Accounting for Data Dependencies within a Hierarchical Dirichlet Process Mixture Model

Dongwoo Kim
KAIST
Department of Computer Science
dw.kim@kaist.ac.kr

Alice Oh
KAIST
Department of Computer Science
alice.oh@kaist.edu

ABSTRACT

We propose a hierarchical nonparametric topic model, based on the hierarchical Dirichlet process (HDP), that accounts for dependencies among the data. The HDP mixture models are useful for discovering an unknown semantic structure (i.e., topics) from a set of unstructured data such as a corpus of documents. For simplicity, HDP makes an exchangeability assumption that any permutation of the data points would result in the same joint probability of the data being generated. This exchangeability assumption poses a problem for some domains where there are clear and strong dependencies among the data. A model that allows for non-exchangeability of data can capture these dependencies and assign higher probabilities to clusters that account for data dependencies, for example, inferring topics that reflect the temporal patterns of the data. Our model incorporates the distance dependent Chinese restaurant process (ddCRP), which clusters data with an inherent bias toward clusters of data points that are near to one another, into a hierarchical construction analogous to the HDP, and we call this new prior the distance dependent Chinese restaurant franchise (ddCRF). When tested with temporal datasets, the ddCRF mixture model shows clear improvements in data fit compared to the HDP in terms of heldout likelihood and complexity. The resulting set of topics shows the sequential emergence and disappearance patterns of topics.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Nonparametric statistics; H.3.3 [**Information Search and Retrieval**]: Clustering, Text Mining

General Terms

Algorithms, Experimentation

Keywords

Latent Topic Modeling, Bayesian Nonparametric Models

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

1. INTRODUCTION

A probabilistic topic model seeks to discover a hidden structure referred to as "topics" from an unannotated set of data. The assumptions of the LDA (latent Dirichlet allocation) family, a popular topic model initially developed for modeling text, is that a topic is represented by a multinomial over the vocabulary, each document is generated from a set of topics, and each word token in the document is assigned to a specific topic. This topic modeling problem is an example of the grouped clustering problem where the data are composed of a set of groups, each data point within a group is assigned to a latent cluster, and these latent clusters are shared across the groups. When applied to text, each document corresponds to a group, each word token within the document corresponds to each data point, a topic corresponds to a latent cluster, and these topics are shared across all the documents in the corpus. The nonparametric extension of the LDA, called the HDP-LDA for Hierarchical Dirichlet process [12], is widely used for the grouped clustering problem, and numerous variants of the HDP have been applied to text modeling [17], sound source modeling [5], activity recognition [6], and computational biology [11]. The nonparametric nature of the HDP means the model can be fitted without specifying the number of clusters a priori because the model infers an appropriate number of clusters.

The HDP assumes that data are exchangeable, and when applied to text, this means that the word tokens, as well as the documents, can be permuted in any way without affecting the outcome of the model. For some applications, the exchangeability assumption is appropriate, but in some other applications, it is an overly simplifying assumption. In discovering topics from text, the exchangeability assumption would result in topic assignments without considering the ordering of the documents in the corpus. This poses a problem for many corpora such as the corpus of conference proceedings where there are clear temporal patterns of documents and topics. For example, when analyzing conference proceedings, "collaborative filtering" came to be a major research topic around the year 1999, and so the documents before the year 1999 are unlikely to be generated from that topic. However, a model that assumes exchangeability of data does not distinguish the temporal index of the documents and would allow older documents to be assigned to the new topic.

We propose a new model, the distance dependent Chinese restaurant franchise (ddCRF), which takes the concept of the distance dependent Chinese restaurant process (ddCRP) [1] into a hierarchical construction of the HDP. In the

ddCRP, the random assignments of the data points to the clusters depend on the distances between the data points, and this allows the ddCRP to result in a better fit than the Chinese restaurant process (CRP), a metaphor of the Dirichlet process (DP) prior with the usual exchangeability assumption. By incorporating this idea into the HDP, we are able to construct a new model that performs better on the topic modeling task than the HDP on a sequential dataset. We show this by testing the ddCRF on the task of topic modeling using four datasets of conference proceedings. The ddCRF outperforms the HDP on measures of heldout likelihood and complexity, and the ddCRF is also able to discover interesting temporal patterns of topics.

