Scalable Dynamic Nonparametric Bayesian Models of Content and Users *

Amr Ahmed[1] **Eric Xing**[2]
[1]Research @ Google , [2]Carnegie Mellon University
amra@google.com, epxing@cs.cmu.edu

Abstract

Online content have become an important medium to disseminate information and express opinions. With their proliferation, users are faced with the problem of missing the big picture in a sea of irrelevant and/or diverse content. In this paper, we addresses the problem of information organization of online document collections, and provide algorithms that create a structured representation of the otherwise unstructured content. We leverage the expressiveness of latent probabilistic models (e.g., topic models) and non-parametric Bayes techniques (e.g., Dirichlet processes), and give online and distributed inference algorithms that scale to terabyte datasets and adapt the inferred representation with the arrival of new documents. This paper is an extended abstract of the 2012 ACM SIGKDD best doctoral dissertation award of Ahmed [2011].

1 Introduction

Our online infosphere is evolving with an astonishing rate. It is reported that there are 50 million scientific journal articles published thus far [Jinha, 2010], 126 million blogs [1], an average of one news story published per second, and around 500 million tweets per day. With the proliferation of such content, users are faced with the problem of missing the big picture in a sea of irrelevant and/or diverse content. Thus several unsupervised techniques were proposed to build a structured representation of users and content.

Traditionally, clustering is used as a popular unsupervised technique to explore and visualize a document collection. When applied in document modeling, it assumes that each document is generated from a single component (cluster or topic) and that each cluster is a uni-gram distribution over a given vocabulary. This assumption limits the expressive power of the model, and does not allow for modeling documents as a mixture of topics.

Recently, mixed membership models [Erosheva *et al.*, 2004], also known as admixture models, have been proposed to remedy the aforementioned deficiency of mixture models. Statistically, an object $\mathbf{w}_d$ is said to be derived from an *admixture* if it consists of a bag of elements, say $\{w_{d1}, \ldots, w_{dN}\}$, each sampled independently or coupled in some way, from a mixture model, according to an *admixing coefficient vector* θ, which represents the (normalized) fraction of contribution from each of the mixture component to the object being modeled. In a typical text modeling setting, each document corresponds to an object, the words thereof correspond to the elements constituting the object, and the document-specific admixing coefficient vector is often known as a *topic vector* and the model is known as latent Dirichlet allocation (LDA) model due to the choice of a Dirichlet distribution as the prior for the topic vector θ [Blei *et al.*, 2003].

Notwithstanding these developments, existing models can not faithfully model the dynamic nature of online content, represent multiple facets of the same topic and scale to the size of the data on the internet. In this paper, we highlight several techniques to build a structured representation of content and users. First we present a flexible dynamic non-parametric Bayesian process called the Recurrent Chinese Restaurant Process for modeling longitudinal data and then present several applications in modeling scientific publication, social media and tracking of user interests.

2 Recurrent Chinese Restaurant Process

Standard clustering techniques assume that the number of clusters is known a priori or can be determined using cross validation. Alternatively, one can consider non-parametric techniques that adapt the number of clusters as new data arrives. The power of non-parametric techniques is not limited to model selection, but they endow the designer with necessary tools to specify priors over sophisticated (possibly infinite) structures like trees, and provide a principled way of learning these structures form data. A key non-parametric distribution is the Dirichlet process (DP). DP is a distribution over distributions [Ferguson, 1973]. A DP denoted by $DP(G_0, \alpha)$ is parameterized by a base measure G_0 and a concentration parameter α. We write $G \sim DP(G_0, \alpha)$ for a draw of a distribution G from the Dirichlet process. G itself is a distribution over a given parameter space θ, therefore we can draw parameters $\theta_{1:N}$ from G. Integrating out G, the

*The dissertation on which this extended abstract is based was the recipient of the 2012 ACM SIGKDD best doctoral dissertation award, [Ahmed, 2011].

