A Time-based Collective Factorization for Topic Discovery and Monitoring in News

Carmen Vaca[†‡*] Amin Mantrach[+] Alejandro Jaimes[+] Marco Saerens[**]

cvaca@fiec.espol.edu.ec, {amantrac,ajaimes}@yahoo-inc.com, marco.saerens@uclouvain.be

[‡]Politecnico di Milano, Milan, Italy
[+]Yahoo Labs, Barcelona, Spain
[**]Université de Louvain, Louvain, Belgium
[†]Escuela Superior Politecnica del Litoral, Guayaquil,Ecuador

ABSTRACT

Discovering and tracking topic shifts in news constitutes a new challenge for applications nowadays. Topics *evolve*, *emerge* and *fade*, making it more difficult for the journalist – or the press consumer – to decrypt the news. For instance, the current *Syrian chemical crisis* has been the starting point of the *UN Russian initiative* and also the revival of the *US France alliance*. A topical mapping representing how the topics evolve in time would be helpful to contextualize information. As far as we know, few topic tracking systems can provide such temporal topic connections. In this paper, we introduce a novel framework inspired from *Collective Factorization* for online topic discovery able to connect topics between different time-slots. The framework learns jointly the topics evolution and their time dependencies. It offers the user the ability to control, through one unique hyper-parameter, the tradeoff between the past accumulated knowledge and the current observed data. We show, on semi-synthetic datasets and on Yahoo News articles, that our method is competitive with state-of-the-art techniques while providing a simple way to monitor topics evolution (including emerging and disappearing topics).

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Text processing*; G.1.3 [**Numerical Linear Algebra**]: Sparse, structured, and very large systems (direct and iterative methods); G.1.6 [**Optimization**]: Gradient methods

General Terms

Algorithms, Theory

*This work was carried out while the author was an intern at Yahoo Labs, Barcelona.

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04
http://dx.doi.org/10.1145/2566486.2568041.

Keywords

Topic discovery; topic monitoring; topic tracking; streaming; collective factorization; online learning

1. INTRODUCTION

Discovering and tracking of topics has become increasingly important for text stream analysis. Automatic extraction of meaningful concepts from large amounts of documents can help detecting events taking place in real time and facilitate the exploration of unlabeled user-generated content archives.

Over the last decade, many strategies for topic detection based on probabilistic models have appeared. In particular, LDA (Latent Dirichlet Allocation), the seminal work by Blei *et al.*[3], has been extended to improve performance or to incorporate contextual information regarding authors and their social network connections [16, 20, 24]. Extensions of LDA have also been proposed for dynamic topic detection [2, 1]. However, these models have a main drawback in that high computational times make them unable to deal with large amounts of documents arriving in real time. Hence, these approaches cannot be applied in many real-world scenarios where data must be processed online and efficiently. Prominent examples are online news outlets and social media where users are continuously producing large amount of data whose topics rapidly grow and fade in intensity across time.

In 2007, the Associated Press, through interviews to young adults news consumers, found that the high amount of information available nowadays do not help the audience to obtain more insight from news [22]. People who participated in the study explained that they would rather prefer less content to be able to discern critical information. Journalists face the same challenge when trying to a) get an overview of the news of the day and b) deriving connections among today stories with previous ones. In this scenario topic discovery frameworks can help to identify a set of keywords related to news documents that let the journalist obtain a quick overview of the stories. However, it is crucial to work on the ability of these frameworks to find connections between stories across time that contextualize better the information stream. As far as we know there are few state-of-the-art topic streaming systems closing this gap.

To address this concern, we model the textual stream using a collective matrix factorization based framework [28] that jointly learns topics evolution and their dependencies. The framework relies on one *unique* parameter controlling

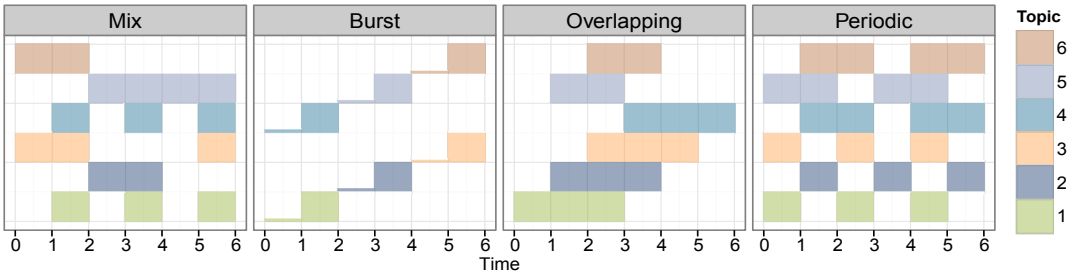

Figure 1: Synthetic dataset representing four scenarios.

the trade-off between the memory of the system and the current observed data. This parameter acts also as a regularization parameter of the introduced loss function. We are able to detect temporal connections between topics discovered in consecutive time-slots. Discovering these connections "online" gives insights to the user for contextualizing the evolving trends: *emerging*, *evolving* and *fading* topics.

Moreover, one of the most important aspects when tracking or discovering topics relies on the underlying temporal patterns guiding the topic evolution. Such temporal patterns correspond not only to the evolution of real world events [8] but also to the way people publish and read information [15, 27]. For instance, some topics may appear as *bursts* as in the case of natural disasters, others vanish and reappear *periodically* as, for example, tv shows broadcasted weekly. During the monitoring, the topics do not necessarily appear in the same time-slot presenting an *overlapping* pattern. Finally, it is also possible to have a combination of these settings where different topical patterns are *mixed* together. We show how our model and state-of-the-art approaches are able to track topics on synthetic datasets (Figure 1) exhibiting those different patterns. We also compare our method with state-of-the-art approaches for topic discovery and tracking on Yahoo News articles collected over a period of two weeks.

Our main contributions are:

- A novel time-based *collective factorization* approach for topic detection and monitoring that models the underlying temporal dynamics in news streams using a unique hyper-parameter that controls the tradeoff between the past accumulated knowledge (i.e. *the memory*) and the current observed data;
- A proof of convergence for a multiplicative-updates based algorithm that may handle a large amount of data in an online environment;
- A benchmark study that compares topic tracking performance of five state-of-the-art techniques on synthetic and real world datasets both in terms of topics discovered and execution time.

The rest of the paper is organized as follows. In the next Section we discuss the related work. In Section 3 we formalize the problem. In Section 4, we introduce a novel time-based collective factorization to control the tradeoff between the memory and the new observed data. In Section 5, we derive a multiplicative-updates based algorithm and discuss its properties. We present the experimental setup for topic discovery and tracking and show the temporal dynamics of topics discovered in one of our datasets in Section 6. Finally, in Section 7 we present our conclusions and future work.

2. RELATED WORK

Online topic detection and tracking methods have been the subject of interesting work in recent years. Recently, topic detection frameworks have been applied for news events detection using microblogging posts [23]. The majority of the topic detection frameworks are extensions of well known algorithms, recognized as state-of-the-art for topic detection in a "non-online" setting. In probabilistic approaches, the LDA (Latent Dirichlet Allocation) [3] has become a reference work in Bayesian learning. The counterpart of LDA in the matrix factorization community, has been NMF (Non negative Matrix Factorization) introduced some years before by [14]. Naturally, both approaches have been considered as starting frameworks to handle the streaming setting.

Blei *et. al* extended LDA for Dynamic Topic Models (DTM) [2]. From this seminal work, other graphical models have been proposed to incorporate the temporal dimension: Trends Over Time[29], Trend Analysis Model (TAM) [12], TM-LDA[30] and temporal Discriminative Probabilistic Model (DPM) [9].

NMF has also been extended to the Online-NMF in [4] to deal with continuously incoming data. The authors added a regularization constraints to the NMF optimization problem in order to take into account of the previously learned topics. This approach is used as one of the baselines in this paper.