This paper is organized in the following structure. Section 2 describes how we build in data dependence into the CRF to make our model, the ddCRF. Section 3 demonstrates the posterior inference procedure for the ddCRF. Section 4 presents the modeling performance of ddCRF by inferring the topics of four different time-varying corpora of conference proceedings. Finally, Section 5 summarizes this work and discusses directions for future work.

2. DISTANCE DEPENDENT CHINESE RESTAURANT FRANCHISE

We propose a new model, the distance-dependent Chinese restaurant franchise (ddCRF) by adopting the concept of the distance dependent Chinese restaurant process (ddCRP) into the CRF, to account for dependencies among data. We preserve the notation of ddCRF as described in the original CRF paper [12]. In this section, we show how the ddCRP can be integrated into the CRF.

2.1 Table-Based Distance Dependent CRP

Blei and Frazier [1] introduced the customer-based distance dependent Chinese restaurant process (ddCRP) where, instead of customers being assigned to tables, they are assigned to other customers or not assigned to anyone. The probability of a new customer being assigned to other customers is proportional to the distances between the new customer and the other customers. Explicit table assignments do not occur in the customer-based CRP, but the connected components of customers implicitly exhibit a clustering property.

This customer-based ddCRP can be reverted to a table-based ddCRP by summing over the distances to each of the customers within the same connected component. Let K be an imaginary number of tables, which would be the same as the number of connected components of customers, and z_i denote the index of the imaginary table of the ith customer. Let D denote the set of all distance measurement between customers, d_{ij} denote the distance between customer i and j, and f denote the decay function which takes a distance as its parameter. The probability of each table for the ith customer is specified as follows:

$$\begin{aligned} p(z_i = k \mid D, z_{1:(i-1)}, \gamma) &= \frac{\sum_{z_j = k} f(d_{ij})}{\gamma + \sum_{j \neq i} f(d_{ij})} \\ p(z_i = K+1 \mid D, z_{1:(i-1)}, \gamma) &= \frac{\gamma}{\gamma + \sum_{j \neq i} f(d_{ij})}. \end{aligned}$$

In general, the decay function mediates how the distances among customers affect the resulting distribution over partitions. We consider two decay functions: the *exponential decay* $f(d_{ij}) = e^{-d_{ij}/a}$, and the *logistic decay* $f(d_{ij}) = exp(-d_{ij}+a)/(1+exp(-d_{ij}+a))$, where a is a decay parameter.

By setting the right combination of the type of decay function and the distance measure, we obtain the special case of sequential CRPs. When we define the distance measure such that $d_{ij} = \infty$ for those $j > i$, using either the logistic or the exponential decay function brings $f(\infty) = 0$, and this results in a sequential CRP.

The partition probability over customers can be computed in the customer-based ddCRP simply as defined in [1]. In the table-based ddCRP, however, to compute the partition probability over customers, we must consider all combinations of table assignments, and the number of combinations increases factorially as the number of customers increases. If we make the assumption of sequential non-exchangeability of data such that the model would be the sequential ddCRP, the partition probability can be computed by

$$p(z_{1:N} \mid D, f, \gamma) = \prod_{i=1}^{N} \left(\frac{\mathbf{1}[z_i = K_{i-1}+1]\gamma}{\sum_{j<i} f(d_{ij}) + \gamma} + \frac{\mathbf{1}[z_i \neq K_{i-1}+1] \sum_{z_j = z_i, j<i} f(d_{ij})}{\sum_{j<i} f(d_{ij}) + \gamma} \right), \quad (1)$$

where K_i is the number of allocated tables until the ith customer sits at a table.

