

Describing the Web in less than 140 Characters

Stéphane Raux

LIAFA and Linkfluence
153 Boulevard Anatole France,
93521 Saint-Denis, France
stephane.raux@liafa.jussieu.fr

Nils Grünwald

Linkfluence
153 Boulevard Anatole France,
93521 Saint-Denis, France
nils.grunwald@linkfluence.net

Christophe Prieur

LIAFA, University Paris-Diderot
175 rue du Chevaleret,
75013 Paris
prieur@liafa.fr

Abstract

Links curation, *i.e.* finding relevant information within the World-Wide-Web and its ever-growing amount of content is a crucial problem for information access. Hyperlinks recommendation has been for a long time a common way to share references between web users, be it by e-mail exchanges, instant messages or forums. We explore in this paper how Social Media extend this recommendation practice by focusing on the citation of hyperlinks on Twitter. We investigate how people deal with the strong limitation of 140 characters per message, showing that this constraint encourages people to perform a good synthesis of the content they are linking to. We take advantage of this practice to efficiently cluster the actual content of the linked pages with an algorithm based on lexical proximities between messages. Our method yields topical clusters that are consistent with the dynamics of user interests with no need to extract text from the pages themselves.

1 Introduction¹

From a user point of view, sharing hyperlinks is a common way to recommend online resources, as they can be sent with little effort (if any), in an email for instance. This link-sharing practice has flourished with the rise of platforms providing more and more cross-users interactions, making Social Media highly reactive to events by producing a constant flow of recommendations. In this regard, services like Twitter are now key pipes in the online information flow, probably less because of the actual content they provide than because of their essential role providing pointers to external content.

Our hypothesis is that Twitter can be used as an efficient index of Web content: the limitation of 140 characters per message forces users to be very concise. Moreover, as most of the time the quoted url are “shortenized” by dedicated services (*e.g.* `bit.ly` or `tinyurl.com`), which provide a short opaque alias for a given url, Twitter users usually summarize the content they are linking in a few words. We take advantage of this behaviour to propose a method which

aggregates the linked content in order to discover emerging topics. The main interest of this approach would be to make content recommendations based on Twitter data. For a given url, it would provide the topic it is about and links to other contents pointing to similar issues.

After presenting related works done on this platform, we will present a method to identify emerging topics as clusters of linked urls, only relying on the texts of tweets recommending them. We show in Section 4 how this method, applied to a corpus of more than 1,500,000 tweets, brings remarkably stable and consistent results. We then conclude with a qualitative discussion of the typology of detected topics (as labelled url clusters).

2 Related works

Recent work on Twitter emphasizes the coexistence of two main types of users: *meformers*, who focus more on social relationships and tend to interact with friends, and *informers*, whose main activity is to share content (Naaman et al. 2010). In the same way, (Cha, Haddadi, and Gummadi 2010) compare the mechanisms of information and popularity and show that this distinction holds also on a large-scale point of view: while celebrities are top ranked by the number of direct messages, information providers are top ranked by the number of *retweets* (message forwarding). In this paper, we will focus on this second category of users. As Cha *et al.* also found that 92 % of retweeted messages contain a url, we can safely postulate that only using tweets containing a url will not lead us to miss important contents.

(Boyd, Golder, and Lotan 2010) show that the *retweet* behaviour is a key user-powered feature in the diffusion of information, and (Lerman, Ghosh, and Rey 2010) consider *retweets* on Twitter as votes by comparing them with votes on Digg. We rely in our approach on the number of tweets and retweets to determine the importance of our clusters. However, as we consider that the semantic relevance of a tweet is mostly independent of its number of retweets, we build our clusters based solely on unique text.

Some authors have analyzed tweets content in order to predict real-world outcomes, for box-office (Asur and Huberman 2010) or public polls (O’Connor et al. 2010). These approaches often involve machine learning and sentiment analysis to provide insights on future outcomes from a training corpus. Other studies focus on analyzing cycles of news

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹This work has been partially supported by the French National Agency of Research (ANR) through grant “Webfluence” #ANR-08-SYSC-009.

activity, by retrospectively tracking memes. (Leskovec, Backstrom, and Kleinberg 2009) have applied this idea on a blog corpus by clustering quotations using graph techniques. Using Twitter, we pursue a related goal, but we build our clusters based on lexical proximities instead of using variants of the same quote.

The problem of the classification of urls by topic has been studied by (Baykan et al. 2009), who use machine-learning techniques, and more recently by (Blanco, Dalvi, and Machanavajjhala 2011), who propose a more scalable algorithm. Both approaches rely only on the urls of the pages: the extraction and the analysis of the full content of the pages is very resource-consuming, and the webpages themselves can change or be removed, preventing from analyzing retrospectively a given set of urls. By using the content of the tweets and not the urls themselves, we propose another way of classifying urls relying only on users feedback.