[1]http://www.blogpulse.com/

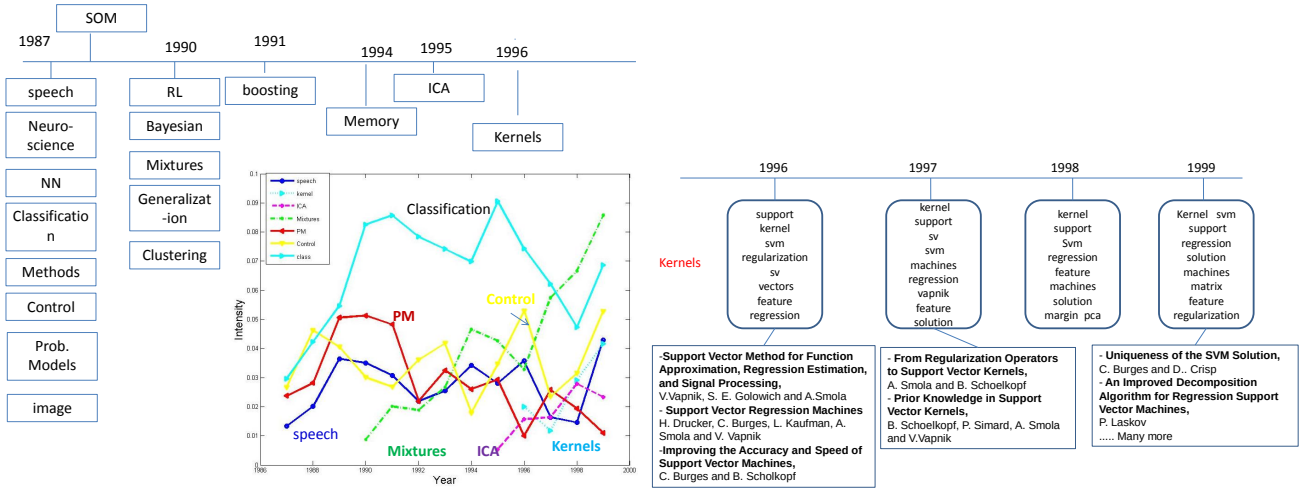

Figure 1: Left: the NIPS conference timeline as discovered by the iDTM. Right the evolution of the Topic Kernel Methods.

parameters $\boldsymbol{\theta}$ follow a Polya urn distribution [Blackwell and MacQueen, 1973], also knows as the Chinese restaurant process (CRP), in which the previously drawn values of θ have strictly positive probability of being redrawn again, thus making the underlying probability measure G discrete with probability one. More formally,

$$\theta_i|\theta_{1:i-1}, G_0, \alpha \sim \sum_k \frac{m_k}{i-1+\alpha}\delta(\phi_k) + \frac{\alpha}{i-1+\alpha} G_0. \quad (1)$$

where $\phi_{1:k}$ denotes the distinct values among the parameters $\boldsymbol{\theta}$, and m_k is the number of parameters θ having value ϕ_k. By using the DP at the top of a hierarchical model, one obtains the Dirichlet process mixture model, DPM [Antoniak, 1974]. The generative process thus proceeds as follows:

$$G|\alpha, G_0 \sim DP(\alpha, G_0), \quad \theta_d|G \sim G, \quad \mathbf{w}_d|\theta_d \sim F(.|\theta_d), \quad (2)$$

where F is a given likelihood function parameterized by θ.

Dirichlet process mixture (or CRP) models provide a flexible Bayesian framework, however the full exchangeability assumption they employ makes them an unappealing choice for modeling longitudinal data such as text streams that can arrive or accumulate as epochs, where data points inside the same epoch can be assumed to be fully exchangeable, whereas across the epochs both the structure (i.e., the number of mixture components) and the parametrization of the data distributions can evolve and therefore unexchangeable. In this section, we present the Recurrent Chinese Restaurant Process (RCRP) [Ahmed and Xing, 2008] as a framework for modeling these complex longitudinal data, in which the number of mixture components at each time point is unbounded; the components themselves can retain, die out or emerge over time; and the actual parametrization of each component can also evolve over time in a Markovian fashion.