Although this commitment to a table-based ddCRP would mean we cannot take advantage of the sampling efficiency of the customer-based ddCRP [1], we propose the table-based distance dependent CRP and use it for the rest of this paper. We do so because in the hierarchical model presented in the next section, it is relatively easy to implement and compute the conditional posterior probabilities.

2.2 CRF to ddCRF

The Chinese restaurant franchise is designed as an approach to the problem of model-based clustering of grouped data. The CRF assumes that the data are exchangeable, but this assumption does not take into account inherent dependencies among data points in some corpora. In order to capture such dependencies, we can incorporate the key idea of ddCRP, which takes into account the non-exchangeability of data, into the CRF. There are three ways to do that.

1. We can model the first (menu) level CRP as a ddCRP. The intuition behind this approach is that the selection of a menu could be influenced by nearby restaurants. If there is a famous menu in a specific region, menus in the other restaurants may be affected by that menu. In this case, however, customers in the restaurants do not exhibit dependences with other customers in the same restaurant.

2. We can model the second (customer) level CRP as a ddCRP. The intuition behind this approach is that the selection of a table by a customer could be affected by other customers already sitting at the tables, but the selection of menu at each table is not affected by menus in the other restaurants.

3. We can model both first and second levels of CRP as ddCRP. By replacing both levels of CRP with ddCRP, we can combine both of the approaches above.

In the rest of this paper we only consider the first approach. Although we only implement and test the first approach, the other two approaches can be implemented in a straightforward way. For the task of topic modeling of a text corpus, replacing the first level CRP by ddCRP indicates that the topic allocation within a document can be influenced by the topics of other documents that are close to the document. Replacing the second level CRP by ddCRP is also interesting, especially for an application such as object recognition in images where the locations and spatial distances of pixels within one image should be considered [16].

The conditional distribution of the ddCRF follows directly from the conditional distribution of the CRF, only we need to consider the decay function and the distances between data points for the ddCRF. Let ϕ_k is an atom drawn from based distribution H, ψ_{jt} is drawn from first level CRP which is a menu of tth table at jth restaurant, and θ_{ij} is drawn from second level CRP which is a ith customer of jth restaurant. In the case where the first level of CRP is replaced by ddCRP, the conditional distribution of second level θ_{ji} only depends on the other Θ. However, the conditional distribution of the first level ψ_{jt} must be computed by considering the distances with other Ψ, hence we have:

$$\begin{aligned}\psi_{jt} \mid \psi_{11}, \psi_{12}, ..., \psi_{21}, ..., \psi_{j,t-1}, \gamma, H \\ \sim \sum_{k=1}^{K} \frac{\sum_{j't' \neq jt, k_{j't'}=k} f(d_{j't',jt})}{\sum_{j't' \neq jt} f(d_{j't',jt}) + \gamma} \delta(\phi_k) \\ + \frac{\gamma}{\sum_{j't' \neq jt} f(d_{j't',jt}) + \gamma} H.\end{aligned}$$

The distance $d_{j't',jt}$ between tables ψ_{jt} and $\psi_{j't'}$ must be carefully defined because it mediates the conditional distribution of ψ_{jt}. It is possible to treat each table in each restaurant to have its own location, or to treat all tables in the same restaurant to share the same location so the distance would be zero between tables at the same restaurant, i.e., $d_{jt',jt} = 0$.

3. POSTERIOR INFERENCE

In this section, we describe a Gibbs sampling method for the ddCRF mixture model. Several posterior approximation techniques including MCMC [12] and variational inference techniques [13] are introduced for the CRF mixture and its related Bayesian non-parametric mixture models.

3.1 Posterior Sampling in the ddCRF Mixture

Let us recall the variables of interest. x_{ji} is the ith data point in the jth group, t_{ji} is an indicator variable for the table index of x_{ji}, k_{jt} is an indicator variable of the menu index of the tth table at jth restaurant, and n_{jtk} is the number of data points at tth table in jth restaurant with dish k. We use dot to represent a marginal sum, and a superscript to denote the counts or indicators of variable excluding the one specified by the superscript.

