Dynamic NMFs with Temporal Regularization for Online Analysis of Streaming Text

Ankan Saha
Department of Computer Science
University of Chicago, Chicago IL 60637
ankans@cs.uchicago.edu

Vikas Sindhwani
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
vsindhw@us.ibm.com

Abstract

Learning a dictionary of basis elements with the objective of building compact data representations is a problem of fundamental importance in statistics, machine learning and signal processing. In many settings, data points appear as a stream of high dimensional feature vectors. Streaming datasets present new twists to the problem. On one hand, basis elements need to be dynamically adapted to the statistics of incoming datapoints, while on the other hand, early detection of rising new trends is desirable in many applications. The analysis of social media streams formed by tweets and blog posts is a prime example of such a setting, where topics of social discussions need to be continuously tracked and new emerging themes are required to be rapidly detected. We formalize such problems in terms of online learning of dynamic non-negative matrix factorizations (NMF) with novel forms of temporal regularization. We describe a scalable optimization framework for our algorithms and report empirical results on detection and tracking of topics over simulated document streams and real-world news stories.

1 Introduction

We consider the problem of building compact, dynamic representations of streaming datasets such as those that arise in social media. By constructing such representations, "signal" can be separated from "noise" and essential data characteristics can be continuously summarized in terms of a small number of human interpretable components. In the context of social media applications, this maps to the discovery of unknown "topics" from a streaming document collection. Each new batch of documents arriving at a timepoint is completely unorganized and may contribute either to ongoing unknown topics of discussion (potentially causing underlying topics to drift over time) and/or initiate new themes that may or may not become significant going forward, and/or simply inject irrelevant noise. In this paper, we describe an online learning framework to consistently reassemble the data stream into coherent threads of evolving components while also serving as an "early warning" system for new, rapidly emerging trends.

While the dominant body of previous work in dictionary learning and topic modeling has focussed on solving batch learning problems, a real deployment scenario in social media applications truly requires forms of online learning. The user of such a system is less interested in a one-time analysis of topics in a document archive, and more in being able to follow ongoing evolving discussions and being vigilant of any emerging themes that might require immediate action. Tracking temporal variations in social media streams is attracting increasing interest[17]. Several papers have proposed dynamic topic and online dictionary learning models (see [3, 11, 4, 9, 14, 18, 2] and references therein) that either exploit temporal order of documents in offline batch mode (using variational inference or Gibbs sampling

techniques) or are limited to handling a fixed bandwidth of topics with no explicit algorithmic constructs to attempt to detect emerging themes early.

In this paper, we propose a framework for online dictionary learning to handle streaming non-negative data matrices with possibly growing number of components. Our methods are rooted in non-negative matrix factorizations (NMF) [12, 16] whose unregularized variants for (generalized) KL-divergence minimization can be shown to be equivalent to pLSI [7], a classic probabilistic topic modeling algorithm. For squared loss, NMF finds a low-rank approximation to a data matrix $\mathbf{X}$ by minimizing $\|\mathbf{X} - \mathbf{WH}\|_{fro}^2$ under non-negativity and scaling constraints on the factors $\mathbf{W}$ and $\mathbf{H}$. It is common to add some form of l_1/l_2 regularization e.g., to encourage sparse factors and prevent overfitting. If $\mathbf{X}$ is an $N \times D$ document-term matrix, then $\mathbf{W}$ is a $N \times K$ matrix of topic encodings of documents while $\mathbf{H}$ is a $K \times D$ matrix of topic-word associations, whose rows are the dictionary elements learnt by the NMF approach.

Given streaming matrices, we learn a sequence of NMFs with two forms of temporal regularization. The first regularizer enforces smooth evolution of topics via constraints on amount of drift allowed. The second regularizer applies to an additional "topic bandwidth" introduced into the system for early detection of emerging trends. Implictly, this regularizer extracts smooth trends of candidate emerging topics and then encourages the discovery of those that are rapidly growing over a short time window. We formulate this setup as minimization of an objective function which can be reduced to rank-one subproblems involving projections onto the probability simplex and SVM-like optimization with additional non-negativity constraints. We develop efficient algorithms for finding stationary points of this objective function. Since they mainly involve matrix-vector operations and linear-time subroutines, our algorithms scale gracefully to large datasets. We empirically study how temporal priors affect the quality of detection and tracking in streaming topic modeling problems.

