Intertemporal topic correlations in online media

A comparative study on weblogs and news websites

Jean-Philippe Cointet
CREA (CNRS/Ecole Polytechnique)
1, rue Descartes
F-75005 Paris, France
cointet@shs.polytechnique.fr

Emmanuel Faure
CREA (CNRS/Ecole Polytechnique)
1, rue Descartes
F-75005 Paris, France
faure@shs.polytechnique.fr

Camille Roth
Department of Sociology
School of Human Sciences
University of Surrey
GU2 7XH, United Kingdom
c.roth@surrey.ac.uk

Abstract

We address the issue of intertemporal topic correlations in a selection of online media consisting of political weblogs and press website content. We wish to investigate in which way various information sources may be correlated, therefore preceding and maybe influencing each other. We use hidden Markov modeling to exhibit dynamic relationships in topic occurrences between distinct groups of weblogs; thereby considering topic distributions over weblog groups as system states, looking for minimal causal states, and exhibiting their transition probabilities. Beyond behavioral correlations between some groups of blogs and online media, we also identify varied and richer types of inter-group patterns. In particular, using a very compact description, we could infer interpretations as to how diverse groups of blogs behave with respect to each other as regards raising and discussing issues.

Keywords

Weblog communities, temporal correlation detection, behavioral patterns, dynamic categorization, hidden Markov models, causal state machines, document mining, social cognition.

1. Introduction

As individuals raise issues, exchange viewpoints and argue about them in a globally distributed and networked fashion, part of this huge discussion forum has reached virtual arenas and social media as well — first through Internet news forums and more recently to weblogs, which are now a fast-growing and massive subpart of the phenomenon. Webloggers “post” dated articles in which they express current opinions and allegedly exhibit their present concerns.

As such, the blogosphere can be studied as a complex system of interrelated content with its own dynamics. On the other hand, one may also wonder whether this system exhibits features which are also apparently present in the larger social system it integrates: being obviously not isolated from the external world, some relationships or correspondences between the blogosphere and its wider environment could reasonably be expected.

More to the point, it has been observed that blog topics exhibit spikes (sudden topics, linked to a current event, or enjoying a limited-time interest) and chatter (more recurring, “long-term” topics) [3, 4] indicating strong correlation and even resonance phenomena in the system dynamics. Following this observation an appealing issue is that of describing how topics are evolving in a given weblogger community. In this direction, it may be particularly interesting to know whether there are synchronized groups of authors, i.e., agents discussing the same topics at similar times or with some delays, more broadly in an identical fashion. Put differently, while there allegedly are influences between bloggers (e.g. through hyperlinks between weblogs, be it direct [5] or indirect [1] influence) as well as from sources which are more or less external from a blogspace standpoint (online or offline news media [8], real-world acquaintances, etc...), we would be primarily willing to detect whether some communities of weblogger exhibit similar behavioral patterns in addressing issues with respect to each other — rather than, say, finding a social network through which information might navigate.

For instance, it is not unrealistic to expect that at least some weblogs will be strongly correlated to classical mass media content, especially to content found in the press at the same given period [8]. Besides, it is likely that the interest of some groups around some topics might trigger other groups to start discussing them. On the whole, our aim is to identify a *broad class of dynamic patterns on topic occurrences* — taking into account sources from inside and outside the blogosphere — in order, then, to study their intertemporal correlations.

To achieve this we hence propose a generalized framework based on causal state machines [2, 11], while our empirical case study will be focused on a selection of French political weblogs, in the context of the forthcoming presidential elections in Spring 2007. In Sec. 2 we present the empirical protocol. Section 3 is devoted to finding semantically-based weblog groupings for which we will, in Sec. 4, appraise and compare dynamic patterns of activity.

2. Empirical setting and protocol

We consider data telling us *which blog* is talking *of what* at *which time*, using a discrete-time crawl of weblog posts and assuming that terms used in the corresponding text are representative of what their authors are dealing with — this is a rather classical assumption in text mining.