The sampling process is used in order to infer about $\boldsymbol{t}$ and $\boldsymbol{k}$ based on the observed data set $\boldsymbol{x}$. Before computing the posterior probabilities of these variables, we need to compute the probability of data point x_{ji} given all other variables. We let H denote the prior distribution over probability of data points, and ϕ_k is drawn from this distribution with probability $p(\phi_k|\eta)$ with hyperparameter η, then each data point belonging to this latent cluster has a probability $p(\cdot|\phi_k)$. Therefore, the conditional density of data x_{ji} given all other variables under mixture component k is computed as:

$$\begin{aligned}p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}, \boldsymbol{k}) &= \frac{p(x_{ji}, \boldsymbol{x}^{-ji}|\boldsymbol{t}, \boldsymbol{k})}{p(\boldsymbol{x}^{-ji}|\boldsymbol{t}, \boldsymbol{k})} \\ &= \frac{\int p(x_{ji}|\phi_k) \prod_{j'i' \neq ji, z_{j'i'}=k} p(x_{j'i'}|\phi_k) p(\phi_k|\eta) d\phi_k}{\int \prod_{j'i' \neq ji, z_{j'i'}=k} p(x_{j'i'}|\phi_k) p(\phi_k|\eta) d\phi_k}.\end{aligned}$$

This equation can be further simplified if we use a conjugate prior H. If we use a dirichlet multinomial conjugacy on H, this equation can be further simplified as follows:

$$p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}, \boldsymbol{k}) = \frac{\sum_{j'i' \neq ji} 1[x_{j'i'} = x_{ji}] + \eta}{\sum_{j'i' \neq ji} 1[k_{j't_{j'i'}} = k_{jt_{ji}}] + K \cdot \eta}.$$

Sampling $\boldsymbol{t}$ Now we show the conditional probability of t_{ji} given other variables by bringing the Chinese restaurant franchise metaphor. The probability of t_{ji} is proportional to the number of customers sitting at the table t times the probability of data point x_{ji} arising from the table.

$$\begin{aligned}&p(t_{ji} = t|\boldsymbol{t}^{-ji}, \boldsymbol{k}, \boldsymbol{x}) \\ &\propto \begin{cases} n_{jt}^{-ji} \cdot p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}^{-ji}, t_{ji} = t, \boldsymbol{k}) \\ \qquad (t \text{ is used before}) \\ \alpha \cdot p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}) \\ \qquad (\text{new } t), \end{cases}\end{aligned} \tag{2}$$

where the probability of the data point x_{ji} drawn from a new table can be calculated by marginalizing over the latent cluster k,

$$\begin{aligned}&p(x_{ji} \mid \boldsymbol{x}^{-ji}, \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}) \\ &= \sum_{k=1}^{K} \frac{\sum_{j't' \neq jt, k_{j't'}=k} f(d_{j't',jt})}{\sum_{j't' \neq jt} f(d_{j't',jt}) + \gamma} \\ &\times p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}, k_{jt^{\text{new}}} = k) \\ &+ \frac{\gamma}{\sum_{j't' \neq jt} f(d_{j't',jt}) + \gamma} \\ &\times p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}, k_{jt^{\text{new}}} = k^{\text{new}}).\end{aligned} \tag{3}$$

Sampling $\boldsymbol{k}$ If t_{ji} is drawn from the second term of Equation 2 then we have to draw $k_{jt^{new}}$ for the new table from the following distribution.

$$\begin{aligned}&p(k_{jt^{\text{new}}} = k|\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{k}^{-jt}) \\ &\propto \begin{cases} \sum_{j't' \neq jt, k_{j't'}=k} f(d_{j't',jt}) \\ \times p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}, k_{jt^{\text{new}}} = k) \\ \qquad (k \text{ is used before}) \\ \gamma \cdot p(x_{ji}|\boldsymbol{x}^{-ji}, \boldsymbol{t}^{-ji}, t_{ji} = t^{\text{new}}, \boldsymbol{k}, k_{jt^{\text{new}}} = k^{\text{new}}) \\ \qquad (\text{new } k). \end{cases}\end{aligned}$$

We use the same computation as in Equation 3 (omitting the common denominator), but we present this again to clarify the sampling procedure. We can also sample k_{jt} by excluding all data items in table jt, and this sampling of new k_{jt} changes the component membership of all data items in table jt.