It should be noted that our work is on a different vein than the topic detection and tracking methods previously developed in the information retrieval community [6, 1]. These methods typically work with an initial collection of training documents known to be on certain topics. Then as streaming documents come in, their proximity to training documents is computed using some measure of document similarity. If the similarity is less than a threshold, a new topic is considered to have emerged, else the document is merged with known topics (and the topic centroid is updated). Such an approach is very sensitive to noise in the stream - each "noisy" document is a potential candidate for a new topic. In contrast, our approach is more robust to noise as it works with a short time window and separates signal from noise based on how well a candidate topic reconstructs the recent elements of the data stream while showing a rising temporal trend. We employ constrained optimization and trend filtering techniques towards these objectives.

In the sequel we abuse notation to denote $\mathbf{h}_i$ as the i^{th} row of $\mathbf{H}$ and $\mathbf{h}_{ij} = \mathbf{H}_{ij}$. Δ_D denotes the D dimensional simplex. $[K]$ refers to the set $\{1, 2 \ldots K\}$ and $\mathbf{0}$ refers to the vector of all 0's of appropriate dimension.

2 Dynamic Dictionary Learning

Let $\{\mathbf{X}(t) \in \mathbb{R}^{N(t) \times D}, t = 1, 2...t, ..\}$ denote a sequence of streaming matrices where each row of $\mathbf{X}(t)$ represents an observation whose timestamp is t. In topic modeling applications over streaming documents, $\mathbf{X}(t)$ will represent the highly sparse document-term matrix observed at time t. We use $\mathbf{X}(t_1, t_2)$ to denote the document-term matrix formed by vertically concatenating $\{\mathbf{X}(t), t_1 \leq t \leq t_2\}$. At the current timepoint t, our model consumes the incoming data $\mathbf{X}(t)$ and generates a factorization $(\mathbf{W}(t), \mathbf{H}(t))$ comprising of $K(t)$ topics. The design of this factorization follows from two considerations: (1) The first $K(t-1)$ topics in $\mathbf{H}(t)$ must be smooth evolutions of the $K(t-1)$ topics found upto the previous timepoint, $\mathbf{H}(t-1)$. We call this the *evolving set* and introduce an evolution parameter, δ which constrains the evolving set to reside within a box of size δ on the probability simplex around the previously found topics. With minor modifications, δ can also be made topic or word-specific e.g., to take topic volatility or word dominance into account. (2) The second consideration is the fast detection of emerging topics. At each timepoint, we inject additional

topic bandwidth for this purpose. We call this the *emerging set*. Thus the topic variable $\mathbf{H}(t)$ can be partitioned into an evolving set of $K(t-1)$ topics, $\mathbf{H}^{ev}$, and an emerging set of K^{em} topics $\mathbf{H}^{em}$. Furthermore, we assume that emerging topics can be distinguished from noise based on their temporal profile. In other words, the number of documents that a true emerging topic associates with begins to rapidly increase. For this purpose, we introduce a short sliding time window ω over which trends are estimated. In the following section, we define a novel regularizer $\Omega(\mathbf{W}^{em})$ that consumes the document-topic associations for the emerging bandwidth and penalizes components that are static or decaying so that learnt emerging topics are more likely to be ones that are rising in strength. (3) We assume that topics in the emerging set become part of the evolving set going forward, unless some of them are discarded as noise by manual guidance from the user or using criteria such as net current strength. In our experiments, we retain all topics in the emerging set.

