Timelines: Recovering Birth and Evolution of Topics In Scientific Literature Using Dynamic Non-Parametric Bayesian Models

Amr Ahmed **Eric Xing**
School of Computer Science
Carnegie Mellon University

With the dramatic increase of digital document collections such as online journal articles, the Arxiv, conference proceedings, blogs, to name a few, there is a great demand for developing automatic text analysis models for analyzing these collections and organizing its content. Statistical admixture topic models [1] were proven to be a very useful tool to attain that goal and have recently gained much popularity in managing large collection of documents. Via an admixture model, one can project each document into a low dimensional space where their latent semantic (such as topical aspects) can be captured. This low dimensional representation can then be used for tasks like classifications, measuring document-document similarity or merely as a visualization tool that gives a bird's eye view of the collection and guides its exploration in a structured fashion.

An admixture topic model posits that each document is sampled from a fixed-dimensional mixture model according to a document's specific mixing vector over the *topics*. The variabilities in the topic mixing vectors of the documents are usually modeled as a Dirichlet distribution [1], although other alternatives have been explored in the literature [2, 3]. The components of this Dirichlet distribution encode the popularity of each of the topics in the collection. However, document collections often come as temporal streams where documents can be organized into epochs; examples of an epoch include: documents in an issue of a scientific journal, the proceeding of a conference in a given year, or the news articles published in a given week. Documents inside each epoch are assumed to be exchangeable while the order between documents is maintained across epochs. With this organizations, several aspects of the aforementioned static topic models are likely to change over time, specifically: topic *popularity*, topic *word distribution* and the *number* of topics.

Several models exist that could accommodate the evolution of some but not all of the aforementioned aspects. In [4], the authors proposed a dynamic topic model in which the topic's word distribution and popularity are linked across epochs using state space models, however, the number of topics are kept fixed. In [5], the authors presented the topics over time model that captures topic popularity over time via a beta distribution, however, topic distributions over words and the number of topics were fixed over time, although the authors discussed a non-parametric extension over the number of topics. On the other hand, several models were proposed that could *potentially* evolve all the aforementioned aspect albeit in a simple clustering settings, i.e. each document is assumed to be sampled from a single topic [6, 7, 8]. Accommodating the evolution of the aforementioned aspects in a full-fledged admixture setting is non-trivial and introduces its own hurdles. Moreover, it is widely accepted [1] that admixture models are superior compared to simple clustering models for modeling text documents, especially for long documents such as research papers.

In this paper we introduce iDTM: infinite dynamic topic models which can accommodate the evolution of the aforementioned aspects. iDTM allows for unbounded number of topics: topics can born and die at any epoch, the topics' word distributions evolve according to a first-order state space model, and the topics' popularity evolve using the rich-gets richer scheme via a Δ-order process. iDTM is built on top of the recurrent Chinese restaurant franchise (RCRF) process which we introduce and define in Section 2. The RCRF profecc introduces dependencies between the atom locations (topics) and weights (popularity) of each epoch-specific CRF process [9]. Inference is car-

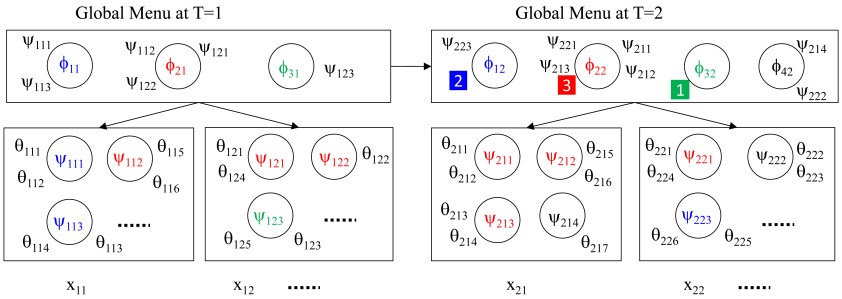

Figure 1: The recurrent Chinese restaurant franchise (RCRF) precoces. The figure shows a first-order process with no decay to avoid cluttering the display,however see the text for the description of a general Δ-order process.

ried via A Gibbs sampling algorithm and we give demonstrations over simulated and real datasets (leaving final full results to be presented at the workshop and the full version of the paper).

1 Settings and Background

In this section, we lay the foundation for the rest of this paper by first detailing our settings and then reviewing the CRFP. We are interested in modeling an ordered set of documents $\boldsymbol{x} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_T})$, where T denotes the number of epochs and $\boldsymbol{x_t}$ denotes the documents at epoch t. Furthermore, $\boldsymbol{x_t} = (\boldsymbol{x_{tj}})_{j=1}^{n_t}$, where n_t is the number of documents at epoch t. Moreover, each document comprises a set of n_{tj} words, $\boldsymbol{x_{tj}} = (x_{tji})_{i=1}^{n_{tj}}$, where each word $x_{tji} \in \{1, \cdots, \mathrm{W}\}$. Our goal is to discover *potentially* an unbounded number of topics $(\boldsymbol{\phi}_k)_{k=1}^{\infty}$ where each topic $\boldsymbol{\phi}_k = (\phi_{k,t_{k_1}}, \cdots, \phi_{k,t_{k_2}})$ spans a set of epoches where $1 \leq t_{k_1} \leq t_{k_2} \leq T$, and $\phi_{k,t}$ is the topic's word distribution at epoch t.