3.2 Sampling Hyperparameters

To improve our model, we place a prior for first and second level concentration parameters γ and α, as used in the original CRF. Sampling α is done in the same way as [12], so we

Table 1: Dataset statistics of four conference proceedings.

Conference	Period	# of articles	# of unique terms	# of tokens
NIPS	1988-1999	1,740	6,455	450K
SIGIR	1978-2010	1,838	1,988	106K
SIGMOD	1978-2010	2,311	2,745	165K
SIGGRAPH	1974-1991	783	2,381	53K

discuss here how to sample γ. We note that the probability of γ is conditionally independent of customer assignments $\boldsymbol{t}$ given dish assignments $\boldsymbol{k}$. From this fact, the probability of γ is

$$p(\gamma|\boldsymbol{k}) \propto p(\boldsymbol{k}|\gamma)p(\gamma),$$

where $p(\gamma)$ is the prior on the concentration parameter γ. As we derived in Equation 1, $p(\boldsymbol{k}|\gamma)$ can be computed in a sequential setup. Therefore, γ can be sampled from its posterior distribution.

$$p(\boldsymbol{k}|\gamma)$$
$$= \prod_{i}^{N} \frac{\mathbf{1}[k_i = K_{i-1}+1]\gamma + \mathbf{1}[k_i \neq K_{i-1}+1]\sum_{k_j=k_i, j<i} f(d_{ij})}{\gamma + \sum_{j<i} f(d_{ij})}$$
$$\propto \gamma^K \left[\prod_{i}^{N}\left(\alpha + \sum_{j<i} f(d_{ij})\right)\right]^{-1}$$

where K_i is the number of allocated tables until the ith customer in the sequential setup. To sample from the continuous variable we use Griddy-Gibbs method in [8]. This method evaluates the probabilities on a finite set of points, approximates the inverse cdf $p(\gamma|\boldsymbol{k})$ using these points, and samples from the approximated inverse cdf.

4. EMPIRICAL STUDY

We now describe the experiments to evaluate the performance of the ddCRF on four different text datasets and show how the ddCRF compares against the HDP. The datasets for the experiments include conferences, SIGIR, SIGMOD, SIGGRAPH abstracts, collected through the ACM digital library,[1] and the NIPS article dataset[2]. These four conferences have long histories, their proceedings are published over 20 years, and like many academic publications, their main topics have shifted through time. The detailed statistics of the four datasets we used in the experiments are in Table 1.

4.1 Experimental Setup

We modeled these datasets using the ddCRF to capture the topic changes within a conference though time. We removed stop words, terms that occurred less than 10 times in NIPS, SIGIR and SIGMOD, and terms that occurred less than 5 times in SIGGRAPH.

We trained the HDP mixture model, the ddCRF mixture model, and the latent Dirichlet allocation(LDA) [2] with these datasets and compared their results. Each of the results are averaged over 20 runs. All models used for the evaluation used a symmetric Dirichlet distribution with parameter of 0.5 for the prior H over topic distribution. It is possible to sample the Dirichlet parameter over H, but in that case, the number of topics increases too much [15], and it is not efficient in practice, so we just leave it as a constant. The concentration parameters α and γ were given vague gamma priors with the scale parameter of 1 and the shape parameter of 1.

[1]http://portal.acm.org/

[2]http://www.cs.utoronto.ca/∼sroweis/nips

4.2 Comparison to HDP

We compare the ddCRF with the HDP using two metrics, heldout likelihood and complexity. We also compare the discovered topics from the two models for a qualitative analysis.