3 A method to detect “Hot Topics”

In this section, we formally define our algorithm for clustering the urls by similar topics. We first model our set of tweets as a bipartite graph of terms and urls, then we propose an original projection of this graph and proceed to a simple clustering.

Bipartite Graphs

A (directed) graph is a pair $G = (V, E)$, where V is the set of *vertices* and $E \subseteq V \times V$ the set of *edges*. One denotes by $N(v) = \{u \in V, (u, v) \in E\}$ the *neighbourhood* of a vertex v , and calls its elements the *neighbours* of v . The number of vertices in $N(v)$ is the *degree* of v : $d(v) = |N(v)|$.

When the set V can be split into two disjoint sets \top (*top* vertices) and \perp (*bottom* vertices) such that all edges are in $\top \times \perp$, then G is said to be *bipartite*. One may then denote G as (\top, \perp, E) . The \perp -*projection* of G is the graph (\perp, E_{\perp}) where two vertices of \perp are linked together if they share a common neighbour in G :

$$E_{\perp} = \{(u, v), \exists x \in \top : (u, x) \in E \text{ and } (v, x) \in E\}.$$

The \top -*projection* of G is defined in the same way, by inverting \top and \perp .

The \perp -projection of G can be weighted by defining a function ω_{\perp} on E_{\perp} in the following way, for any $(u, v) \in E_{\perp}$:

$$\omega_{\perp}(u, v) = \sum_{x \in N(u) \cap N(v)} \frac{1}{d(u)} \times \frac{1}{d(x)}.$$

This weight function is actually the probability to reach v in two hops from u .

If G itself is weighted with a function ω on its edges with real values, then one can generalize the definition of the weight function on E_{\perp} as follows:

$$\omega_{\perp}(u, v) = \sum_{x \in N(u) \cap N(v)} \frac{\omega(u, x)}{k(u)} \times \frac{\omega(x, v)}{k(x)},$$

where $k(u)$ is the *weighted degree* of a vertex u , defined as the sum of the weights of its incident edges: $k(u) = \sum_{v \in N(u)} \omega(u, v)$.

A detailed survey in the context of social networks can be found in (Latapy, Magnien, and Del Vecchio 2008).

Clustering a Url Graph

We work with a dataset of tweets mentioning (unshortened) urls and thus consider a bipartite graph $G = (\top, \perp, E)$ where \top is a set of words quoted in the tweets, \perp the set of urls and E the set of pairs (w, u) such that the word w appears in a tweet mentioning the url u . This graph is weighted with a weight function ω whose value on a pair (w, u) is the *TF.IDF* score of the word w within all the unique words in the tweets linked to the url u , filtered by a blacklist of common empty words in French and English.

Now the weighted \perp -projection of G defined as above is a weighted directed graph of urls where $\omega_{\perp}(u, v)$ expresses to what extent url u is described with the same words as url v . We call *hot topics*, clusters of urls in the graph built in the following way.

After removing self loops from G_{\perp} , we use a classical *Union-Find* algorithm and a so-called *specificity function* on each cluster. Initially each url is a cluster by itself, and therefore very specific. Then each edge $(u, v) \in E_{\perp}$ is checked in descending order of edges weights. If u and v do not belong to the same cluster and merging their two clusters does not produce a cluster with a specificity lower than a given threshold λ , then we merge them.

We used as specificity function the number of words linked to all the urls of a given cluster. More formally, it is the size of the intersection of the neighbourhoods (in the bipartite graph) of all the urls in the cluster. All the results given in this paper have been obtained with a threshold of 2 (all clusters have at least two words in common).

4 Validation

The following methodology has been used to validate the algorithm. We used a set of 11,258 websites which was qualitatively selected to be a consistent sample of the most active content-producing websites on the French-speaking Web. They were chosen, categorized and labeled through automated crawling, topological analysis and manual selection by documentalists in 19 thematic categories, in such a way that websites belong to the same category if they show preferential attachment and semantic coherence.

Among these categories, we selected three categories rather strongly influenced by the news and therefore good candidates for trends and events detection: *Tech. Lovers* (blogs mainly publishing posts about new technologies), *Native Media* (news media which only exist on the Web), and *Citizenship* (people chiefly interested in political news and trends).

We captured every tweet containing at least one link to any website belonging to one of the three selected categories (see table 1) over a period of 6 months from June, the 14th 2010 to January, the 9th 2011, using the `backtweet.com` service which allows to find all recent tweets pointing to a url.