The discussion above motivates the following objective function that is optimized at every timepoint t.

$$(\mathbf{W}^{\star}, \mathbf{H}(t)) = \underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} \|\mathbf{X}(t-\omega, t) - \mathbf{WH}\|_{fro}^2 + \mu\Omega(\mathbf{W}) \tag{1}$$

This objective function is minimized under the following constraints.

$$\mathbf{W}, \mathbf{H} \geq 0 \tag{2}$$

$$\sum_{j=1}^{D} \mathbf{H}_{ij} = 1 \qquad \forall i \in [K(t-1) + K^{em}] \tag{3}$$

$$\min(\mathbf{H}_{ij}(t-1) - \delta, 0) \leq \mathbf{H}_{ij} \leq \max(\mathbf{H}_{ij}(t-1) + \delta, 1), \qquad \forall i \in [K(t-1)], \forall j \in [D] \tag{4}$$

We then extract $\mathbf{W}(t)$ from the bottom rows of $\mathbf{W}^{\star}$ that correspond to $\mathbf{X}(t)$. The system is then said to have tagged the i^{th} document (row) in $\mathbf{X}(t)$ with the most dominating topic $\operatorname{argmax}_j \mathbf{W}(t)(i,j)$ which gives a clustering of documents. Note that the regularizer, $\Omega(\mathbf{W})$, defined in the next section, implicitly only operates on those columns of $\mathbf{W}$ that correspond to emerging topics.

The solution $\mathbf{W}^{\star}, \mathbf{H}(t)$ are also used for initializing parts of $\mathbf{W}, \mathbf{H}$ in the next run (details ommitted). This hot-start mechanism significantly accelerates convergence. In the next section, we define the emergence regularization operator $\Omega(\mathbf{W})$ and then present our optimization algorithm.

3 Emergence Regularization

In this section, we formulate the regularization operator $\Omega(\mathbf{W})$ by chaining together trend extraction with a margin-based loss function to penalize static or decaying topics. We begin with a brief introduction to trend filtering.

Hodrick-Prescott (HP) Trend Filtering: Let $\{y_t\}_{t=1}^T$ be a univariate time-series which is composed of an unknown, slowly varying trend component $\{x_t\}_{t=1}^T$ perturbed by random noise $\{z_t\}_{t=1}^T$. Trend Filtering is the task of recovering the trend component $\{x_t\}$ given $\{y_t\}$. The Hodrick-Prescott filter is an approach to estimate the trend assuming that it is smooth and that the random residual is small. It is based on solving the following optimization problem:

$$\underset{\{x_t\}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{T}(y_i - x_i)^2 + \lambda \sum_{t=2}^{T-1} \left((x_{t+1} - x_t) - (x_t - x_{t-1})\right)^2 \tag{5}$$

Let us introduce the second order difference matrix $\mathbf{D} \in \mathbb{R}^{(T-2)\times T}$ such that

$$\mathbf{D}(i,i) = 1 \qquad \mathbf{D}(i,i+1) = -2 \quad \text{and} \qquad \mathbf{D}(i,i+2) = 1 \qquad \forall i \in [T-2]$$

Then, it is easy to see that the solution to the optimization problem of Equation 5 is given by, $\mathbf{x} = [I + 2\lambda\mathbf{D}^{\top}\mathbf{D}]^{-1}\mathbf{y}$, where we use the notation $\mathbf{y} = (y_1 \dots y_T)^T, \mathbf{x} = (x_1 \dots x_T)^T$. We use F to denote $[I + 2\lambda\mathbf{D}^{\top}\mathbf{D}]^{-1}$, the linear smoothing operator associated with the

Hodrick-Prescott Filter. Given the time series $\mathbf{y}$, the Hodrick-Prescott (HP) trend estimate simply is $\mathbf{x} = F\mathbf{y}$.

Loss Function for Measuring Emerging Trend: Let $\mathbf{x} = F\mathbf{y}$ be the HP trend of the time series $\mathbf{y}$. Let $\mathcal{D}$ be the forward difference operator, i.e., the only non-zero entries of $\mathcal{D}$ are: $\mathcal{D}_{i,i} = -1$ and $\mathcal{D}_{i,i+1} = 1$. If $\mathbf{z} = \mathcal{D}\mathbf{x}$, then $z_i = x_{i+1} - x_i$ reflects the discrete numerical gradient in the trend x. Given z_i, we define a margin based loss function (the ℓ_2 hinge loss), $L(z_i) = c_i \max(0, \delta - z_i)^2$, where if the growth in the trend at time i is *sufficient*, i.e., greater than δ, the loss evaluates to 0. If the growth is insufficient, the loss evaluates to $c_i(\delta - z_i)^2$ where c_i is the weight of timepoint i which typically increases with i. For a vector $\mathbf{z}$, the loss is added over the components. In terms of the original time series $\mathbf{y}$, this loss function is,

$$L(\mathbf{y}) = \sum_{i=1}^{T-1} c_i \max(0, \delta - (\mathcal{D}F\mathbf{y})_i)^2 \tag{6}$$