Heldout Likelihood: Heldout likelihood is widely used in the topic modeling community to compare how well the trained model explains the heldout data (cf. [9, 14, 12]). A better model will give rise to a higher likelihood of heldout documents on average.

$$\text{Heldout likelihood} = log\ p(W|M_{\text{train}}),$$

where M_{train} denotes the model already trained by a training data, and W denotes the heldout data. To calculate the heldout likelihood we used the last 10% of documents for testing and 90% of documents for training.

Figure 1 shows the heldout likelihood of the datasets. For all datasets the ddCRF shows better heldout likelihood than the HDP regardless of the decay parameter and the decay function.[4] The results exhibit that the heldout likelihood gets lower when the decay parameter increases on average. Therefore our model inferences more time sensitive topics with lower decay parameters.

Complexity: Bayesian nonparametrics and the related methods are often used as an alternative of the model selection and integrate over all complexities of a model. If there are two or more models that produce similar results in terms of heldout likelihood, the less complex model is preferred. To measure the complexity of models, we compute a complexity of each model, as defined in [15]. From the posterior topic assignment of the Gibbs sample, we compute the complexity as follows:

$$\text{Complexity} = K + \sum_{k}\sum_{d}\mathbf{1}[(\sum_{n}\mathbf{1}[z_{d,n}=k]) > 0],$$

where K is the number of allocated topics. This measure considers how many topics are used to explain each document and sum it through the entire corpus. A lower complexity indicates that the model uses fewer topics to represent the corpus, and a higher complexity indicates that the model decomposes the data into many dimensions.

Figure 2 shows the model complexities for the four different datasets. The average complexities of the ddCRF are better than the HDP in all cases except the SIGMOD dataset.

[4]We also compared the results with LDA, and it showed that ddCRF outperforms LDA with the same number of topics.

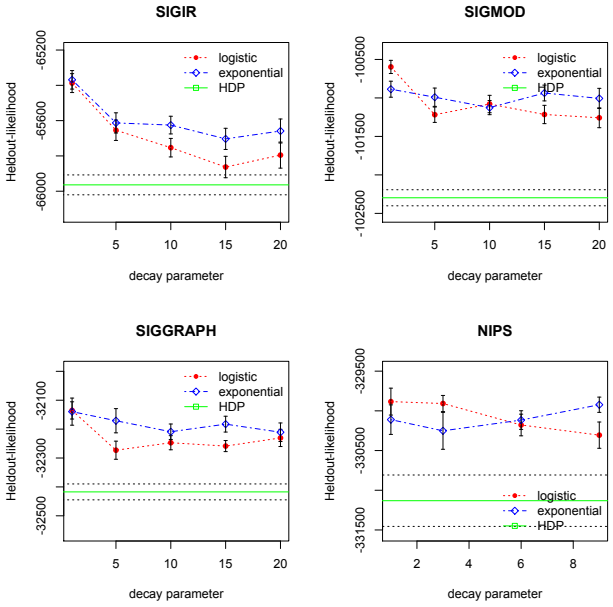

Figure 1: Heldout-Likelihood. Higher is better. The ddCRF outperforms the HDP regardless of the decay function for all datasets.

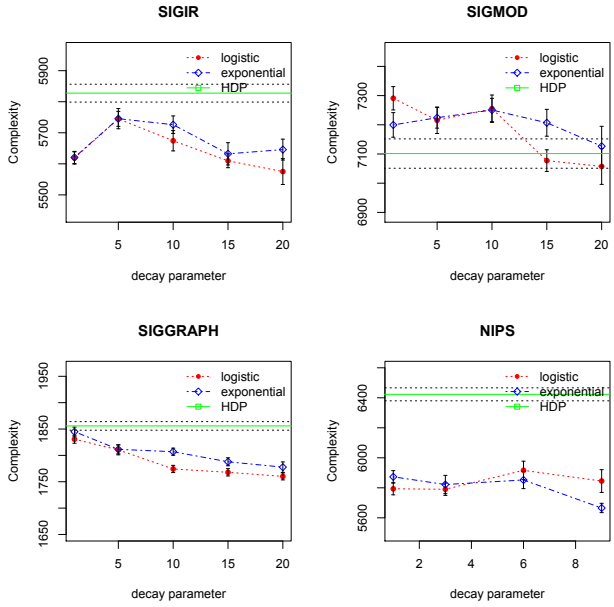

Figure 2: Model complexity where lower values indicate that better (simpler) models were learned. The ddCRF exhibits lower complexity than the HDP for all datasets except for the SIGMOD dataset.