Recently, the dictionary learning paradigm has been used as the basis of an online learning strategy for NMF [17]. This approach consists of minimizing the (desired) expected cost when the training set size goes to infinity. A publicly available sparse modeling software implementing this strategy is used as baseline in this paper. The classical online strategy used in this software relies on the assumption that the unknown relationship between the observations and the hidden factors is stationary (i.e. is not time dependent). In other words, document in different time-slots should be exchangeable. However, when facing evolving topics this assumption may be considered too strong [2]. Therefore, this approach has been extended to handle evolving input streams in [11]. Additionally, in [25] the authors proposed a dynamic NMF framework with a complex temporal regularization. This algorithm is benchmarked on different datasets in this paper. However, none of the approaches cited so far model explicitly how topic connects across consecutive time-slots.

As far as we know, the only topic tracking system going in this direction has been [30], where the authors propose to directly learn the topic transition matrix in social me-

dia using a least-squares based technique. The proposed approach consists of independently learning the topics at each time step using a standard Bayesian model (namely the LDA) and, subsequently, from the obtained topic distributions learn a topic transition matrix by solving a square-loss optimization problem. This approach suffers from a main drawback: instead of directly learning from the input data streams it assumes the existence of a fast and reliable LDA model. Therefore, in this paper, we propose to learn (1) the documents-by-topics and topics-by-words decompositions at each time step, as well as (2) their associated transition matrix directly from the current input stream. In this way, we can track topics and their transitions using a unified framework.

In this paper, we propose a novel time-based *collective matrix factorization* framework for discovering topic and their connections along time. The idea of collective matrix factorization has been first formalized by Singh and Gordon [28] as general framework for multi-relational factorization models. They subsume models on any number of relations as long as their loss function is a twice differentiable decomposable loss. In their work, they address the problem of items recommendation. The collective factorization approach proposed in this work is based on a similar idea of joint (collective) factorization. However, instead of factorizing simultaneously several matrices, we factor one unique matrix by several time-based factors (a mathematical formalization is given further in Section 4).

3. PROBLEM STATEMENT

The problem we are tackling may be formalized as follows: a collection of documents arrives continuously in batches. Each batch is represented by a data matrix $\mathbf{X}^{(t)}$ of size $N_d^{(t)} \times N_f$, where $N_d^{(t)}$ is the number of documents produced at time step t and N_f is the number of features in a coding scheme. For instance, the frequency of 1-grams in a bag-of-words representation. We assume that the dictionary (i.e. N_f) is known in advance. This is indeed a realistic assumption, when processing English news, for example, the dictionary can be extracted from a set of independent articles judged as representative of the domain.

The complete data matrix $\mathbf{X}$, obtained by concatenating vertically the matrices $\mathbf{X}^{(t)}$ along the time steps, is considered huge and practically difficult to store and to handle. The simplest approach to topic detection consists of directly learning from the global matrix $\mathbf{X}$. However, in the real world, we are observing evolving topics and trends [19]. Hence, using out-of-date data to estimate current trends may lead to wrong inference. Another typical strategy, consists of directly learning topics from the current batch of data while ignoring the trends history.

One is therefore faced with the tradeoff between past and present observations. Completely forgetting the past might result in loss of crucial contextual information. However, using models learned on outdated data will lead to failures in the detection of sudden events like natural disasters. In this context, we expect from an online model to detect, at specific time steps, dominant trends that a global trained model may have missed.

In this work, we propose a new collective NMF-based framework for evolving input data streams. Non-negative Matrix Factorization (NMF) aims at decomposing a matrix $\mathbf{X}$ in two non-negative, lower dimensional matrices $\mathbf{W}$ and $\mathbf{H}$, such that their product can well approximate the original matrix $\mathbf{X}$, *i.e.*, $\mathbf{X} \approx \mathbf{WH}$.

Unlike other matrix factorization techniques, such as SVD, it imposes non-negativity constraints on the resulting matrices. These constraints result in an additive effect that leads to a so-called "additive parts-based" representation of the data [6]. The discovered factors are sparse and easily interpretable, *i.e.*, the basis vectors naturally correspond to conceptual properties of the data. Moreover, the sparsity of the factors results in easier application to new data.

In our context, the NMF consists of treating the data matrix observed at time t $\mathbf{X}^{(t)}$ as a product of two factors $\mathbf{W}^{(t)}\mathbf{H}^{(t)}$, with $\mathbf{W}^{(t)} \geq 0$ and $\mathbf{H}^{(t)} \geq 0$ (i.e. non-negative entries), where $\mathbf{W}^{(t)}$ has a size of $N_d^{(t)} \times K$ and $\mathbf{H}^{(t)}$ has a size of $K \times N_f$, where K represents the number of topics. Usually, K is much smaller than N_f.

The input data matrix as well as its decomposition are likely to be sparse at each time step. Therefore, we may consider the approach as a way to compress the information accumulated over time. In the context of "topic discovery", the goal is to predict $\mathbf{H}^{(t)}$, the words distribution for each topic at time t, which should be as close as possible to the real topics known at time t (we will refer to it as $\mathbf{H}_{true}$). Another related application is "topic tracking", where typically for a given set of known topics, $\mathbf{H}_{known}$, we would like to track their level of activity in time. In this case, the task consists of predicting at each time step $\mathbf{W}^{(t)}$ that contains a per-document topic distribution at time t. As we are also interested to connect topic shifts along time, we also would like to automatically learn how $\mathbf{H}^{(t)}$ relates to $\mathbf{H}^{(t-1)}$.

4. JOINT PAST-PRESENT DECOMPOSITION MODEL

Starting from the observation that there is valuable information to extract from past and present, we model the trade-off between both realities. The first reality, admits a *present decomposition* at time t:

$$\mathbf{X}^{(t)} \approx \mathbf{W}^{(t)}\mathbf{H}^{(t)} \tag{1}$$

However, we would like to say something about the current data $\mathbf{X}^{(t)}$ in terms of the accumulated history (i.e. the *past*, the *memory*). Important information about the past is revealed by $\mathbf{H}^{(t-1)}$, the previous discovered topics. Although the observed data is dynamic, we may comfortably assume that the topics evolved smoothly during one time step, and that the current topics are related to those that appeared in the previous time-slots. Therefore, we suppose that the new data may also be decomposed in terms of the previous topics, leading to a *past* decomposition of the same data matrix expressed by the following equation:

$$\mathbf{X}^{(t)} \approx \mathbf{W}^{(t)}\mathbf{M}^{(t)}\mathbf{H}^{(t-1)} \tag{2}$$

with $\mathbf{H}^{(t-1)}$ given. The proposed model directly explains the current data ($\mathbf{X}^{(t)}$) jointly by the present and the past through a mapping factor ($\mathbf{M}^{(t)}$). The matrix $\mathbf{M}^{(t)}$ is a topic-transition matrix trying to capture how much the current topic distribution ($\mathbf{H}^{(t)}$) may be linearly explained from the previous one ($\mathbf{H}^{(t-1)}$). In this way, we jointly learn topics evolution given by $\mathbf{H}^{(t)}$ and their temporal dependencies given by $\mathbf{M}^{(t)}$. Notice that we impose $\mathbf{M}^{(t)} \geq 0$ to stay

in a strictly non-negative decomposition framework. The joint constraint proposed in this model is soft as it operates indirectly through $\mathbf{W}^{(t)}$, common to both decompositions. $\mathbf{W}^{(t)}$ is the membership matrix quantifying to which extent a document belongs to a specific topic. In this model, this membership depends on both, the current topic distribution: $\mathbf{H}^{(t)}$, and the past one: $\mathbf{H}^{(t-1)}$. Hence, the model is called the *Joint Past Present* (JPP) decomposition.

To summarize, we propose to decompose collectively:

$$\begin{cases} \mathbf{X}^{(t)} \approx \mathbf{W}^{(t)}\mathbf{H}^{(t)} & (3) \\ \mathbf{X}^{(t)} \approx \mathbf{W}^{(t)}\mathbf{M}^{(t)}\mathbf{H}^{(t-1)} & (4) \end{cases}$$

The idea of imposing a common $\mathbf{W}^{(t)}$ for both decomposition comes from traditional collective factorization techniques [28]. However, instead of simultaneously factorizing several matrices, we factorize one unique matrices by severals different time-based factors $\mathbf{H}^{(t)}$ and $\mathbf{H}^{(t-1)}$. In this way, we can connect time-based factors in one unique factorization. As far as we know, such a time-based decomposition has never been proposed.