In RCRP, documents are assumed to be divided into epochs (e.g., one hour or one day); we assume exchangeability only within each epoch. For a new document at epoch t, a probability mass proportional to α is reserved for generating a new cluster. Each existing cluster may be selected with probability proportional to the sum $m_{kt} + m'_{kt}$, where m_{kt} is the number of documents at epoch t that belong to cluster k, and m'_{kt} is the prior weight for cluster k at time t. If we let c_{td} denotes the cluster assingment of document d at time t, then:

$$c_{td}|\mathbf{c_{1:t-1}}, \mathbf{c_{t,1:d-1}} \sim \text{RCRP}(\alpha, \lambda, \Delta) \quad (3)$$

to indicate the distribution

$$P(c_{td}|\mathbf{c_{1:t-1}}, \mathbf{c_{t,1:d-1}}) \propto \begin{cases} m'_{kt} + m_{kt}^{-td} & \text{existing cluster} \\ \alpha & \text{new cluster} \end{cases} \quad (4)$$

As in the original CRP, the count m_{kt}^{-td} is the number of documents in cluster k at epoch t, not including d. The temporal aspect of the model is introduced via the prior m'_{kt}, which is defined as

$$m'_{kt} = \sum_{\delta=1}^{\Delta} e^{-\frac{\delta}{\lambda}} m_{k,t-\delta}. \quad (5)$$

This prior defines a time-decaying kernel, parametrized by Δ (width) and λ (decay factor). When $\Delta = 0$ the RCRP degenerates to a set of independent Chinese Restaurant Processes at each epoch; when $\Delta = T$ and $\lambda = \infty$ we obtain a global CRP that ignores time. In between, the values of these two parameters affect the expected life span of a given component, such that the lifespan of each storyline follows a power law distribution [Ahmed and Xing, 2008]. In addition, the distribution ϕ_k of each component changes over time in a markovian fashion, i.e.: $\phi_{kt}|\phi_{k,t-1} \sim P(.|\phi_{k,t-1})$. In the following three sections we give various models build on top of RCRP and highlight how inference is performed and scaled to the size of data over the internet.

3 Modeling Scientific Publications

With the large number of research publications available online, it is important to develop automated methods that can discover salient topics (research area), when each topic

started, how each topic developed over time and what are the representative publications in each topic at each year. Mixed-membership models (such as LDA) are static in nature and while several dynamic extensions have been proposed ([Blei and Lafferty, 2006]), non of them can deal with evolving all of the aforementioned aspects. While, the RCRP models can be used for modeling the temporal evolution of research topics, it assumes that each document is generated from a single topic (cluster). To marry these two approaches, we first introduce Hierarchical Dirichlet Processes (HDP [Teh *et al.*, 2006]) and then illustrate our proposed model.

Instead of modeling each document $\mathbf{w}_d$ as a single data point, we could model each document as a DP. In this setting,each word w_{dn} is a data point and thus will be associated with a topic sampled from the random measure G_d, where $G_d \sim DP(\alpha, G_0)$. The random measure G_d thus represents the document-specific mixing vector over a potentially infinite number of topics. To share the same set of topics across documents, we tie the document-specific random measures by modeling the base measure G_0 itself as a random measure sampled from a $DP(\gamma, H)$. The discreteness of the base measure G_0 ensures topic sharing between all the documents.

Now we proceed to introduce our model, iDTM [Ahmed and Xing, 2010b] which allows for infinite number of topics with variable durations. The documents in epoch t are modeled using an epoch specific HDP with high-level base measure denoted as G_0^t. These epoch-specific base measures $\{G_0^t\}$ are tied together using the RCRP of Section 2 to evolve the topics' popularity and distribution over words as time proceeds. To enable the evolution of the topic distribution over words, we model each topic as a logistic normal distribution and evolve its parameters using a Kalman filter. This choice introduces non-conjugacy between the base measure and the likelihood function and we deal with it using a Laplace approximate inference technique proposed in [Ahmed and Xing, 2007].