2 The Recurrent Chinese Restaurant Franchise Process

The recurrent Chinese restaurant franchise (RCRF) process shown in Figure 1 is a generalization of the CRF process introduced in [9]. Figure 1 depicts a RCRF process of order one for clarity, however, in this section we give a description of a general process of order Δ. The RCRF operates in epochs, where customers are not allowed to stay in any of the restaurants after the end of the epoch. At the end of epoch $t-1$, the consumption of the dishes ordered from the menu in any of the previous Δ epochs is analyzed. First any dish that was not ordered at least once in the last Δ epochs is removed, and then a time-weighted average usage of dish k is calculated as $m^{'}_{kt}$ where:

$$m^{'}_{kt} = \sum_{\delta=1}^{\Delta} \exp^{\frac{-\delta}{\lambda}} m_{k,t-\delta} \tag{1}$$

Where λ is the decay factor of the exponential decay time-kernel, Δ is its finite width, and $m_{k,t-\delta}$ is the number of tables served dish k in epoch $t-\delta$. Those dishes are inherited in the global menu at time t. Customer x_{tji} entering restaurant j at epoch t can sit on table b that has n_{tjb} customers and serves dish ψ_{tjb} with probability $\frac{n_{tjb}}{i-1+\alpha}$. He then shares this dish with those customers. Alternatively, he can choose to sit on a new table, b_{tj}^{new} with probability $\frac{\alpha}{i-1+\alpha}$ and orders a new dish. He has three alternatives. First, he can order an inherited dish ϕ_{kt} which is already served in at least one table in any restaurant, i.e. $m_{kt} > 0$ with probability $\frac{m_{kt}+m^{'}_{kt}}{\sum_{l=1}^{K_t} m_{lt}+m^{'}_{lt}+\gamma}$, K_t is the total number of dishes in the global menu at epoch t. Second, he can choose an inherited dish not served in any table in any restaurant, i.e. $m_{kt} = 0$, with probability $\frac{m_{kt}}{\sum_{l=1}^{K_t} m_{lt}+m^{'}_{lt}+\gamma}$, however in this case he can choose the cooking style of this dish: $\phi_{kt} \sim \mathrm{P}(.|\phi_{k,t-1})$. Finally, he can order an unplanned dish from the global menu, $\phi_{K_t^{new}} \sim H$, with probability $\frac{\gamma}{\sum_{l=1}^{K_t} m_{lt}+m^{'}_{lt}+\gamma}$ and increment K_t. Putting everything together, we have:

$$\theta_{tji}|\theta_{tj,1:i-1}, \alpha, \boldsymbol{\psi}_{t-\Delta:t} \sim \sum_{b=1}^{b=B_{tj}} \frac{n_{tjb}}{i-1+\alpha}\delta_{\psi_{tjb}} + \frac{\alpha}{i-1+\alpha}\delta_{\psi_{tjb^{\mathrm{new}}}} \tag{2}$$

$$\psi_{tjb^{new}}|\boldsymbol{\psi},\gamma \sim \sum_{k:m_{kt}>0} \frac{m_{kt}+m'_{kt}}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma}\delta_{\phi_{kt}} + \sum_{k:m_{kt}=0} \frac{m_{kt}+m'_{kt}}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma}\mathrm{P}(.|\phi_{k,t-1}) + \frac{\gamma}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma} H \quad (3)$$

where the first summand in (3 is over dishes served in at least one table, and the second summand is over dishes inherited but not yet ordered. Moreover, we have conveniently defined m'_{kt} to be zero for newly born dishes at epoch t, and m_{kt} to be zero for inherited but not yet ordered dishes at epoch t.

The hyperparameters of the RCRF process are given by: the concentration parameters α, γ which are endowed with an uninformative gamma prior, the variance parameter of the base measure H which is taking as a normal distribution, the parameters of the the topic's dynamic kernel $P(.|\phi)$ which is modeled as a radome walk, and the parameters of the time-kernel (λ, Δ). Inference is carried via a Gibbs sampling algorithm that takes into consideration the non-conjugacy between the normal base measure and the multinomial emission (details are removed for lack of space).