In order to solve the problem, we need to define a specific loss function $\mathcal{L}(\mathbf{X}^{(t)}; \mathbf{W}^{(t)}; \mathbf{H}^{(t)}; \mathbf{M}^{(t)}; \mathbf{H}^{(t-1)})$ which quantifies the distance between the original matrix, $\mathbf{X}^{(t)}$, and the obtained decompositions in Equations (3) and (4). This loss function may, for instance, be the Frobenius-norm or the KL divergence. In the mean time, we would like to add regularization constraints on the possible solutions in order to either enforce sparsity or to reduce the complexity of the model. Setting all these factors together, the resulting optimization problem aims to minimize the following loss function:

$$\begin{aligned} L = \underset{\mathbf{W}^{(t)},\mathbf{H}^{(t)},\mathbf{M}^{(t)}}{\arg\min} & \|\mathbf{X}^{(t)} - \mathbf{W}^{(t)}\mathbf{H}^{(t)}\|_F^2 \\ & + \|\mathbf{X}^{(t)} - \mathbf{W}^{(t)}\mathbf{M}^{(t)}\mathbf{H}^{(t-1)}\|_F^2 \\ & + \lambda\|\mathbf{M}^{(t)} - \mathbf{I}\|_F^2 + \alpha\|\mathbf{H}^{(t)}\|_1 + \beta\|\mathbf{W}^{(t)}\|_1 + \gamma\|\mathbf{M}^{(t)}\|_1 \end{aligned} \quad (5)$$

subject to $\mathbf{W}^{(t)} \geq 0, \mathbf{H}^{(t)} \geq 0$ and $\mathbf{M}^{(t)} \geq 0$ where $\|.\|_F$ represents the Frobenius norm and $\|.\|_1$ stands for the $l_1$norm. The $l_1$norm based regularization has the known effect to promote sparsity which is desired when modeling topics [3, 17]. In theory, using the $l_1$norm also on $\mathbf{M}^{(t)}$ as the effect of promoting a smooth evolution (one topic should evolve from a little number of existing topics). The temporal regularization $\lambda\|\mathbf{M}^{(t)} - \mathbf{I}\|_F^2$ controls how much the user wants to bias the decomposition towards $\mathbf{H}^{(t-1)}$. The λ parameter $\in (0, \infty)$ balances present and past information; it quantifies the extent to which the model is past (i.e. $\lambda \to \infty$) or present oriented (i.e. $\lambda \to 0$).

5. DERIVED ALGORITHM

The problem (5) is not convex for all parameters $\mathbf{W}^{(t)}$, $\mathbf{H}^{(t)}$, $\mathbf{M}^{(t)}$ simultaneously. However, we can find a local minimum for the objective function using a multiplicative-updates as introduced by [14].

Considering the Karush-Kuhn-Tucker (KKT) first-order conditions applied to our problem, we derive:

$$\mathbf{W}^{(t)} \geq \mathbf{0}, \ \ \mathbf{H}^{(t)} \geq \mathbf{0}, \ \ \mathbf{M}^{(t)} \geq \mathbf{0}, \quad (6)$$

$$\nabla_{\mathbf{W}^{(t)}} L \geq \mathbf{0}, \ \ \nabla_{\mathbf{H}^{(t)}} L \geq \mathbf{0}, \nabla_{\mathbf{M}^{(t)}} L \geq \mathbf{0}, \quad (7)$$

$$\mathbf{W}^{(t)} \odot \nabla_{\mathbf{W}^{(t)}} L = \mathbf{0}, \ \ \mathbf{H}^{(t)} \odot \nabla_{\mathbf{H}^{(t)}} L = \mathbf{0}, \ \ \mathbf{M}^{(t)} \odot \nabla_{\mathbf{M}^{(t)}} L = \mathbf{0} \quad (8)$$

where $\odot$ is the element-wise product.

From the loss function in Equation (5), we derive the gradients according to each parameter:

$$\nabla_{\mathbf{H}^{(t)}} L = \mathbf{W}^{(t)^{\mathrm{T}}}\mathbf{W}^{(t)}\mathbf{H}^{(t)} - (\mathbf{W}^{(t)^{\mathrm{T}}}\mathbf{X}^{(t)} - \alpha) \quad (9)$$

$$\nabla_{\mathbf{W}^{(t)}} L = \mathbf{W}^{(t)}(\mathbf{H}^{(t)}\mathbf{H}^{(t)^{\mathrm{T}}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}\mathbf{H}^{(t-1)^{\mathrm{T}}}\mathbf{M}^{(t)^{\mathrm{T}}}) - (\mathbf{X}^{(t)}\mathbf{H}^{(t)^{\mathrm{T}}} + \mathbf{X}^{(t)}\mathbf{H}^{(t-1)^{\mathrm{T}}}\mathbf{M}^{(t)^{\mathrm{T}}} - \beta) \quad (10)$$

$$\nabla_{\mathbf{M}^{(t)}} L = (\mathbf{H}^{(t-1)}\mathbf{H}^{(t-1)^{\mathrm{T}}})\mathbf{M}^{(t)^{\mathrm{T}}}(\mathbf{W}^{(t)^{\mathrm{T}}}\mathbf{W}^{(t)}) + \lambda\mathbf{M}^{(t)^{\mathrm{T}}} - (\mathbf{H}^{(t-1)}\mathbf{X}^{(t)^{\mathrm{T}}}\mathbf{W}^{(t)} + \lambda\mathbf{I} - \gamma) \quad (11)$$

By substituting the corresponding gradients in Equation(8), we derive the following update Equations:

$$\mathbf{H}^{(t)} = \mathbf{H}^{(t)} \odot \frac{[(\mathbf{W}^{(t)^{\mathrm{T}}}\mathbf{X}^{(t)} - \alpha)]}{[(\mathbf{W}^{(t)^{\mathrm{T}}}\mathbf{W}^{(t)}\mathbf{H}^{(t)})]} \quad (12)$$

$$\mathbf{W}^{(t)} = \mathbf{W}^{(t)} \odot \frac{[(\mathbf{X}^{(t)}\mathbf{H}^{(t)^{\mathrm{T}}} + \mathbf{X}^{(t)}\mathbf{H}^{(t-1)^{\mathrm{T}}}\mathbf{M}^{(t)^{\mathrm{T}}} - \beta)]}{[(\mathbf{W}^{(t)}(\mathbf{H}^{(t)}\mathbf{H}^{(t)^{\mathrm{T}}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}\mathbf{H}^{(t-1)^{\mathrm{T}}}\mathbf{M}^{(t)^{\mathrm{T}}}))]} \quad (13)$$

$$\mathbf{M}^{(t)} = \mathbf{M}^{(t)} \odot \frac{([\mathbf{H}^{(t-1)}\mathbf{X}^{(t)^{\mathrm{T}}}\mathbf{W}^{(t)} + \lambda\mathbf{I} - \gamma)]}{[(\mathbf{H}^{(t-1)}\mathbf{H}^{(t-1)^{\mathrm{T}}})\mathbf{M}^{(t)^{\mathrm{T}}}(\mathbf{W}^{(t)^{\mathrm{T}}}\mathbf{W}^{(t)}) + \lambda\mathbf{M}^{(t)^{\mathrm{T}}}]} \quad (14)$$

These last Equations lead to Algorithm 1.

Theorem 1 *The loss function L in Equation (5) is non increasing under the update rules in Equations (12), (13) and (14). The loss function L is invariant under these updates if and only if* $\mathbf{H}^{(t)}$, $\mathbf{W}^{(t)}$ *and* $\mathbf{M}^{(t)}$ *are at a stationary point of the function.*

The proof of this theorem if given in the Appendix.

6. EXPERIMENTS AND DISCUSSIONS

6.1 Datasets

We evaluate our method using two different text corpora whose documents are labeled with time stamps and categories:

1. **Yahoo News**. We crawled 13,319, publicly available, news articles published on the Yahoo RSS feeds[1] between September 19th and October 2nd, 2012. The documents are annotated by experts with one or more categories out of 76 available labels. The collection was preprocessed to remove punctuation, stop-words and numbers. After lemmatization, term frequency vectors were produced out of each document.