We thus use a hand-made selection of some political weblogs (blogs whose posts almost exclusively deal with political life and issues [6]) among those we consider fairly active: we select blogs publishing political posts at a pace of at least 20 posts per month and having at least five comments per post

ICWSM ’2007 Boulder, CO – USA

on average, and make sure the breakdown of these sources is roughly representative of the context of the French political arena. We also harvest a selection of news media websites as an indicator of topics addressed by the press at the same time.[1] These press websites can also be considered as content-publishing weblogs. The number of sources is denoted by B: in our data $B = 39$, with 33 weblogs and 6 news websites.

We gather post contents on a daily basis for all these sources. We apply a simplistic linguistic treatment on the corresponding text in that we stem (lemmatize) words and group some synonymous terms in order to build term classes. Among the most frequent terms, we only consider a hand-made selection of $C = 75$ topics[2]. Additionally, we gather data on T days, which is the total observation time — in our case, $T = 30$; data collection starts on November 1st, 2006 and ends at the end of that month, on Nov 30th.

We illustrate our approach using this empirical data, which thus formally consists of a third-order tensor $\mathbf{W}$ where $\mathbf{W}_{i,j,k} = n$ if blog i features n occurrences of term j at time k.

3. Static case

3.1 Semantic profiles

The goal of this section is to first distinguish categories based on semantic similarity: we will next use these static semantic categories as groups among which to find inter-temporal correlations. We therefore compute a temporal aggregation of terms used by each weblog, so that we obtain a "semantic profile" matrix $\mathbf{w}$ from $\mathbf{W}$ such that $\mathbf{w}_{i,j}$ equals the number of occurrences of term j in blog i for the whole period. More precisely, $\mathbf{w}_{i,j} = \sum_{k=1}^{T} \mathbf{W}_{i,j,k}$, and $\mathbf{w}$ is thus a $B \times C$ matrix.

As such, this data is equivalent to a set of documents for which we know word occurrence frequencies. So far, such data has been extensively studied in information retrieval, following the famous vector-space model [9]. Here, we therefore consider each weblog as a vector in a vector-space where terms are vectorial directions — the C-dimensional vector $\mathbf{w}_i$ associated to blog i can be seen as its semantic profile.

3.2 Semantic clustering

To build categories of documents/blogs, we first compute the matrix of similarities between weblogs, i.e., between their semantic profiles/vectors. To this end, we provide:

(i) a normalization procedure, so that term occurrences are weighted properly. We adopt the "tf·idf" canonical approach, replacing $\mathbf{w}$ by a tf.idf-weighted matrix $\hat{\mathbf{w}}$. More precisely, this consists in weighting the term frequency "tf" (so that most used terms in a given blog are more important) with the inverse document frequency "idf", or frequency of the term in the whole corpus (so that rarer terms in the corpus weigh more: this takes into account the discriminating power of terms which, while usually rare in the corpus, are being abnormally mentionned in a given document).[3]

(ii) a similarity measure, so that weblog profiles can be compared. The canonical approach uses a "cosine" distance, i.e., similarity between blogs i and j is denoted by $\mathbf{s}(i,j) = \dfrac{\hat{\mathbf{w}}_i \cdot \hat{\mathbf{w}}_j}{\|\hat{\mathbf{w}}_i\| \|\hat{\mathbf{w}}_j\|}$.

Then, we simply compute the dendrogram associated to the similarity matrix $\mathbf{s}$. We can cut the dendrogram at a level providing a desired number of clusters $\mathcal{B}$. We choose $\mathcal{B} = 3$ categories and get the corresponding possible semantic community description. We hereafter denote them with α, β and γ respectively (although we shall not try providing a qualitative interpretation of our results according to these categories, we can *roughly* describe them as apparent left-wing, apparent right-wing and none-of-these). Considering our news media selection as yet another category — called "press" — we eventually have $\mathcal{B} = 4$ categories.