Emergence and disappearance of topics: One strength of our model is that the model imposes a sequential assumption to the first level CRP, and it disallows a word to be assigned to a topic that first appears at a later point in the dataset. Therefore the posterior topic assignment explicitly shows when the topic first appears.

The emergence of topics, combined with the topic trends over time, shows the strength of the ddCRP to model the sequential non-exchangeability of data. To measure the topic trends over time, we define topic *intensity* computed from the posterior sample assignment. At each time slice t, the intensity of topic k is computed by the number of terms assigned to a topic k over the total number of terms at time slice t.

We choose two topics from the SIGIR dataset based on the training results of the ddCRF and look for the most similar topics from the results of the HDP, similarity measured in terms of JS divergence. Figure 3 shows the intensity of those four topics over time. The figures on the left show the topics and their intensities as found by the HDP, and the figures on the right show the topics and their intensities as found by the ddCRF. The topics on the top are about "collaborative filtering", and the topics on the bottom are about "spam filtering". The "spam filtering" topic, which emerged with the rapid growth of email spams, was found by the ddCRF around the year 2000, but the similar topic found by the HDP seems to be a mix of a topic related to spam filtering and another topic related to news articles. The spam filtering topic in fact became a major topic in the SIGIR conference in 2000, and the HDP did not capture this phenomenon well. Similarly, for the "collaborative filtering" topic which became a major research topic in the year 2000, the ddCRF correctly identified the emergence of the topic, whereas the HDP was unable to do so.

We also found interesting patterns of topic emergences from the SIGMOD corpus as shown in Figure 4. The ddCRF identified the "web search" topic emerging around 1994 and the "xml" related topic emerging around 2000. When we modeled the same dataset using the HDP, however, these two topics seem to mix together into one topic.

While the emergence of new topics is simple to identify, the disappearance of topics cannot be explicitly captured from the posterior assignment. With certain decay function and parameter, however, we can deduce when the topic disappeared. An example of a topic that disappears is "File structure and record" in the SIGIR corpus, which last appeared in the year 1988.

5. SUMMARY AND FUTURE WORK

We introduced the distance dependent Chinese restaurant franchise, a hierarchical Bayesian nonparametric model that accounts for dependencies among data. By generalizing the widely used HDP to assume non-exchangeability of data, our model captures temporal patterns in sequential data much better than the existing topic models, HDP and LDA.

One property of the ddCRF that distingishes from the previos attempts is that it can accomodate spatial dependencies in addition to the temporal dependencies. We only considered temporal dependencies here, but modeling spatial dependencies will be the most promising future application of our model. Further, applications of the probabilistic topic model are not restricted to analyzing a document corpus. One of the advantages of the probabilistic topic modeling framework is the flexibility to extend the basic LDA and HDP models. These variants are widely applied to diverse fields[10, 4, 3, 7]. Our model can also be applied to various types of data to uncover other meaningful structure from