[1]http://developer.yahoo.com/rss/

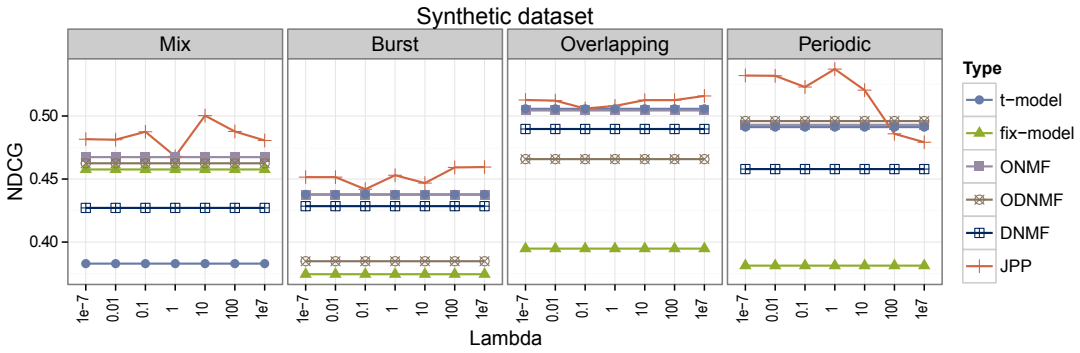

Figure 2: Impact of the λ parameter on the JPP model observed in four synthetic datasets with different evolution patterns. NDCG values are averaged over 15 folds.

input : $\mathbf{X}^{(t)}, \mathbf{H}^{(t-1)}, \lambda, \epsilon$
output: $\mathbf{W}^{(t)}, \mathbf{H}^{(t)}, \mathbf{M}^{(t)}$

$\mathbf{W}^{(t)}, \mathbf{H}^{(t)}, \mathbf{M}^{(t)} \leftarrow$ *random non-negative init*;
$\delta^{'} \leftarrow maxInt, \delta \leftarrow \frac{\delta^{'}}{2}$;
$\beta \leftarrow$ *to choose in [0.001,0.05] [5]*;
$\lambda \leftarrow$ *to choose in* $[0, \infty[$;
while $abs(\delta^{'} - \delta) \geq \epsilon$ **do**
  $\mathbf{H}^{(t)} \leftarrow \mathbf{H}^{(t)} \odot \frac{[(\mathbf{W}^{(t)\mathrm{T}}\mathbf{X}^{(t)} - \alpha)]}{[(\mathbf{W}^{(t)\mathrm{T}}\mathbf{W}^{(t)}\mathbf{H}^{(t)})]}$;
  $\mathbf{W}^{(t)} \leftarrow \mathbf{W}^{(t)} \odot \frac{[(\mathbf{X}^{(t)}\mathbf{H}^{(t)\mathrm{T}} + \mathbf{X}^{(t)}\mathbf{H}^{(t-1)\mathrm{T}}\mathbf{M}^{(t)\mathrm{T}} - \beta)]}{[(\mathbf{W}^{(t)}(\mathbf{H}^{(t)}\mathbf{H}^{(t)\mathrm{T}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}\mathbf{H}^{(t-1)\mathrm{T}}\mathbf{M}^{(t)\mathrm{T}}))]}$;
  $\mathbf{M}^{(t)} \leftarrow \mathbf{M}^{(t)} \odot \frac{([\mathbf{H}^{(t-1)}\mathbf{X}^{(t)\mathrm{T}}\mathbf{W}^{(t)} + \lambda\mathbf{I} - \gamma)]}{[(\mathbf{H}^{(t-1)}\mathbf{H}^{(t-1)\mathrm{T}})\mathbf{M}^{(t)\mathrm{T}}(\mathbf{W}^{(t)\mathrm{T}}\mathbf{W}^{(t)}) + \lambda\mathbf{M}^{(t)\mathrm{T}}]}$;
  $\delta^{'} \leftarrow \delta$;
  $\delta \leftarrow \mathcal{L}(\mathbf{X}^{(t)}; \mathbf{W}^{(t)}; \mathbf{H}^{(t)}; \mathbf{M}^{(t)}; \mathbf{H}^{(t-1)})$;
end

Algorithm 1: Joint Past Present decomposition Algorithm.

2. **Semi-synthetic datasets**. Four semi-synthetic datasets were generated by extracting the top six topics in the first six timeslots of the TDT2 collection. TDT2 is the NIST Topic Detection and Tracking [2] text corpora (collections of broadcast news recordings and transcripts). It contains stories extracted from six different news sources published during the first semester of 1998. To match the topic dynamics presented in Figure 1: *mixed, burst, overlapping* and *periodic*, we remove documents corresponding to inactive topics in each timeslot. For instance, consider the topic 1 in the *overlapping* dataset, we make it inactive at times 3, 4 and 5 by removing all the documents relevant to this topic in these timeslots.

In the next sections, we evaluate topic discovery (Section 6.2) and topic tracking (Section 6.3).

6.2 Topic Discovery

The topic discovery (or detection) task consists of detecting novel, previously unknown topics [7]. In this context, the goal is to predict $\mathbf{H}^{(t)}$, the word distribution for each topic at time t, as close as possible to the real word distribution, the ground-truth $\mathbf{H}_{\text{true}}$. However, this ground-truth is usually unknown. Hence, inspired by [25], we simply estimate each row in $\mathbf{H}_{\text{true}}$ as the center of mass of all documents belonging to the topic up to time t, exploiting the document annotations available in the datasets. Such annotations classify documents as related to specific events or categories. Considering that it is common to define a topic using the list of the top five or ten words [21, 26], we consider as the ground-truth the top 10 words appearing in each of the calculated centroids.

6.2.1 Baselines for Comparison

The performance of the JPP algorithm to predict $\mathbf{H}^{(t)}$ is compared with five different baseline methods:

- The t-model, consists of a basic NMF launched on the same input data stream $\mathbf{X}^{(t)}$. The NMF algorithm is implemented with multiplicative-updates rules and $l1$-norm regularization [5];
- The *fix*-model is calculated by learning a topic distribution on data seen on previous timeslots. It uses the same NMF implementation than the t-model;
- An extension of the Online-NMF [4] (ONMF). It is built adding a constraint to NMF, minimizing the *Frobenius*-norm between the previous fixed topic distribution $\mathbf{H}^{(t-1)}$ and the one we want to discover $\mathbf{H}^{(t)}$. We implemented it using multiplicative-updates rules with $l1$-norm regularization;
- The online dictionary learning NMF (ODNMF) of [17] included in the sparse modeling software optimization toolbox. We use this model in batch mode with a warm-up initialization strategy. In other words, at time step t we start the learning process from the solution learned up to time step $t-1$.
- The Dynamic NMF (DNMF) of [25] using the previous fixed topic distribution $\mathbf{H}^{(t-1)}$ to discover $\mathbf{H}^{(t)}$. It uses also the document arrival time stamps. In case of Yahoo News, we use the hour of publication as given by the feed. For the semi-synthetic datasets, we use the day of publication during the week.

[2] http://www.nist.gov/speech/tests/tdt/tdt98/index.htm

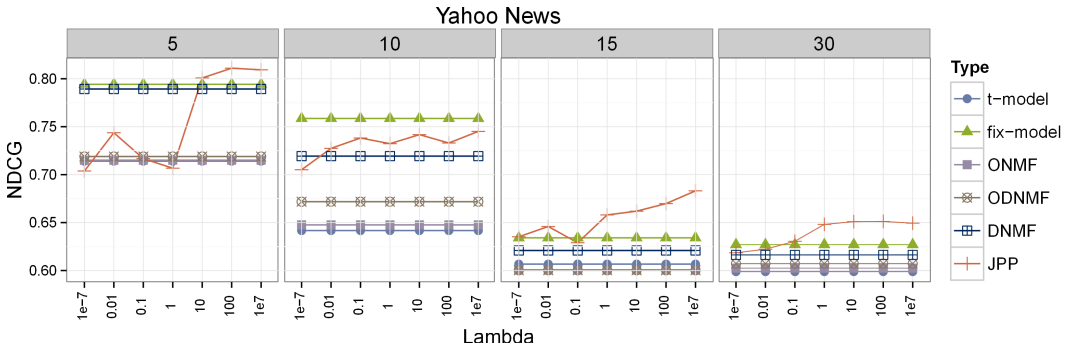

Figure 3: Performance comparison among algorithms for experiments on the Yahoo News dataset. Results are shown for k ranging from 5 to 30 and NDCG scores are averaged over 91 folds.