4. Dynamic case

4.1 Rationale

Given these semantic categories on weblogs and online press (considered as an "institutional content source"), we now wish to know if there exist some groups whose topics are synchronized with or influenced by some other groups — the influence is understood here as the fact that (i) the use by some groups of sources of a given term precedes at a later time a usage of this very term in some other groups, and that (ii) this sort of relationship can be observed on several terms. In other words, we look for dynamic patterns on correlated topic appearances, which are actually intertemporal patterns.

More to the point, such correlation may indicate a direct causal relationship (a topic becomes popular in a given media source, then becomes popular in another group discussing it — typically, media bringing up some issue which is commented on by "individual" webloggers — more broadly, it can reveal an admittedly more influencial group), an indirect one (a topic becomes popular in a given social group which overlaps some weblogger community), or just a different delay in addressing a similar issue (some common exogenous cause triggers a much earlier reaction from a webloger group than from another one).

4.2 Dynamic patterns

4.2.1 *Causal state machines*

To detect these patterns and possibly exhibit intertemporal causal relationships between them, we use a method for automatically reconstructing "causal state machines" as proposed by Crutchfield and Young in [2]. We first present the theoretical background, then detail how we apply it to our problem.

The bottom line of this approach is to:

1. consider discretized states for a given system, and identify equivalence classes of states — two states being considered equivalent if they statistically induce the same potential series of future states.

2. then, create a graph featuring transitions between state classes — a state class "causing", or inducing, another state class (that is, it builds a grammar of possible states) — and compute their probabilities.[4]

[1] More precisely: Europe 1, France Info, France Inter, L'Express, Le Figaro, Le Monde, Le Point, Libération.

[2] Detailed list of terms is provided at this address: `http://snafca.free.fr/icwsm/list.pdf`

[3] The coefficients $\mathbf{w}_{i,j}$ are thus actually replaced by $\hat{\mathbf{w}}_{i,j} := \dfrac{\mathbf{w}_{i,j}}{\sum_{j=1}^{C} \mathbf{w}_{i,j}} \cdot \log \dfrac{B}{d_j}$ where B is the total number of blogs, and d_j the number of blogs where term j appears [9].

[4] The construction respects the hidden Markov model properties[10], with the future being independent of the past.

In other words, it is possible to, at the same time, detect dynamic patterns of states in a system and appraise the (causal) relationships between these patterns, in terms of transition probabilities. It has been illustrated and implemented by Shalizi & Shalizi in [11], who provided in particular the CSSR algorithm which made this method solidly operational.[5] Better, it is also possible to detect inter-temporal behavioral correlations between different instances of an allegedly similarly behavioring system.

Let us translate this in the present weblog framework: if we assume that weblog groups behave similarly one with respect to each other (e.g. "group *press* raising an issue induces group α to talk of that same topic") on a significant number of topics, we can describe the symbolic dynamics of topics usage among sources as an Hidden-Markov model made of set of transition probabilities between the different causal states.

Similar behaviors among weblogs should obviously not occur at the same time for all topics — not all topics are "hot" at the same time — but the method makes it possible to detect similar behaviors intertemporally, with delays. By looking at several terms, which can actually be considered several instances of a similarly behaving system, it should become more likely to find these correlations.[6]

More precisely, the CSSR algorithm yields state classes of any symbolic dynamics along with their associated transition probabilities; practically, this method detects causal states made of discrete symbolic sequences of size L. Considering the sample sequence on Tab. 1, for states of length $L = 1$ CSSR may find the following possible causal states: $S1 = \{a\}$, $S2 = \{b\}$, $S3 = \{c\}$ and $S4 = \{d\}$. Following [11], we know that the causal state sequence $\{S_t\}$ is a Markov process, so we can represent the dynamics of the observed process —here, the dynamics of topic occurrences in weblogs— as a random function of the causal state process, as illustrated on Fig. 1.

emitted symbols:	a	a	b	c	c	c	b	d	d	d	a	...
B_1	0	0	1	1	1	1	1	0	0	0	0	...
B_2	0	0	0	0	1	1	1	1	1	1	0	...