Optimization Problem: As documents arrive over $t \in [T]$, we use $\mathbf{S}$ to denote a $T \times N$ time-document matrix, where $\mathbf{S}(i,j) = 1$ if the document j has time stamp i. Noting that each column $\mathbf{w}$ of $\mathbf{W}$ denotes the document associations for a given topic, $\mathbf{Sw}$ captures the time series of total contribution of the topic $\mathbf{w}$. Finally, we concretize (1) as the following optimization problem

$$\underset{\mathbf{W},\mathbf{H}\geq\mathbf{0}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{WH}\|_{fro}^2 + \mu \sum_{\mathbf{w}_i \in \mathbf{W}^{em}} L(\mathbf{S}\mathbf{w}_i) \tag{7}$$

subject to constraints in equations 3 and 4.

4 Optimization algorithms

We approximate $\mathbf{X}$ as the sum of rank-one matrices $\mathbf{w}_i\mathbf{h}_i^\top$ and optimize cyclically over individual $\mathbf{w}_i$ and $\mathbf{h}_i$ variables while keeping all other variables fixed. This results in three specific sub-problems, each of which requires an efficient projection of a vector onto an appropriate space. Optimization of rank-one subproblems has been previously shown to be very effective for standard NMFs [10, 5] and is also reminiscent of the K-SVD approach for dictionary learning [8].

Optimization over $\mathbf{h}_i$: Holding all variables except $\mathbf{h}_i$ fixed and omitting additive constants independent of $\mathbf{h}_i$, (7) can be reduced to $\operatorname{argmin}_{\mathbf{h}_i \in \mathfrak{C}} \left\|\mathbf{R} - \mathbf{w}_i\mathbf{h}_i^\top\right\|_{fro}^2$ where $\mathbf{R} = \mathbf{X} - \sum_{j \neq i} \mathbf{w}_j\mathbf{h}_j^\top$ is the residual matrix, independent of $\mathbf{h}_i$. Simple algebraic operations yield that the above is equivalent to

$$\underset{\mathbf{h}_i \in \mathfrak{C}}{\operatorname{argmin}} \left\|\mathbf{h}_i - \mathbf{R}^\top\mathbf{w}_i / \|\mathbf{w}_i\|^2\right\|^2 \tag{8}$$

Case 1: $\mathbf{h}_i$ is evolving: For an evolving topic, the optimization needs to be performed under the constraints (4) and (3). Thus the optimum $\mathbf{h}_i^\star$ is obtained by projection onto the set $\mathfrak{C} = \{\mathbf{h}_i : \mathbf{h}_i \in \Delta_D, l_j \leq \mathbf{h}_{ij} \leq u_j\}$ for appropriate constants l_j and u_j. This is equivalent to a projection onto a simplex with box constraints. Adapting a method due to [15], we can find the minimizer in $O(D)$ time *i.e.* linear in the number of coordinates.

Case 2: $\mathbf{h}_i$ is emerging: For an emerging topic, $\mathfrak{C} = \{\mathbf{h}_i : \mathbf{h}_i \in \Delta_D\}$ and the optimization (8) becomes equivalent to a projection onto the simplex Δ_D. The same algorithm [15] again gives us the minimizer in linear time $O(D)$.