We applied this model to the collection of papers published in the NIPS conference over 18 years. In Figure 1 we depict the conference timeline and the evolution of the topic 'Kernel Methods' alone with popular papers in each year.

In addition to modeling temporal evolution of topics, in [Ahmed *et al.*, 2009] we developed a mixed-membership model for retrieving relevant research papers based on multiple modalities: for example figures or key entities in the paper such as genes or protein names (as in biomedical papers). Figures in biomedical papers pose various modeling challenges that we omit here for space limitations.

4 Modeling Social Media

News portals and blogs/twitter are the main means to disseminate news stories and express opinions. With the sheer volume of documents and blog entries generated every second, it is hard to stay informed. This section explores methods that create a structured representation of news and opinions.

Storylines emerge from events in the real world, such as the Tsunami in Japan, and have certain durations. Each story can be classified under multiple topics such as disaster, rescue and economics. In addition, each storyline focuses on certain

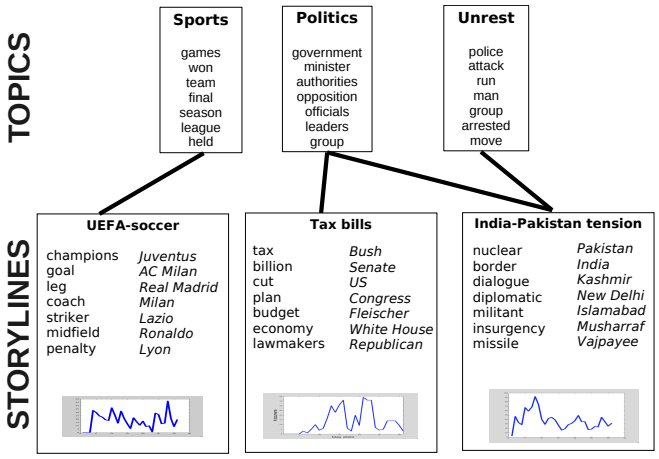

Figure 2: Some example storylines and topics extracted by our system. For each storyline we list the top words in the left column, and the top named entities at the right; the plot at the bottom shows the storyline strength over time. For topics we show the top words. The lines between storylines and topics indicate that at least 10% of terms in a storyline are generated from the linked topic.

words and named entities such as the name of the cities or people involved in the event. In [Ahmed *et al.*, 2011b,a] we used RCRP to model storylines. In a nutshell, we emulate the process of generating news articles. A story is characterized by a mixture of topics and the names of the key entities involved in it. Any article discussing this story then draws its words from the topic mixture associated with the story, the associated named entities, and any story-specific words that are not well explained by the topic mixture. The latter modification allows us to improve our estimates for a given story once it becomes popular. In summary, we model news story clustering by applying a topic model to the clusters, while simultaneously allowing for cluster generation using RCRP.

Such a model has a number of advantages: estimates in topic models increase with the amount of data available. Modeling a story by its mixture of topics ensures that we have a plausible cluster model right from the start, even after observing only one article for a new story. Third, the RCRP ensures a continuous flow of new stories over time. Finally, a distinct named entity model ensures that we capture the characteristic terms rapidly. In order to infer storyline from text stream, we developed a Sequential Monte Carlo (SMC) algorithm that assigns news articles to storylines in real time. Applying our online model to a collection of news articles extracted from a popular news portal, we discovered the structure shown in Figure 2. This structure enables the user to browse the storylines by topics as well as retrieve relevant storylines based on any combination of the storyline attributes. Note that named entities are extracted by a preprocessing step using standard extractors. Quantitatively, we compared the accuracy of our *online* clustering with a strong *offline* algorithm [Vadrevu *et al.*, 2011] with favorable outcome.