time $\longrightarrow$

Table 1: *Sample temporal evolution of the presence of some given topic in two blogs B_1 and B_2 (or "emitted symbols" by both sources, using the classical HMM theory terminology)*

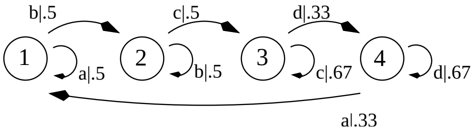

Fig. 1: *Causal state machine corresponding to the example of Tab. 1 for $L = 1$, transition probabilities are shown along with corresponding emitted symbols*

4.2.2 Alphabet of weblog topics

Switching back to our empirical study, we first focus on a given term (class) c. For each term, we thus consider all the possible cases of appearance of this term over the $\mathcal{B} = 4$ different categories: there are obviously $2^{\mathcal{B}}$ possible combinations. This makes up our *alphabet*, as shown on Tab. 2 — for clarity reasons we choose capital letters if the term appears in the press, and small letters otherwise.

Practically, to express term occurrences using our alphabet, we must aggregate the data over different categories of sources. Therefore we sum all frequencies in all blogs belonging to a category and normalize this sum by the number of blogs in the category. We get a matrix $\mathbf{M}^c$, associated to a term class, where $\mathbf{M}^c_{j,k}$ equals the averaged frequency of term c in posts published by group j dated at time k. The next step is to create a binary pattern for the expression (or non-expression) of a term over the various blog groups — in other words, we binarize the matrix $\mathbf{M}^c$ in order to translate term occurrences into the above-mentionned alphabet. Because some terms are more frequent than others, we again multiply the previous normalized value by the "idf" value (see Sec. 3.2) in order to discretize our dataset with a unique threshold for all concepts. Using this given threshold value we binarize $\mathbf{M}^c$: we put 1 if the occurrence frequency in a group is above this threshold, 0 otherwise.

We can eventually represent the time series of sequential alphabetic states of all terms by creating a matrix μ where rows are terms and columns are timesteps, each matrix cell being a symbol of the alphabet. If $\mu_{c,t}$ equals "f", for instance, this means that at time t, the term c appeared in blog groups α and γ.

4.2.3 Results

The algorithm provides causal states as shown on Tab. 2. Recall that these states are equivalence classes. Some are pretty easy to interpret: for instance, $S4$ corresponds to a unique letter in the alphabet e. This causal state is thus associated to the presence of a term in a unique group of blogs, γ. Similarly $S3$, which is associated to "h", corresponds to a term present in every weblog group, but absent from the press. $S0$ is associated to both "a" and "G", i.e., terms being absent of every weblog category ("a") or present everywhere except in category α ("G"). $S1$ is a more complex state, as it features symbols corresponding to topic occurrence in at most 2 groups, not often the press. $S2$ will be discussed below.

The CSSR algorithm also provides us with the probabilities of transition between these causal states. In order to focus on significant transitions (or influences) as well as to facilitate result analysis, we deliberately chose to remove any transition whose probability is under 10%. The transition graph of the resulting causal state machine is plotted on Fig. 2.

Interpretation. We observe on this graph that $S3$, which is associated to the presence of the term everywhere except in the press, has probability 0.30 to stay in the same state at the next step. Another transition possibility is to switch to state $S1$, with the same probability, or to $S2$, with probability 0.15. More precisely, if, the next day, weblogs categories α or β stop using this term with probability 0.15 each, then the state switches to $S1$ which is typical of minor topics. There is also probability 0.15 for the press to talk about it the next day, if so we switch to $S2$.