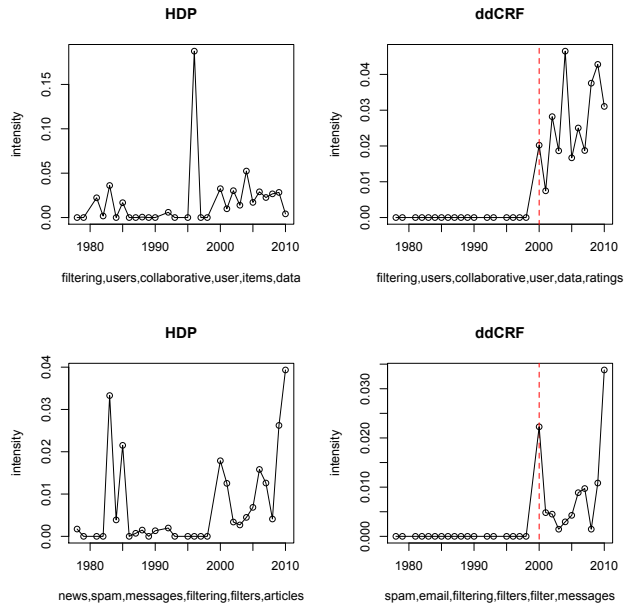

Figure 3: Topic proportions over time identified by the HDP and the ddCRF from the SIGIR corpus. The words beneath the x-axis are the top probability words in each of the topics. As the dotted (red) lines in the ddCRF figures show, the ddCRF clearly captures when the topics first emerged.

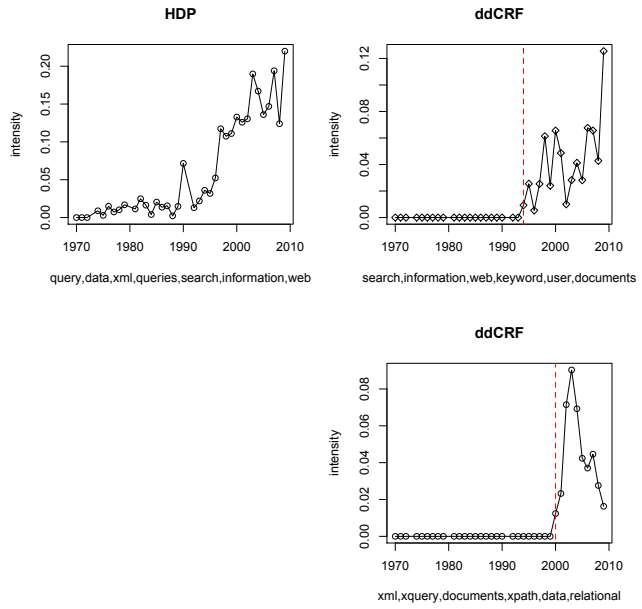

Figure 4: Topic proportions over time identified from the SIGMOD. As the dotted (red) lines in the figures show, the ddCRF captures the emergences of topics about "web search" in 1994 and "xml" in 2000.

data. More specifically, we plan to explore spatial dependencies in modeling of images and spatiotemporal dependencies in geo-tagged data (e.g, tweets). Further, we can explore other approximation inference methods such as a variational method.

6. ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Tehcnology (2010-0025706).

7. REFERENCES

[1] D. Blei and P. Frazier. Distance dependent chinese restaurant processes. *ICML*, 2010.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, pages 993–1022, Jan 2003.

[3] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. *ICML*, 2007.

[4] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. *NIPS*, 2005.

[5] M. Hoffman, D. Blei, and P. Cook. Finding latent sources in recorded music with a shift-invariant hdp. *DAFx*, 2009.

[6] D. Hu, X. Zhang, J. Yin, V. Zheng, and Q. Yang. Abnormal activity recognition based on hdp-hmm models. *International Joint Conferences on Artificial Intelligence*, 2009.

[7] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. *CIKM*, 2009.

[8] C. Ritter and M. Tanner. Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler. *Journal of the American Statistical Association*, 87(419):pp. 861–868, 1992.

[9] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *UAI*, 2004.

[10] R. Socher, S. Gershman, A. Perotte, and P. Sederberg. A bayesian analysis of dynamics in free recall. *NIPS*, 2009.

[11] K. Sohn and E. Xing. A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *Annals of Applied Statistics*, 3(2):791–821, 2009.

[12] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Jan 2006.

[13] Y. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. *NIPS*, 20, 2008.

[14] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why Priors Matter. *NIPS*, 2009.

[15] C. Wang and D. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. *NIPS*, 2010.

[16] X. Wang and E. Grimson. Spatial latent dirichlet allocation. *NIPS*, 2007.

[17] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. *KDD*, 2010.