Second, we address the problem of ideology-bias detection in user generated content such as microblogs. We follow the notion of ideology as defines by Van Dijk [Dijk, 1998]: "a set of general abstract beliefs commonly shared by a group

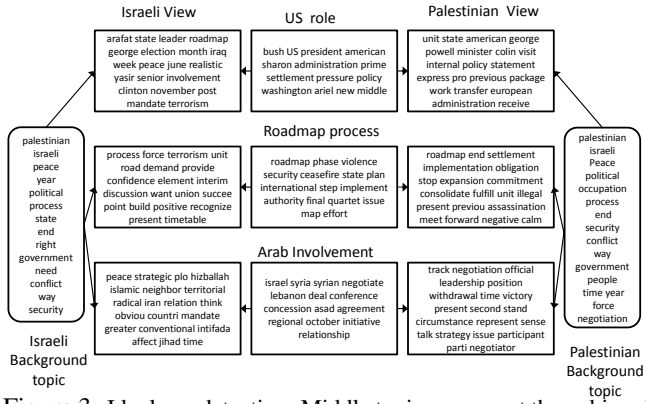

Figure 3: Ideology-detection. Middle topics represent the unbiased portion of each topic, while each side gives the Israeli and Palestinian perspective.

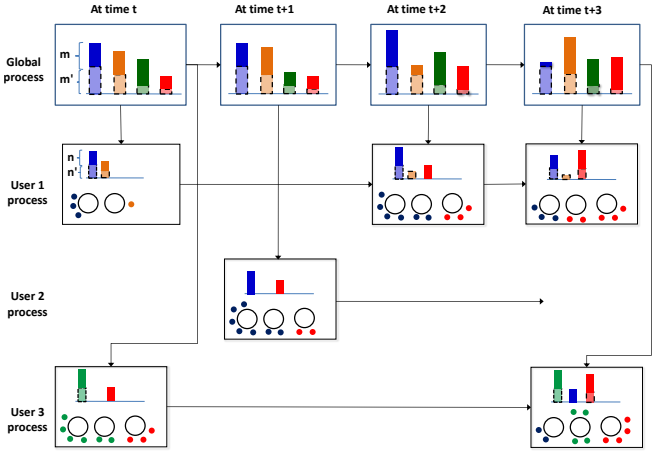

Figure 4: A Fully evolving non-parametric process. Top level process evolves the global topics via an RCRP. Each row represents a user process evolving using an RCR process whose topics depends both on the the global topics at each epoch and the previous state of the user at previous epochs. The user process is sparser than the global process as users need not appear in each epoch, moreover users can appear (and leave) at any epoch.

of people." In other words, an ideology is a set of ideas that directs one's goals, expectations, and actions. For instance, *freedom of choice* is a general aim that directs the actions of "liberals", whereas *conservation of values* is the parallel for "conservatives". In Ahmed and Xing [2010a] we developed a multi-view mixed-membership model that utilizes a factored representation of topics, where each topic is composed of two parts: an unbiased part (shared across ideologies) and a biased part (different for each ideology). Applying this model on a few ideologically labelled documents as seeds and many unlabeled documents, we were able to identify how each ideology stands with respect to mainstream topics. For instance in Figure 3 we show the result of applying the model to a set of articles written on the middle east conflict by both Israeli and Palestinian writers. Given a new documents, the model can 1) detect its idealogical bias (if any), 2) point where the bias appears (i.e. highlight words and/or biased sentences) and 3) retrieve documents written on the same topic from the opposing ideology. Our model achieves state of the art results in task 1 and 3 while being unique in solving task 2.

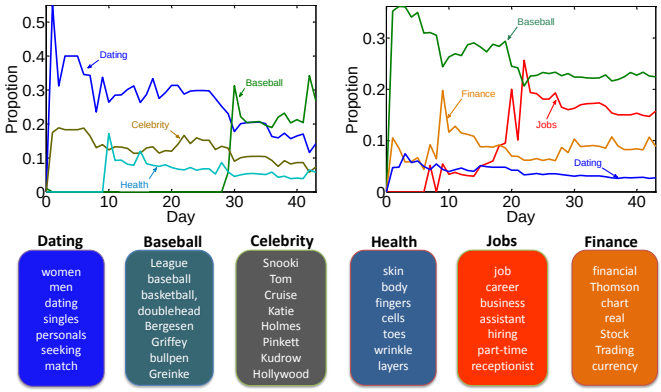

Figure 5: Dynamic interests of two users.