Metric	Model	5	10	15	30
microF1	t-model	0.43 ± 0.08	0.36 ± 0.09	0.33 ± 0.06	0.32 ± 0.03
	fix-model	0.46 ± 0.02	0.48 ± 0.02	0.32 ± 0.01	0.33 ± 0.01
	ONMF	0.44 ± 0.08	0.37 ± 0.09	0.32 ± 0.05	0.32 ± 0.04
	ODNMF	0.53 ± 0.11	0.41 ± 0.01	0.34 ± 0.05	0.35 ± 0.04
	DNMF	0.45 ± 0.10	0.39 ± 0.08	0.33 ± 0.06	0.32 ± 0.03
	JPP	**0.55** ± 0.06	0.47 ± 0.07	**0.41** ± 0.05	0.36 ± 0.03
MAP	t-model	0.45 ± 0.08	0.37 ± 0.11	0.34 ± 0.06	0.32 ± 0.04
	fix-model	0.56 ± 0.01	**0.52** ± 0.03	0.35 ± 0.01	0.33 ± 0.01
	ONMF	0.45 ± 0.07	0.38 ± 0.10	0.33 ± 0.05	0.33 ± 0.05
	ODNMF	0.46 ± 0.11	0.45 ± 0.12	0.37 ± 0.09	0.34 ± 0.05
	DNMF	0.47 ± 0.11	0.40 ± 0.09	0.33 ± 0.07	0.32 ± 0.03
	JPP	**0.59** ± 0.06	0.49 ± 0.09	**0.43** ± 0.06	**0.36** ± 0.04
NDCG	t-model	0.71 ± 0.07	0.64 ± 0.10	0.61 ± 0.05	0.60 ± 0.04
	fix-model	0.79 ± 0.01	0.76 ± 0.02	0.63 ± 0.01	0.63 ± 0.01
	ONMF	0.72 ± 0.05	0.65 ± 0.09	0.60 ± 0.04	0.60 ± 0.05
	ODNMF	0.70 ± 0.09	0.71 ± 0.09	0.64 ± 0.07	0.62 ± 0.05
	DNMF	0.72 ± 0.09	0.67 ± 0.07	0.60 ± 0.06	0.61 ± 0.04
	JPP	**0.81** ± 0.04	0.75 ± 0.07	**0.68** ± 0.05	**0.65** ± 0.04

Table 1: Topic detection evaluation, using three metrics, on the Yahoo News dataset. The bold-faced numbers indicate that JPP is significantly better than the other methods (p value <0.01 in Wilcoxon paired test). The values are averaged over 91 folds Yahoo News. λ is set to 10^7.

6.2.2 *Experimental setup*

We learn the models using the same time-window for all the algorithms (one week for the semi-synthetic data sets, and one day for Yahoo News). The time window depends on the maximum amount of data we can easily store and manage, and the frequency at which the user wants to discover or track topics.

We set the parameters of the different methods in the following way: after validation on an independent news data set, the $l1$-norm regularization parameters of the t-model, the *fix*-model, ONMF and JPP are fixed to 0.05 and kept to this value along all the experiments. All other parameters of the baseline methods are internally tuned on three time-slots discarded afterwards from the evaluation.

6.2.3 *Parameter analysis*

Having a clear intuition about the hyper-parameter λ constitutes a major challenge for the reusability of the JPP algorithm. When facing a concrete data set the end-user should be able to easily set the appropriate value for λ. We run an experiment on each synthetic dataset to improve our understanding on how the model behaves on different evolving patterns. We learn all the baseline models and JPP for different values of λ using a 'moving window' for the starting period. We iteratively start from $1, 2, ..., N_t - 1$ and use the remaining time steps as the *evaluation period*. This strategy results in defining $N_t(N_t - 1)/2$ folds (where N_t is the total number of time slots) on which the performance metrics are averaged. For instance, we use $N = 14$ days for Yahoo News dataset and, thus, we average over 91 folds. We learn the *fix*-model in the *starting period*. Later, on each time step of the *evaluation period*, we learn the t-model, ONMF, ODNMF, DNMF and JPP with $\mathbf{H}^{(t-1)}$ as memory parameter. For each row of $\mathbf{H}^{(t)}_{\text{true}}$ the closest recovered topic is calculated (i.e. the one for which the cosine similarity with the ground-truth is maximal). Finally, the detected topics are evaluated against the ground-truth using different performance metrics commonly used in information retrieval: microF1, Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) [18].

The periodic dataset consists of topics that emerge and fade alternately, a scenario in which focussing completely on the past will fail during transiting periods (when the topic just appear or vanish), while being completely present oriented leads to miss the opportunity of learning from past occurred topics. A cursory inspection of the averaged NDCG scores, shown in Figure 2, confirms this intuition where the best value of λ for the *periodic* pattern is 1 which reflects

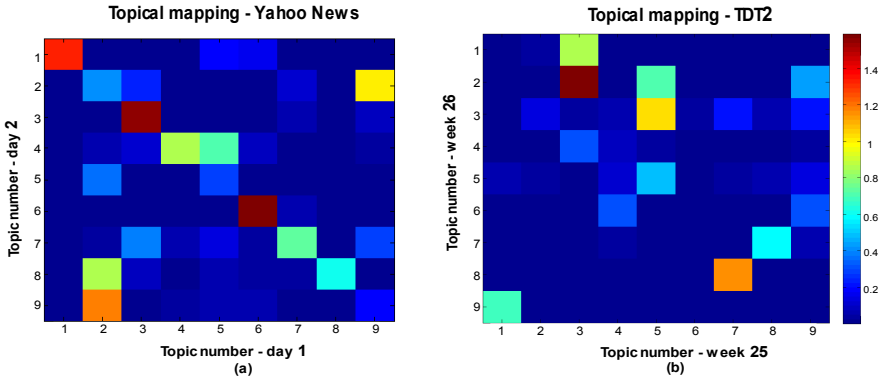

Figure 4: Heatmap of matrix $\mathbf{M}^{(t)}$. Color intensity in the entry m_{ij} shows the extent to which topic i discovered at time t can be explained by the topic j discovered at time $t-1$. Therefore, full blue rows represent new emerging topics, full blue columns represent fading topics while brighter cells i, j indicate an evolution of topic j to topic i. (a) $\mathbf{M}^{(2)}$ matrix obtained by the JPP algorithm from the Yahoo News dataset. It shows the mapping between $k = 9$ topics discovered at time *2* with those obtained at time *1* (b) $\mathbf{M}^{(26)}$matrix obtained by the JPP algorithm from the TDT2 dataset. It shows the mapping between $k = 9$ topics discovered at time *26* with those obtained at time *25*. Keywords representing each topic are given in Tables 2 and 3.

Trend	week 25	week 26	Topic
Evolving	3	2	percent, yen, economy, market, economic, japan, Japanese hong, kong, handover, china, chief, economy
	7	8	nuclear, india, Pakistan, tests, test, indias, weapons Pakistan, india, Kashmir, sharif, border, Pakistani, nuclear
	5	3	chinese, china, cuba, rights, human, Clinton, tjananmeh dalai, lama, Tibet, tibetam, chinese, contacts, lamas
Emerging	-	4	Cities, expensive, ranking, Geneva, study, city, report
	-	7	aids, fallows, hiv, tobacco, Viagra, drug, complaint
Fading	2	-	Netanyahu, Israel, Israeli, Palestinian, Arafat, Peace
	6	-	Game, bulls, Jordan, Malone, Chicago, Pippen

Table 2: Topics mapping for two consecutive weeks (25 and 26) on the TDT2 dataset extracted from the matrix $\mathbf{M}^{(26)}$ generated in the JPP algorithm (Figure 4).

a balance between past-orientation and present-orientation. In this case, the performance of JPP is significantly better on all metrics for $\lambda \leq 10$ according to a paired signed Wilcoxon test with a p value < 0.01.