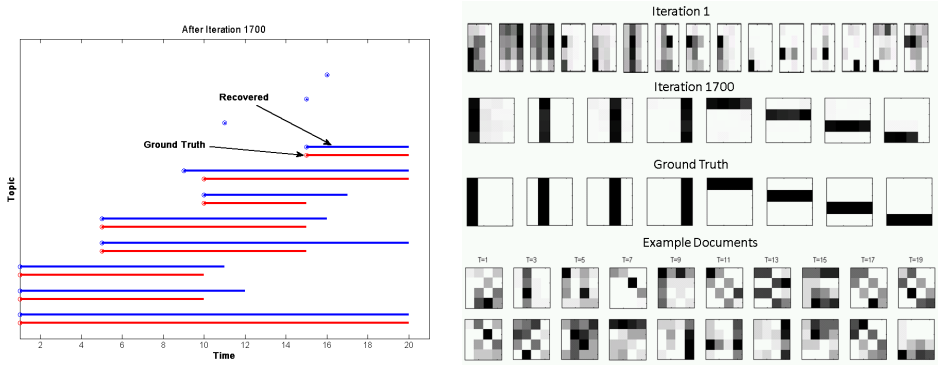

Figure 2: Illustrating Simulation results. **Left**: Topic's death-birth over time (topics numbered from bottom-top). Ground truth is shown in red and recovered in blue. Each topic puts its math uniformly on 4-words : either a row or a column shown as black squares in the figure. **Right**: from top to bottom, topics' distribution after iteration 1, a posterior sample, ground truth (numbered from left to right), and finally a set of documents at different time epochs from the training data. Each document is represented as a unigram distribution over the 16-word vocabulary where the lighter the color assigned to a word, the lower its probability under this document

3 Results

Here we show simple demonstrations of the model when applied to simulated and real data. First, Figure 2 shows how the model was able to recover the death and birth of topics in a simulated corpus. Second, Figure 3 shows some of the topics, with their life span, discovered from the NIPS12 data collection. The result obtained using the model can be used to recover the timeline of topics in the corpus: when each topic was born, as well as the timeline of each topis: what are the key papers in each topic at each year. Moreover, we can also identify seminal papers as the papers that spawn new topics or cause dramatic change in the language model of each topic which can be measured as the KL divergence between the multinomial distribution of the topic at two successive time epochs. Due to lack of space, we defer full analysis of the result to be presented at the workshop and in the final version of the paper.

References

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

[2] D. Blei and J. Lafferty. A correlated topic model of science. In *Annals of Applied Statistics*, volume 1, pages 17–35, 2007.

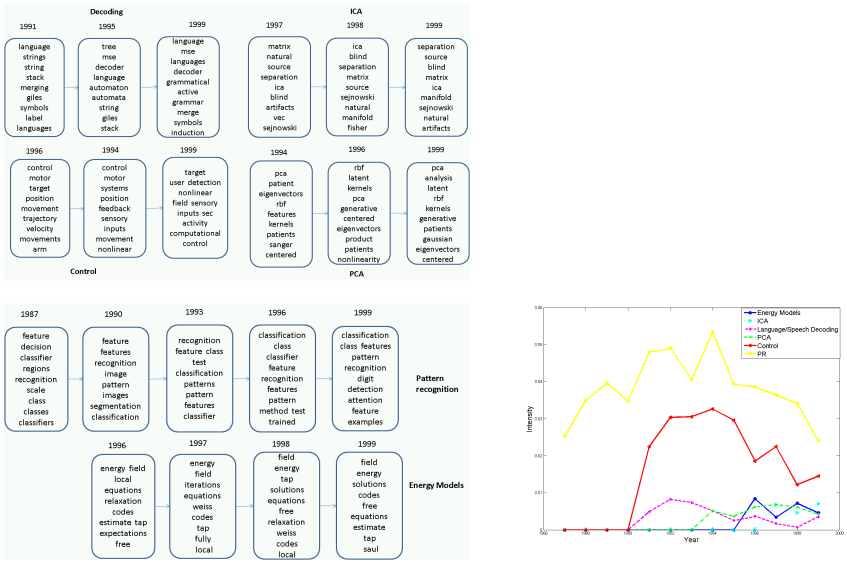

Figure 3: Illustrating Results over the NIPS dataset. **Left**: The word distribution of some topics over their lifespan. **Right**: The evolution of topic's popularity (intensities)

[3] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. 2006.

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

[5] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *KDD*, 2006.

[6] A. Ahmed and E.P. Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process with application to evolutionary clustering. In *SDM*, 2008.

[7] F. Caron, M. Davy, and A. Doucet. Generalized polya urn for time-varying dirichlet processes. In *UAI*, 2007.

[8] N. Srebro and S. Roweis. Time-varying topic models using dependent dirichlet processes. In *Technical report, Department of Computer Science, University of Toronto*, 2005.

[9] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. In *Journal of the American Statistical Association*, volume 101, pages 1566–1581, 2006.