Optimization over evolving $\mathbf{w}_i$: When $\mathbf{w}_i \in \mathbf{W}^{ev}$, the second term in (7) does not contribute and the corresponding optimization problem boils down to $\mathbf{w}_i^\star = \operatorname{argmin}_{\mathbf{w}_i \geq \mathbf{0}} \left\|R - \mathbf{w}_i\mathbf{h}_i^\top\right\|^2$. Similar to (8), simple algebraic operations yield that the above minimization is equal to the following simple projection problem, $\operatorname{argmin}_{\mathbf{w}_i \geq \mathbf{0}} \left\|\mathbf{w}_i - \mathbf{R}\mathbf{h}_i / \|\mathbf{h}_i\|^2\right\|^2$ for appropriate residual matrix $\mathbf{R}$. The projection set

$\mathfrak{C}$ now is just the non-negative orthant, for which there is a closed form minimizer: $\mathbf{w}_i = \max\left(0, \frac{1}{\|\mathbf{h}_i\|^2}(\mathbf{R}\mathbf{h}_i)\right)$, where sparse matrix-vector products against $\mathbf{R}$ can be efficiently computed, i.e., $\mathbf{R}\mathbf{h} = \mathbf{X}\mathbf{h} - \sum_{j \neq i} \mathbf{w}_j(\mathbf{h}_j^\top \mathbf{h})$.

Emerging $\mathbf{w}_i$: When $\mathbf{w}_i \in \mathbf{W}^{em}$, the second term in (7) is active and the corresponding optimization problem looks like: $\operatorname{argmin}_{\mathbf{w}_i \geq \mathbf{0}} \left\|\mathbf{R} - \mathbf{w}_i\mathbf{h}_i^\top\right\|^2 + \mu L(S\mathbf{w}_i)$. Omitting the terms independent of $\mathbf{w}_i$, simple algebra yields that the above is equivalent to: $\operatorname{argmin}_{\mathbf{w}_i \geq \mathbf{0}} \left\|\mathbf{w}_i - R\mathbf{h}_i/\left\|\mathbf{h}_i\right\|^2\right\|^2 + \mu L(S\mathbf{w}_i)/\left\|\mathbf{h}_i\right\|^2$. Noting that we choose L to be the ℓ_2 hinge loss, this leads to,

$$\operatorname*{argmin}_{\mathbf{w}_i \geq \mathbf{0}} \left\|\mathbf{w}_i - R\mathbf{h}_i/\left\|\mathbf{h}_i\right\|^2\right\|^2 + \frac{\mu}{\left\|\mathbf{h}_i\right\|^2} \sum_{j=1}^{T-1} c_j \max(0, \delta_j - \mathbf{q}_j^\top \mathbf{w}_i)^2$$

where $\mathbf{q}_j^\top = (DFS)_{j,:}$ the j^{th} row of DFS. This can be converted into a generic minimization problem of the following form,

$$\min_{\mathbf{w} \geq \mathbf{0}} J(\mathbf{w}) = \sum_i \max\left(0, c_i(\delta_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)\right)^2 + \frac{\lambda}{2}\left\|\mathbf{w} - \mathbf{w}_0\right\|^2 \tag{9}$$

for some constant $\mathbf{w}_0$. This is precisely the SVM optimization problem with additional non-negativity constraints on $\mathbf{w}_i$. This objective is minimized using a projected gradient algorithm on the primal objective directly, as it is smooth and therefore the gradient is well defined. Thus $\mathbf{w}^{(k+1)} = \prod\left(\mathbf{w}^{(k)} - \eta_k \nabla J(\mathbf{w}^{(k)})\right)$ where $\prod$ is the projection operator $\prod(s) = \max(s, 0)$. The best rate η_k at the k^{th} step is chosen according to [13].

5 Empirical Studies

The goal of our empirical study is to understand the influence of temporal regularization (evolution and emergence parameters) on the effectiveness of topic detection and tracking. To enable quantitative evaluation, we presented two topic-labeled datasets to our algorithm as streams and the resulting topics generated by the system were benchmarked against ground truth topic assignments.