As we are interested in capturing and comparing transient burst of topics linking, we choose to work on the dataset at the week scale. After some tests, it was found to be a good

compromise to avoid the different activity levels between the days of the week while allowing to capture short-time events.

Category	Sites	Urls	Tweets
Tech. Lovers	666	141,280	1,054,575
Native Media	943	98,731	414,232
Citizenship	422	27,109	100,366

Table 1: General statistics for each category

Several treatments were applied to these urls in order to better aggregate the tweets linking the same urls. First, as we have seen previously, the use of shortenizers is widespread on Twitter, and thus the links extracted from the tweets need to be expanded before aggregating them. We did this by systematically following all redirections for every link in order to get the final url. Second, these urls are still often not ideal for aggregation, because many twitter clients add new specific parameters to the query and fragment parts of the url, for logging purposes. We use a set of heuristics to delete them and get a canonical version of each link.

Stability of the Algorithm

We first apply our clustering method to each category, then we study the distribution of the size of the clusters obtained by the algorithm. We focus on relative distributions in order to compare categories with very different sizes. We find high heterogeneity: most of the urls belong to small clusters of size 1 or 2, and a few belong to large clusters which may include tens of urls.

We compare the distributions obtained on each week and discover that the distributions are very stable during the whole period. Table 2 summarizes the mean value of the cumulative distributions for each category.

Category	> 1	> 2	> 5	> 10
Tech. Lovers	73.9 %	44.0 %	17.9 %	7.1 %
Native Media	66.3 %	36.7 %	13.0 %	4.6 %
Citizenship	57.5 %	26.5 %	9.7 %	3.7 %

Table 2: Mean value of the cumulative distribution of the size of the clusters, for each week of the period

Moreover, if we consider each week period, we observe that the mean proportion of single urls fluctuates between 26.1 % and 43.7 %, and that our algorithm is stable when clustering over a given category of sites. This is interesting, as there is no need to define a given number of clusters. This is not the case for a lot of other clustering methods like *k-means*, or with machine-learning techniques.

Lexical Consistency of the Clusters

What remains to be measured is whether the clusters obtained are indeed a good match for our stated objectives. To do so, we have downloaded the set of webpages linked during 3 full weeks of the tweets of our dataset. The main content of each page was then extracted using shallow text features (Kohlschütter, Fankhauser, and Nejd1 2010), and we used this text to compute the mean cosine similarity between

Category	1×	2×	5×	10×
Tech. Lovers	77.7 %	64.5 %	35.4 %	19.5 %
Native Media	77.0 %	64.7 %	38.8 %	25.4 %
Citizenship	86.3 %	63.3 %	25.2 %	18.4 %

Table 3: Percentage of the clusters of each category on the week from 2011-01-03 to 2011-01-09 with mean similarity at least N times the baseline

every page linked by a cluster of tweets. The extraction step was needed to reduce as much as possible the fake similarity induced by identical navigational content on the webpages from the same website. The links for which the extraction failed to produce content, or which were not text (images, videos, *etc.*), were simply dropped from the dataset for this experiment.

We postulate that clusters that relate to a single item of the agenda of a community should link pages sharing a significant part of their vocabulary. To verify this, on the same corpus of texts we compute a baseline similarity score for each week of activity, defined as the mean cosine similarity score between every pair of texts in the dataset, and compare it to the mean similarity score in the corpus (see table 3).

The results confirm our postulate: the mean similarity between contents linked by clusters is consistently and significantly higher than the baseline score in each dataset. Further measurements also show that the quality rises with the size of the clusters, with clusters of size 2 being the less reliable in terms of results. We observe a high standard deviation on small clusters which can be explained by several factors. High scores are common because of the widespread habit of lengthy quotations between blog posts, and the extensive use of cable releases to write news posts in the Native Media category. Some websites also allow access to the same content through different urls, which leads to perfect similarity scores when clustered together.

Volume of tweets and popular topics

In order to evaluate if our method brings new information for end users, we aggregated the number of tweets received by each url in order to measure the total number of tweets received by each cluster. As all the communities are dominated by a small number of clusters which get most of the attention, we checked whether these clusters are important just because they contain a popular url, or if some clusters are composed of many urls receiving few tweets each.

We called *top url*, the url receiving the most tweets in each cluster, and we computed the Spearman’s rank correlation coefficient between the number of tweets received by the cluster and the number of tweets received by these *top urls*. We observed that for all communities, this coefficient becomes very low if we only consider the top 10 % of the clusters. This shows that our algorithm is also able to put in light topics that are not promoted by highly followed actors, but which are nevertheless important enough to gather a large attention through many more modest contributions.