5 Modeling User Interests

Historical user activity is key for building user profiles to predict the user behaviour and affinities in many web applications such as targeting of online advertising, content personalization and social recommendations. User profiles are temporal, and changes in a user's activity patterns are particularly useful for improved prediction and recommendation. For instance, an increased interest in car-related web pages suggests that the user might be shopping for a new vehicle.

In Ahmed *et al.* [2011c] we present a comprehensive statistical framework for user profiling based on the RCRP model which is able to capture such effects in a fully unsupervised fashion. Our method models topical interests of a user dynamically where both the user association with the topics and the topics themselves are allowed to vary over time, thus ensuring that the profiles remain current. For instance if we represent each user as a bag of the words in their search history, we could use the iDTM model described in Section 3. However, unlink research papers that exist in a given epoch, users exist along multiple epoch (where each epoch here might denote a day). To solve this problem we extend iDTM by modeling each user himself as a RCRP that evolves over time as shown in Figure 4. To deal with the size of data on the internet, we developed a streaming, distributed inference algorithm that distribute users over multiple machines and synchronizing the model parameters using an asynchronous consensus protocol described in more details in [Ahmed *et al.*, 2012; Smola and Narayanamurthy, 2010]. Figure 5 shows qualitatively the output of our model over two users. Quantitatively the discovered interests when used as features in an advertising task results in significant improvement over a strong deployed system.

6 Conclusions

Our infosphere is diverse and dynamic. Automated methods that create a structured representation of users and content are key to help users staying informed. We presented a flexible nonparametric Bayesian model called the Recurrent Chinese Restaurant Process and showed how using this formalism (in addition to mixed-membership models) can solve this task. We validated our approach on many domains and showed how to scale the inference to the size of the data on the internet and how to performing inference in online settings.

References

A. Ahmed and E. P. Xing. On tight approximate inference of the logistic normal topic admixture model. In *AISTATS*, 2007.

A. Ahmed and E. P. Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *SDM*, pages 219–230. SIAM, 2008.

A. Ahmed and E. P. Xing. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *EMNLP*, 2010.

A. Ahmed and E. P. Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In *UAI*, 2010.

A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *KDD*, pages 39–48. ACM, 2009.

A. Ahmed, Q. Ho, J. Eisenstein, E. P. Xing, A. J. Smola, , and C. H. Teo. Unified analysis of streaming news. In *in WWW 2011*, 2011.

A. Ahmed, Q. Ho, C. hui, J. Eisenstein, A. Somla, and E. P. Xing. Online inference for the infinte topic-cluster model: Storylines from text stream. In *AISTATS*, 2011.

A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *KDD*, pages 114–122, 2011.

A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A.J. Smola. Scalable inference in latent variable models. In *Web Science and Data Mining (WSDM)*, 2012.

A. Ahmed. *Modeling Users and Content: Structured Probabilistic Representation, and Scalable Online Inference Algorithms*. PhD thesis, School of Computer Science, Carnegie Mellon University, 2011.

C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

D. Blackwell and J. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, volume 148, pages 113–120. ACM, 2006.

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

T. A. Van Dijk. Ideology: A multidisciplinary approach. 1998.

E. Erosheva, S. Fienberg, and J. Lafferty. Mixed membership models of scientific publications. *PNAS*, 101(1), 2004.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

A. E. Jinha. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.

A. J. Smola and S. Narayanamurthy. An architecture for parallel topic models. In *Very Large Databases (VLDB)*, 2010.

Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(576):1566–1581, 2006.

S. Vadrevu, C. H. Teo, S. Rajan, K. Punera, B. Dom, A. J. Smola, Y. Chang, and Z. Zheng. Scalable clustering of news search results. In *in WSDM 2011*, 2011.