Datasets: We used two datasets for our experiments. The **Simulation** dataset consists of 1000 documents with 2500 terms divided into 25 topics accummulated over 31 days. We generated a (nearly) low-rank document-term matrix, $\mathbf{X} = \mathbf{W}\mathbf{H} + \mathbf{S}$, where $\mathbf{S}$ is a noise matrix with sparsity 0.001 and non-zero elements randomly drawn from a uniform distribution on the unit interval. This dataset comprises of 25 topics whose term-distributions (as specified by the 25 rows of $\mathbf{H}$) are random 2500-dimensional points on the topic simplex with sparsity 0.01. These topics are then randomly mixed (as specified in $\mathbf{W}$) to create the documents such that each topic dominates 40 documents with atleast 80% mixing proportions and each document on average contains 2.5 topics. These documents are then associated with timestamps such that topic $i, i > 5$ steadily emerges at timepoint i with a time profile as shown in the left subfigure in Figure 1. These emerging topics arise in the background of 5 initial static topics leading to an overall profile of temporal dynamics as shown (stacked area chart) in the right subfigure of Figure 1. We choose the hinge parameter to be $\mu = 5$ and emerging bandwidth of 1 per timepoint for this dataset. In our experiments, we use a sliding window of $\omega = 7$ timepoints. The second dataset is drawn from the Nist Topic Detection and Tracking (**TDT2**) corpus[1] which consists of news stories in the first half of 1998. In our evaluation, we used a set of 9394 documents represented over 19528 terms and distributed into the top 30 TDT2 topics over a period of 27 weeks. We choose the hinge parameter to be $\mu = 20$ and emerging bandwidth of 2 per week for this dataset. In our experiments, we use a sliding window of $\omega = 4$ weeks.

Evaluation Metrics: For tracking, we use $F1$ scores, as commonly reported in topic detection and tracking (TDT) literature. We point the reader to [4] for a precise definition of microaveraged $F1$ used in our experiments. We define a second performance

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

metric to capture how rapidly an emerging topic is "caught" and communicated to the user. Recall that a topic is communicated by the top keywords that dominate the associated term distribution in $\mathbf{H}(t)$. We first define true topic distributions as $\mathbf{H}^{true}(t) = \text{argmin}_{\mathbf{H}>0} \|\mathbf{X}(1,t) - \mathbf{W}^{true}\mathbf{H}\|^2_{fro}$ where $\mathbf{W}^{true}$ is set using true topic labels. Next, for each true topic i, we compute *first detection time*, which is the first timepoint at which the system generates a topic distribution in $\mathbf{H}(t)$ that is within a threshold of ϵ from the true topic, as measured by symmetric KL-divergence. We then record the percentage of documents missed before detection, and take the average of this miss rate across all true topics.

Results and Discussion: Figure 2 shows tracking performance as a function of the evolution parameter δ. When $\delta = 0$, the system freezes a topic as soon as it is detected not allowing the word distributions to change as the underlying topic drifts over time. When $\delta = 1$, the system has complete freedom in retraining topic distributions causing no single channel to remain consistently associated with an underlying topic. It can be seen that both these extremes are suboptimal. Tracking is much more effective when topic distributions are allowed to evolve under sufficient constraints in response to the statistics of incoming data. In Figure 2 we turn to the effectivness of emergence regularization. The figure shows how much information on average is missed before underlying topics are first detected, as a function of the emergence parameter μ. We see that increasing μ, for a fixed choice of δ, typically reduces miss rates causing topics to be detected early. As δ is increased, topics become less constrained and therefore provide additional bandwidth to drift towards emerging topics, therefore lowering the miss rate curves. However, this comes at the price of reduced tracking performance. Thus, for a fixed amount of available topic bandwidth, there is a tradeoff between tracking and early detection that can be navigated with the choice of μ and δ. Finally, the top keywords as per true word distributions (estimated by $\mathbf{H}^{true}$) and best matched system generated topics show excellent agreement (not shown for lack of space).

Figure 1: Temporal profile of an emerging topic and overall dynamics in Simulated dataset.