On the burst pattern, the model is robust for the different values of λ with a small advantage for past oriented models. This shows that the model is able to learn from the little amount of information occurring before the burst itself. This contrasts with ONMF and DNMF whose results are similar or worse (for DNMF) than the t-model highlighting their difficulties with early burst detection scenarios. On this pattern, JPP is significantly better on all metrics for all values of λ except for 0.1 where the t-model and ONMF achieve similar performance to JPP.

For the overlapping pattern, the model is robust to the choose of λ and does not suffer from multiple topics appearing in the same time. JPP is significantly better, on all metrics for $\lambda \geq 10$, than other methods except for the *t-model*.

On the mix pattern, JPP is achieving its best performance for $\lambda = 10$. However, JPP is significantly better than state-of-the-art on all metrics for all values of λ except for 1 where ODNMF achieves similar performance.

In summary, on these constructed patterns, JPP achieves very good performance by simply setting λ to a high value, in which case it outperforms the state-of-the-art. In the unique case of periodic patterns, it may be useful to balance more towards the current observation using a smaller value of λ. In the remainder, we propose to study the performance of the model with regard to state-of-the-art techniques for topic discovery.

6.2.4 *Discovery evaluation*

We conducted a second experiment using a real-world dataset: Yahoo News. Here, we aim to evaluate the performance of the JPP algorithm in a topic discovery task with respect to the baseline techniques. For assessing the overall models ability at recovering topics, we report results for three metrics: microF1, Mean Average Precision (MAP) and NDCG.

For the experimental setup, we follow the procedure described in the previous section. At the starting point, we learn the *fix*-model. On each of the remaining time-slots, we learn the t-model, ONMF, ODNMF, DNMF and JPP using $\mathbf{H}^{(t-1)}$ as memory parameter. At each iteration, the topics recovered by each model are compared to $\mathbf{H}^{(t)}_{\text{true}}$ in the same way as explained before. This operation is repeated for different number of topics (5, 10, 15, 30).

For 5, 15 and 30 topics the JPP algorithm is significantly outperforming the t-model, the *fix*-model, ONMF, ODNMF

Trend	19th Sep.	20th Sep.	Topic
Evolving	9	2	muslim, film, protest ,embassy, prophet ,islam, french film, protest, muslim,police, protest,islam, anti, pakistan
	7	7	syria, damascus, strike, helicopter, syrian, mine, iran, nuclear, rebel, opposition, syria, assad, security, iranian
Evolving-merging	4	-	romney, obama, republican,campagin,immigration,presidential, voter,candidate
	5	-	percent, bank, oil, price, market, economy, rate, obama
	-	4	romney, obama, campaign,tax, percent,republican, million,bank,president
Evolving-splitting	2	-	game, league, play, team, win, season, club, goal, player
		8	england, terry, chelsea, cup, captain, over, twenty
		9	game, season, win, score, liverpool, goal, second

Table 3: Topics mapping for two consecutive days (19th and 20th September 2012) on the Yahoo news dataset extracted from the matrix $\mathbf{M}^{(2)}$ generated by the JPP algorithm (Figure 4).

and DNMF (according to a paired signed Wilcoxon test with a p value < 0.01), for all λ greater that 10 (Table 1). For 10 topics the *fix*-model outperforms significantly the other models.

The best performance, generally obtained for a high value of λ on these different media sources, emphasizes that memorizing past events is worthwhile for news data.

6.2.5 Mapping evolving topics

Monitoring requires the ability to map, if possible, the set of topics discovered at time t with those discovered at time *t-1*. The matrix $\mathbf{M}^{(t)}$ used in our framework (Equation (2)) provides a clear insight into this mapping letting us to label topics at time t as *emerging, evolving* or *fading*. The entry m_{ij} of $\mathbf{M}^{(t)}$ indicates to which extent topic i at time t can be explained by the topic j discovered previously at time $t-1$. In a heatmap of $\mathbf{M}^{(26)}$ (Figure 4), full blue rows represent new emerging topics, full blue columns represent fading topics while brighter cells tell us that the topic in column j has evolved into topic in row i.

We analyze the contents of $\mathbf{M}^{(2)}$ and $\mathbf{M}^{(26)}$ computed on Yahoo News and the complete TDT2 data set respectively for 9 topics during two consecutive time slots (from day 1 to 2 and week 25 to 26). The topical mapping obtained exhibits the three trending patterns: *evolving, emerging* and *fading* using a λ value set to 10 (i.e. the best choice of λ for Yahoo News with k$\sim$9, see Figure 3).

A closer inspect to the topics extracted from the TDT2 dataset (Table 2) shows the evolution of topic 5 that refers to *china*[3], *tjananmeh* square and *human rights*, to topic 3 dealing with *dalai lama* and the province of *Tibet*. A better example of evolution is given by topic 7 dealing with the *nuclear tests* in *India* and *Pakistan* and its evolution to the problem of the *Kashmir* territory disputed by both countries. The last example are the evolution of news subjects from the *Japanese economy market* to the *hong kong china economy* which plays as *japan* an important role in the east asia *economy*. The system helps also the news consumer to figure out which are the novel topics of the week: for instance *expensive*, *city*, *report* for *Geneva*. While some last week topics seems to be no more frontpage headlines, for instance the *Israel Palestinian* conflict or the *Chicago bulls Game* with Michael *Jordan*.

Events extracted from the Yahoo News dataset shows patterns helping the reader to contextualize and connect topics occurring the 19th and the 20th September 2012. The two interesting patterns are: (1) two different topics merging at the next time-slot, and (2) one unique topic splitting in two in the next period of time (Table 3). The first case is illustrated by topics 4 and 5 that merge: topic 4 relates with the *campaign* that opposed *obama* and *romney* centered around *immigration*; while topic 5 relates with the *obama economy* policy and his strategy for *market* and *banks*. These two topics merge in the next time step forming one unique topic about the *campaign* but this time centered around the *economy* and *tax* policy. The second pattern is illustrated by topic 2, concerned with the english football premier *league*, that splits into topic 9 covering still mainly the same topic ($game, season, goal$) and the more specific topic 8 about John *terry*, the *captain* of the *chelsea* football club. A deeper analyze of the related articles indicates that the news at that time period have been covering legal issues faced by John *terry* justifying the detection of this specific topic.

We also observe two cases of simple evolution: topic 9 and 7 evolving to topic 2 and 7 respectively. Topic 9 describes the *protests* at the *embassy* against a *french* magazine and a *film* that offended *muslims* showing images of the *prophet*. The topic shifts towards *protests* happening in *pakistan* about this case. Topic 7 and its shifts connect events happening in *syria*.

6.2.6 Running time analysis

During the discovery process, the computation time is crucial for online services. Depending on the source of syndication, e.g. news or social media, the number of observations to handle on a small period of time (e.g. minutes or hours) can grow from a few hundreds to a few thousand observations. In order to assess the ability of the proposed approach to handle a large amount of information produced in a short time period we divided the Yahoo News data in two time slots. The first time period, with 1,171 news articles, is kept as the starting period while the documents arriving from day 2 to day 14th, 12,148 news articles, are merged into one unique time-slot. On this time slot, we measure the computation time taken by the JPP algorithm in seconds for an increasing value of λ for 5 topics averaged on 10 runs. The average computation time decreases from 249 seconds to 64 seconds as λ increases (Figure 5). For high values of λ, the required computation time is even lower than for the basic NMF implementation (64 seconds for JPP while 180s for NMF). This is a benefit of the temporal regularization of the objective function, making it converge faster to a local minimum. Notice the high computation time required by DNMF, approximatively 8 times slower than JPP for a high value of λ (in which case JPP delivers its best performance). ODNMF is the fastest approach of our benchmarks

[3]Terms in italic correspond to the keywords of topics descriptions as found by the algorithm, see Tables 2 and 3

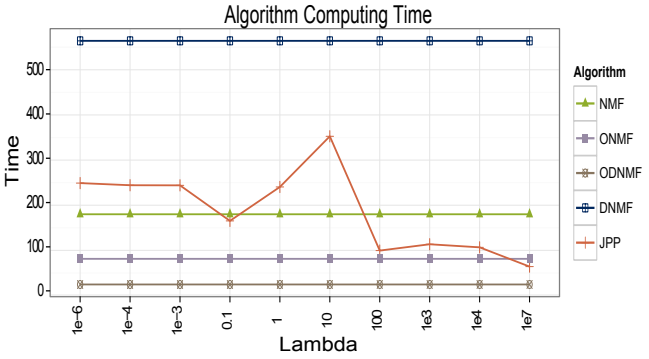

Figure 5: Computation time averaged over 10 runs of each algorithm for 5 topics in the Yahoo News dataset where documents are grouped in two timeslots. Experiments were run on a Intel Xeon CPU X5650, 2.67GHz(6 cores), 64GB of RAM.

with an average computing time of around 25 seconds. This can be partly explained by an efficient implementation of the software and the warm-up initialization strategy helping to converge faster.

