conf. on World Wide Web*, 491–501.

Java, A.; Kolari, P.; Finin, T.; and Oates, T. 2006. Modeling the spread of influence on the blogosphere. In *Proc. of the Int. Conf. on World Wide Web*.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2005. Influential nodes in a diffusion model for social networks. In *Proc. of the Int. Coll. on Automata, Languages and Programming*.

Khantha, M., and Balakrishnan, V. 1985. Reflection principles for biased random walks and application to escape time distributions. *Journal of Statistical Physics* 41:811–824.

Kiefer, P.; Stein, K.; and Schlieder, C. 2006. Visibility analysis on the web using co-visibilities and semantic networks. In *Semantics, Web and Mining*, LNCS 4289. 34–50.

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

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2006. The dynamics of viral marketing. In *Proc. of the 7th ACM Conf. on Electronic commerce*, 228–237. ACM.

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