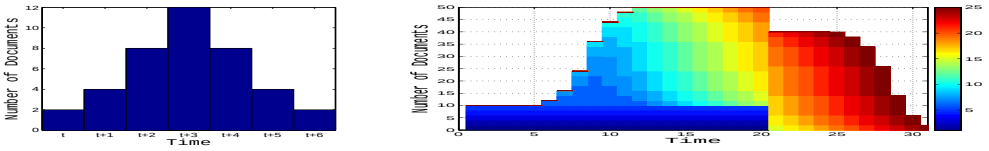
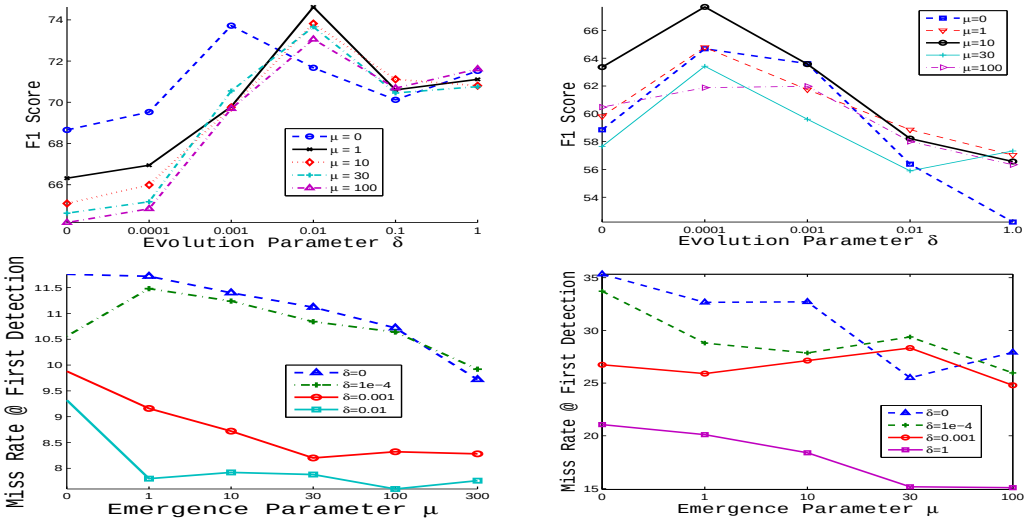


Figure 2: Evolution & Emergence : Simulated (left), TDT2 (right)



References

[1] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer International Series on Information Retrieval. Kluwer Academic Publ, 2002.

[2] Loulwah AlSumait, Daniel Barbara, and Carlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *International Conference on Data Mining*, 2008.

[3] D. Blei and J. Lafferty. Dynamic topic models. In *ICML*, 2006.

[4] Tzu-Chuan Chou and Meng Chang Chen. Using Incremental PLSI for Treshhold-Resilient Online Event Analysis. *IEEE transactions on Knowledge and Data Engineering*, 2008.

[5] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Non-negative and Tensor Factorizations: Applications to Exploratory Multiway Data Analysis and Blind Source Separation*. Wiley, 2009.

[6] Margaret Connell, Ao Feng, Giridhar Kumaran, Hema Raghavan, Chirag Shah, and James Allan. UMass at TDT 2004. 2004.

[7] C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorizations and probabilistic latent semantic analysis. *Computational Statistics and Data Analysis*, 2008.

[8] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[9] A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou. Topic evolution in a stream of documents. In *SDM*, 2009.

[10] Ngoc-Diep Ho, Paul Van Dooren, and Vincent D. Blondel. Descent methods for non-negative matrix factorization. *Numerical Linear Algebra in Signals*, abs/0801.3199, 2007.

[11] Matthew D. Hoffman, David M. Blei, and Frances Bach. Online learning for latent dirichlet allocation. In *NIPS*, 2010.

[12] D. Lee and H.S. Seung. Learning the parts of objects using non-negative matrix factorizations. *Nature*, 1999.

[13] C.J. Lin. Projected gradient methods for non-negative matrix factorization. In *Neural Computation*, 2007.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *JMLR*, 2010.

[15] P. M. Pardalos and N. Kovoor. An algorithm for singly constrained class of quadratic programs subject to upper and lower bounds. *Mathematical Programming*, 46:321–328, 1990.

[16] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 267–273, 2003.

[17] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *ACM International Conference on Web Search and Data Minig (WSDM)*. Stanford InfoLab, 2011.

[18] Limin Yao, David Mimno, and Andrew McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946. ACM, 2009.