Id	Size	Tweets	Retweets	Specific Words	Description
#1	4	510	289	lepost guerre 4chan lol	4chan users at war with France
#2	3	295	212	wtf femme	Sexual harassment affairs
#3	13	195	117	noir jeudi	French NGO lobbying for low rents for students
#4	14	194	89	michael youn	French comic actor
#5	7	193	84	optunisia anonymous	Repression in Tunisia
#6	23	142	25	business net journal	Articles from business website
#7	7	134	38	facebook marchands	E-commerce solutions on Facebook
#8	3	134	83	rolex seguela	Controversal declaration by a well-known French personality
#9	4	129	71	pen marine	News about French far-right leader Marine Le Pen
#10	2	120	101	bcp semble pouvoir	Noisy cluster

Table 4: Top 10 clusters in tweets volume on the Native Media category from 2011-01-03 to 2011-01-09

5 Discussion

We will now explore qualitatively the “hot topics” detected for a given category and check their relevance. Table 4 presents the top 10 clusters on the Native Media category in the last week of our time period. For each cluster, it gives its rank and size, the total number of tweets linking one of its urls (this defines the rank) and how many of them are *retweets*. The so-called specific words are the words shared by all the urls of a cluster (recall that the clusters are built according to these words). Since they are not enough to understand what the clusterized urls are about, we added a quick manual summary to make them more understandable².

As can be seen from this list, the clusters topics are easy to recognize and summarize. Only one cluster, **#10**, is really noisy, because the words on which it was built are a French idiom and have no relation to any topic. Cluster **#6** has all of its links pointing to the same website, the name of which is present in all the tweets as a hashtag. This is the main source of noise in our clusters and some methods could be used to mitigate its effect, for example by preventing the aggregation of urls pointing only to the same website.

The other clusters are interesting and help define what exactly constitutes a “hot topic”. A first kind of clusters (**#1**, **#5**, **#8**) is event-driven and closely linked to the news. A second type of clusters is more related to the long-term agenda of the category. Clusters **#3** and **#7** illustrate this point, with the majority of their links going to broad and more reflexive papers. These clusters are more stable and long-lived than the previous ones, and can sometimes last several weeks.

The results are similar for the tech-lovers and the Citizenship categories, though each specific agenda colors the kind of clusters we obtain. Citizenship has clusters following closely the political news and events, and on the Tech. Lovers category each new product makes a neat cluster, while some clusters have broader signification, for example regrouping everything about the Apple App Store.

Some other categories from our larger research project were also explored and revealed interesting clusters with some specific types, for example the clusters of the Cooking category do not follow the news agenda but can be heavily influenced by seasonality or periodic events like holidays or celebrations.

²The whole lists of clusters along with the actual urls are available at <http://www.liafa.jussieu.fr/~raux/icwsm/>.

To summarize, our method provides url clusters that have qualitatively significant relevance, detecting both time-specific topics and long-range issues. Studying how these topics evolve will help following trends and allow to analyze attention transfers of Twitter users from one topic to another. Finally, keeping the focus on the users could help to better understand online community dynamics, for instance by comparing the evolution of the links between users and topics with links between users and groups in social network services like Facebook or Flickr.

References

- Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. In *Int. Conference on Web Intelligence and Intelligent Agent Technology*, 492–499.
- Baykan, E.; Henzinger, M.; Marian, L.; and Weber, I. 2009. Purely URL-based topic classification. *WWW '09*.
- Blanco, L.; Dalvi, N.; and Machanavajjhala, A. 2011. Highly Efficient Algorithms for Structural Clustering of Large Websites. In *WWW '11*.
- Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet , Tweet , Retweet : Conversational Aspects of Retweeting on Twitter. In *HICSS-43*.
- Cha, M.; Haddadi, H.; and Gummadi, K. P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.
- Kohlschütter, C.; Fankhauser, P.; and Nejdl, W. 2010. Boilerplate Detection using Shallow Text Features. In *WSDM*.
- Latapy, M.; Magnien, C.; and Del Vecchio, N. 2008. Basic notions for the analysis of large two-mode networks. *Social Networks* 30(1):31–48.
- Lerman, K.; Ghosh, R.; and Rey, M. 2010. Information Contagion : An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *ICWSM*, 90–97.
- Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Memetracking and the dynamics of the news cycle. *KDD '09* 497.
- Naaman, M.; Boase, J.; Lai, C.-h.; and Brunswick, N. 2010. Is it Really About Me? Message Content in Social Awareness Streams. In *CSC '10*.
- O’Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*, 122–129.