6.3 Topic Tracking

The topic tracking task consists of associating incoming stories with topics that are known to the system (see [7]). The known topics are provided by the end-user in terms of keywords or by any automatic system. Having a set of known topics $\mathbf{H}_{\text{known}}$ the goal is to associate with them new incoming data. In other words, the goal is to predict $\mathbf{W}^{(t)}$ from a given $\mathbf{X}^{(t)}$ and $\mathbf{H}_{\text{known}}$. In this context, starting from Equation (5) we derive a simple extension of the JPP algorithm for topic tracking where $\mathbf{H}$ is a fixed parameter. Following the same strategy, we can derive a classification algorithm for NMF. In this setting, when $\mathbf{H}$ is a fixed parameter, the Online-NMF reduces to NMF.

As experiments, we test both algorithms (i.e. JPP and NMF) on Yahoo News data sets selecting the documents on the top 15 and 30 topics. The first half of the crawl was used to tune the λ parameter. On the second half, the starting time-slot was used as training set to compute $\mathbf{H}_{\text{known}}$. The learned topics are afterwards tracked on the remaining periods where we predict $\widehat{\mathbf{W}}^{(t)}$. Table 4 reports the averaged performance on these periods in terms of microF1, MAP and NDCG. We use $N = 7$ days for Yahoo News dataset and, thus, we average over 21 folds.

The JPP algorithm performs better than NMF for all performance metrics (according to a paired signed Wilcoxon test with a p value < 0.01). Figure 6 reports the tracking of the 5 top topics of Yahoo News data set during the 6 last days of the crawl. The trends of each topic (increasing and decreasing) are correctly detected by the JPP algorithm. The ground-thruth on this Figure is the average of number documents belonging to a specific topic at each time step. The topic intensity is an estimation of the p(topic|time step) computed from $\mathbf{W}$ where (after a $l1$-norm normalization of each row) we consider each entry w_{ij} to be an estimation of p(topic|document).

This experiment is a proof of concept, validating the applicability of the regularization framework proposed in this paper for a tracking (or more generally temporal text classification) system. As future work, it could be interesting to adapt the proposed temporal regularization to other related algorithms as for instance the PLSA.

k	Alg.	microF1	macroF1	MAP	NDCG
15	NMF	0.56 ± 0.02	0.54 ± 0.02	0.68 ± 0.02	0.80 ± 0.01
	JPP	**0.58** ± 0.02	**0.56** ± 0.02	**0.70** ± 0.02	**0.81** ± 0.01
30	NMF	0.45 ± 0.04	0.42 ± 0.03	0.58 ± 0.03	0.73 ± 0.02
	JPP	**0.48** ± 0.03	**0.46** ± 0.03	**0.61** ± 0.03	**0.75** ± 0.02

Table 4: Topic classification evaluation on the Yahoo News data set using four metrics. The bold-faced numbers indicate that JPP is significantly better than the other methods (p value <0.01 in Wilcoxon paired test). The values are averaged over 21 folds.

7. CONCLUSIONS

In this work, we introduced a novel collective time-based collective factorization algorithm for topic discovery and monitoring of evolving input streams. Our approach in based on an NMF multiplicative-updates framework having one *unique* hyper-parameter controlling the trade-off between the memory and the current observation. The model provides for free a simple way to discover trends: *emerging*, *evolving* and *fading* topics. This gives the opportunity to the news consumer to construct a topical map helping him in contextualizing the continuous flow of information.

We showed, on different media sources, that the model automatically finds a good balance between current and past observations; henceforth, outperforming in many cases the state-of-the-art on both tasks: topic discovery and topic tracking. In terms of computation time the approach can easily handle a large number of documents (more than 10 thousands) in a few minutes and therefore could easily be used to discover topics in an online environment.

8. ACKNOWLEDGMENTS

This research was partially funded by ESPOL, the Ecuadorian agency SENESCYT and partially funded by the European Union 7th Framework Programme ARCOMEM and Social Sensor projects, by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037 “Social Media”. We are thankful to Barla Cambazoglu for providing the Yahoo News crawl data. We are also thankful to Nicola Barbieri, Jean-Michel Renders, Ilaria Bordino for their comments.

9. REFERENCES

[1] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.

[2] D.M. Blei and J.D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] Bin Cao, Dou Shen, Jian-Tao Sun, Xuanhui Wang, Qiang Yang, and Zheng Chen. Detect and track latent factors with online nonnegative matrix factorization.

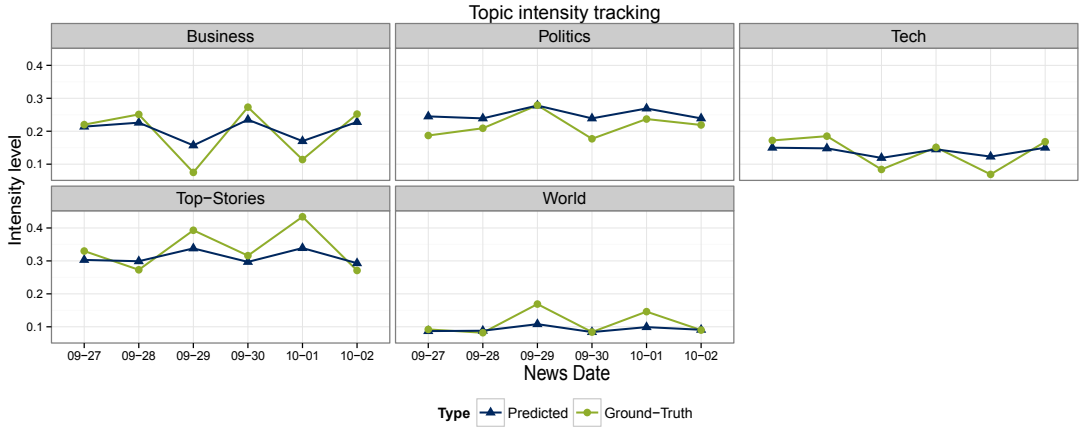

Figure 6: Evolution of the topic intensity on Yahoo News categories from Sep 27th to October 2nd, 2012. The five top categories (Business, Politics, Tech, Top-Stories, World) are shown.

In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12*, 2007.

[5] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[6] Lee Daniel and Seung Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[7] Jonathan G. Fiscus, George R. Doddington, John S. Garofolo, and Alvin F. Martin. Nist's 1998 topic detection and tracking evaluation (tdt2). In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9*, 1999.

[8] T. Fukuhara, T. Murayama, and T. Nishida. Analyzing concerns of people using weblog articles and real world temporal data. In *Proceedings of WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2005.

[9] Qi He, Kuiyu Chang, Ee-Peng Lim, and A. Banerjee. Keep it simple with time: A reexamination of probabilistic topic detection models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(10):1795 –1808, oct. 2010.

[10] Ngoc-Diep Ho. *Nonnegative matrix factorization algorithms and applications*. PhD thesis, Université Catholique de Louvain, 2008.

[11] Shiva Prasad Kasiviswanathan, Prem Melville, Arindam Banerjee, and Vikas Sindhwani. Emerging topic detection using dictionary learning. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28*, pages 745–754, 2011.

[12] Noriaki Kawamae. Trend analysis model: trend consists of temporal words, topics, and timestamps. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 317–326. ACM, 2011.

[13] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13, (NIPS), Denver, CO, USA*, pages 556–562, 2000.

[14] D.D. Lee, H.S. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[15] J. Lehmann, B. Gonçalves, J.J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 251–260. ACM, 2012.

[16] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning*, pages 665–672. ACM, 2009.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[18] Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[19] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, pages 1155–1158. ACM, 2010.