$S4$ is a causal state in which only group γ discusses the topic. We can see that with a probability 0.27, the state is

[5] Available at http://bactra.org/CSSR

[6] Eventually, although not undertaken here, it should be possible to explicitly assess the correlations between these different instances — pretty much like in [7] for studying delays in neural assembly synchronization.

alphabet	a	b	c	d	e	f	g	h	A	B	C	D	E	F	G	H
α	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
β	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
γ	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
press	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
causal state	S0	S1	S1	S1	S4	S1	S1	S3	S1	S2	S1	S2	S1	S2	S0	S2

Table 2: *Chosen alphabet and associated causal states; for instance, symbol "f" corresponds to the occurrence of a term in weblog groups α and γ only*

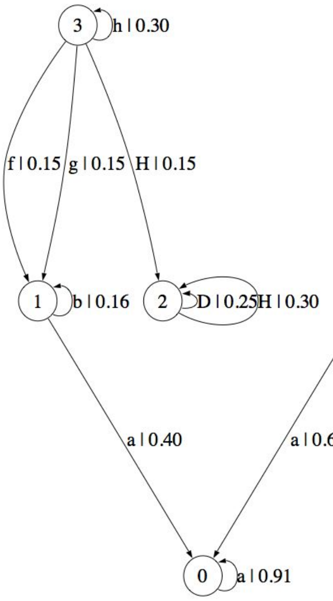

Fig. 2: *Empirical causal state machine (featuring states $S0$, $S1$, $S2$, $S3$, and $S4$), most significant transitions (i.e., whose probability is over .1) and corresponding emitted symbols*

exactly stationary (which means that the system keeps the same topic occurrence pattern). It also has strong chance (0.62) to switch to state $S0$ by emitting "a" — nobody mentions the concept — which means that everybody, even γ, stops talking about it. Once in this causal state, the probability of quiting it is very little as the probability to remain in state $S0$ is very high (0.91); again, emitting "a".

The last causal state $S2$ encloses all cases for which at least α and the press mention the term. It is interesting to note that the two main ways of remaining in this state consist in emitting symbols D or H with high probabilities (respectively 0.25 and 0.30). With H, every group mentions the term, it is plausibly a case of "buzz" around a topic. We may also note that the only significant paths leading to this state are those mentioned above: either by switching from state $S3$ to H with probability 0.15, or through a kind of autocatalytic popularity around a term with transition "$S2 \rightarrow S2$" ending by state H with probability 0.30.

5. Conclusion

We extracted occurrence statistics on a set of 75 terms from a corpus made of 33 French political weblogs and 6 online press sources. Static categorization of political weblog data enabled us to build semantically coherent groups of weblogs. Using this aggregated description we managed to describe the system as an hidden Markov model. Most interestingly, we were able to provide its causal states and transition probabilities and exhibit not only behavioral correlations between some groups of blogs and online media, but also identify varied and richer types of inter-group patterns. In particular, using a very compact description, this methodology enabled us to infer interpretations as to how diverse groups of blogs behave with respect to each other as regards raising and discussing issues.

Acknowledgments. We acknowledge remarks from Paul Bourgine and partial funding from CNRS and EU project "PATRES".

References

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on Weblogging Ecosystem, 13th WWW*, 2004.

[2] J. P. Crutchfield and J. Young. Inferring statistical complexity. *PRL*, 63(2):105–108, 1989.

[3] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *11th ACM SIGKDD*, pages 78–87. ACM Press, 2005.

[4] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of WWW2004*, NYC, NY, USA, May 17-22 2004.

[5] S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu. Conversations in the blogosphere: An analysis "from the bottom up". In *Proceedings of HICSS-38*, 2005.

[6] M. Hindman, K. Tsioutsiouliklis, and J. A. Johnson. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political Science Association*, 2003.

[7] K. Klinkner, C. Shalizi, and M. Camperi. Measuring shared information and coordinated activity in neuronal networks. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 667–674. MIT Press, Cambridge, MA, 2006.

[8] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs(AAAI-CAAW), Palo Alto, CA*, 2006.

[9] G. Salton, A. Wong, and C. S. Yang. Vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[10] C. R. Shalizi. *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata.* PhD thesis, University of Wisconsin at Madison, U.S.A., 2001. Chap. 11.

[11] C. R. Shalizi and K. L. Shalizi. Blind construction of optimal non-linear recursive predictors for discrete sequences. In M. Chickering and J. Halpern, editors, *Uncertainty in Artificial Intelligence: Proceedings of the 20th Conference*, pages 504–511, 2004.