[20] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30(1):249–272, 2007.

[21] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[22] Bree Nordenson. Overload! *Columbia Journalism Review*, 47(4):30–32, 2008.

[23] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to

twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[24] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 569–577. ACM, 2008.

[25] Ankan Saha and Vikas Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization. In *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12*, pages 693–702, 2012.

[26] Yuichiro Sekiguchi, Harumi Kawashima, Hidenori Okuda, and Masahiro Oku. Topic detection from blog documents using usersÕ interests. In *Mobile Data Management, 2006. MDM 2006. 7th International Conference on*, pages 108–108. IEEE, 2006.

[27] D.A. Shamma, L. Kennedy, and E.F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 355–358. ACM, 2011.

[28] Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *ACM Conference on Knowledge Discovery and Data Mining*, 2008.

[29] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 424–433, 2006.

[30] Yu Wang, Eugene Agichtein, and Michele Benzi. Tm-lda: efficient online modeling of latent topic transitions in social media. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16*, pages 123–131, 2012.

APPENDIX

We prove here that the loss function L in Equation (5) is non increasing under the update rules in Equations (12), (13) and (14). A similar proof has been derived for the standard NMF [10].

Notice, first, that we may rewrite the cost function L as follows:

$$L = \underset{\mathbf{W}^{(t)},\mathbf{H}^{(t)},\mathbf{M}^{(t)}}{\arg\min} \sum_{i=1}^{n} \|\mathbf{X}_{i:}^{(t)} - \mathbf{W}_{i:}^{(t)}\mathbf{H}^{(t)}\|_2^2 + \sum_{i=1}^{n} \|\mathbf{X}_{i:}^{(t)} - \mathbf{W}_{i:}^{(t)}\mathbf{M}^{(t)}\mathbf{H}^{(t-1)}\|_2^2 + \lambda\|\mathbf{M}^{(t)} - \mathbf{I}\|_F^2 + \alpha\|\mathbf{H}^{(t)}\|_1 + \beta\sum_{i=1}^{n}\|\mathbf{W}_{i:}^{(t)}\|_1 + \gamma\|\mathbf{M}^{(t)}\|_1 \quad (15)$$

While fixing $\mathbf{H}^{(t)}$ and $\mathbf{M}^{(t)}$, one can separately minimize L with respect to each row $\mathbf{w}^{\mathrm{T}}$ of $\mathbf{W}^{(t)}$ and $\mathbf{x}^{\mathrm{T}}$ of $\mathbf{X}^{(t)}$:

$$\underset{\mathbf{w}^{\mathrm{T}}}{\arg\min}\, L(\mathbf{w}^{\mathrm{T}}) = \underset{\mathbf{w}^{\mathrm{T}}}{\arg\min} \|\mathbf{x}^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}}\mathbf{H}^{(t)}\|_2^2 + \|\mathbf{x}^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}}\mathbf{M}^{(t)}\mathbf{H}^{(t-1)}\|_2^2 + \beta\|\mathbf{w}^{\mathrm{T}}\|_1$$

Consider a current approximation $\hat{\mathbf{w}}^{\mathrm{T}} > 0$ of the solution and formulate the following problem:

$$\underset{\mathbf{w}^{\mathrm{T}}}{\arg\min}\, \widehat{L}(\mathbf{w}^{\mathrm{T}}) = \underset{\mathbf{w}^{\mathrm{T}}}{\arg\min} \|\mathbf{x}^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}}\mathbf{H}^{(t)}\|_2^2 + \|\mathbf{x}^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}}\mathbf{M}^{(t)}\mathbf{H}^{(t-1)}\|_2^2 + \beta\|\mathbf{w}^{\mathrm{T}}\|_1 + (\mathbf{w}^{\mathrm{T}} - \hat{\mathbf{w}}^{\mathrm{T}})^{\mathrm{T}}\mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}}(\mathbf{w}^{\mathrm{T}} - \hat{\mathbf{w}}^{\mathrm{T}}) \quad (16)$$

where $\mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}} = \mathbf{Diag}(x) - (\mathbf{H}^{(t)}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}})$ with **Diag** the diagonal operator creating a diagonal matrix from an input vector, and with $x = \frac{[\hat{\mathbf{w}}^{\mathrm{T}}(\mathbf{H}^{(t)}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}})]}{[\hat{\mathbf{w}}^{\mathrm{T}}]}$. Since we can prove that $\mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}}$ is semidefinite positive (see [13]), it follows that $\widehat{L}(\mathbf{w}^{\mathrm{T}}) \geq L(\mathbf{w}^{\mathrm{T}})$ for all $\mathbf{w}^{\mathrm{T}}$ with $\widehat{L}(\hat{\mathbf{w}}^{\mathrm{T}}) = L(\hat{\mathbf{w}}^{\mathrm{T}})$.

In order to obtain a minimizer $\mathbf{w}^{\mathrm{T}*}$ of $\widehat{L}$,

We set $\nabla_{\mathbf{w}^{\mathrm{T}}}\widehat{L}$ to zero, and obtain:

$$\nabla_{\mathbf{w}^{\mathrm{T}}}\widehat{L} = \nabla_{\mathbf{w}^{\mathrm{T}}}L + (\mathbf{w}^{\mathrm{T}} - \hat{\mathbf{w}}^{\mathrm{T}})\mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}} = 0 \quad (17)$$

and deduce the following minimizer $\mathbf{w}^{*\mathrm{T}}$

$$\mathbf{w}^{*\mathrm{T}}(\mathbf{H}^{(t)}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}} + \mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}}) = \mathbf{x}^{\mathrm{T}}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{x}^{\mathrm{T}}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}} - \beta + \mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}}\hat{\mathbf{w}}^{\mathrm{T}} \quad (18)$$

Since

$$\hat{\mathbf{w}}^{\mathrm{T}}(\mathbf{H}^{(t)}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}}) + \mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}} = \mathbf{Diag}\big(\hat{\mathbf{w}}^{\mathrm{T}}(\mathbf{H}^{(t)}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}})\big)\mathbf{Diag}(\hat{\mathbf{w}}^{\mathrm{T}})^{-1} \quad (19)$$

, and $\mathbf{S}_{\hat{\mathbf{w}}^{\mathrm{T}}}\hat{\mathbf{w}}^{\mathrm{T}} = 0$, we conclude

$$\mathbf{w}^{*\mathrm{T}} = \hat{\mathbf{w}}^{\mathrm{T}} \odot \frac{[\mathbf{x}^{\mathrm{T}}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{x}^{\mathrm{T}}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}} - \beta]}{[\hat{\mathbf{w}}^{\mathrm{T}}(\mathbf{H}^{(t)}{\mathbf{H}^{(t)}}^{\mathrm{T}} + \mathbf{M}^{(t)}\mathbf{H}^{(t-1)}{\mathbf{H}^{(t-1)}}^{\mathrm{T}}{\mathbf{M}^{(t)}}^{\mathrm{T}})]} \quad (20)$$

which corresponds well to the update rule obtained Equation (13). Since $\mathbf{w}^{*\mathrm{T}}$ is a global minimizer of $\widehat{L}(\mathbf{w}^{\mathrm{T}})$, we have $\widehat{L}(\mathbf{w}^{*\mathrm{T}}) \leq \widehat{L}(\hat{\mathbf{w}}^{\mathrm{T}})$. Moreover, $\widehat{L}(\mathbf{w}^{\mathrm{T}})$ has been constructed is order to satisfy $\widehat{L}(\mathbf{w}^{\mathrm{T}}) \geq L(\mathbf{w}^{\mathrm{T}})$ for all $\mathbf{w}^{\mathrm{T}}$. This implies $L(\mathbf{w}^{*\mathrm{T}}) \leq \widehat{L}(\mathbf{w}^{*\mathrm{T}}) \leq \widehat{L}(\hat{\mathbf{w}}^{\mathrm{T}}) = L(\hat{\mathbf{w}}^{\mathrm{T}})$ which proves that L is nondecreasing under the update rule Equation (13). The same approach may be followed to prove that L is nondecreasing under the update rule in Equation (14) for $\mathbf{M}^{(t)}$. For update Equation (12) the proof remains the same as for NMF and can be found in [13].