석사 학위논문
Master's Thesis

거리 의존관계를 이용한 비모수적 베이지안 확률 모형

Distance Dependent Chinese Restaurant Franchise

김 동 우 (金 東 祐 Kim, Dongwoo)
전산학과
Department of Computer Science

KAIST

2011

거리 의존관계를 이용한 비모수적 베이지안 확률 모형

Distance Dependent Chinese Restaurant Franchise

Distance Dependent Chinese Restaurant Franchise

Advisor : Professor Oh, Alice

by

Kim, Dongwoo

Department of Computer Science

KAIST

A thesis submitted to the faculty of KAIST in partial fulfillment of the requirements for the degree of Master of Science in Engineering in the Department of Computer Science . The study was conducted in accordance with Code of Research Ethics[1].

2010. 12. 21.

Approved by

Professor Oh, Alice

[Advisor]

---

[1]Declaration of Ethical Conduct in Research: I, as a graduate student of KAIST, hereby declare that I have not committed any acts that may damage the credibility of my research. These include, but are not limited to: falsification, thesis written by someone else, distortion of research findings or plagiarism. I affirm that my thesis contains honest conclusions based on my own careful research under the guidance of my thesis advisor.

거리 의존관계를 이용한 비모수적 베이지안 확률 모형

김 동 우

위 논문은 한국과학기술원 석사학위논문으로
학위논문심사위원회에서 심사 통과하였음.

2010년 12월 21일

심사위원장 오 혜 연 (인)

심사위원 문 수 복 (인)

심사위원 정 교 민 (인)

MCS
20093057

김 동 우. Kim, Dongwoo. Distance Dependent Chinese Restaurant Franchise. 거리 의존관계를 이용한 비모수적 베이지안 확률 모형. Department of Computer Science . 2011. 25p. Advisor Prof. Oh, Alice. Text in English.

ABSTRACT

Topic models provide a simple way to analyze large volumes of unlabeled documents by automatically identifying the latent semantics of the corpus. Such models have been widely applied to text modeling, cognitive science, computational biology, and many others where there are meaningful patterns hidden in the data. Driven by an ever increasing amount of information available and also by efforts of researchers who have built many tools for topic modeling, there is a wide and fast spread of variants and applications of topic models. This thesis proposes a new model, the distance dependent Chinese restaurant franchise (ddCRF), in which the model considers the distance between the latent variables. This thesis starts with the Chinese restaurant process (CRP), which is a non-parametric prior for Bayesian models, extends it to the Chinese restaurant franchise (CRF), which is a hierarchical non-parametric prior for Bayesian topic models, and finally incorporates the distance dependent Chinese restaurant process (ddCRP) into the CRF to build the ddCRF. For posterior inference in ddCRF, which is an important computational issue in probabilistic generative topic models, this thesis proposes Markov chain Monte Carlo (MCMC) algorithms. The resulting model reflects the intuition that topics in nearby documents are more likely to be similar, and when it is applied to a corpus collected over several years in which the documents exhibit the phenomena of emergence and disappearance of topics through time, the ddCRF produces much clearer patterns than previously proposed models for capturing such temporal patterns. The improved performance of the ddCRF in modeling such corpora is shown with four different corpora of conference proceedings, SIGIR, SIGMOD, SIGGRAPH and NIPS. The ddCRF performs better than the CRF and the most widely used topic model, latent Dirichlet allocation (LDA), in terms of held-out likelihoood and complexity. Another advantage of the ddCRF over LDA, dynamic topic model, and other parametric models, is that the number of topics, which is an important parameter in LDA, need not be fixed a priori in the ddCRF because it is a non-parametric model that infers the appropriate number of topics for a corpus.

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Hierarchical Dirichlet process(HDP) mixture models are valuable tools for the grouped clustering problems where the data are composed of a set of groups, each data point within a group is assigned to the latent cluster, and these latent clusters are shared across the groups. The HDP and its variants are widely applied to text modeling[26], sound source modeling[10], activity recognition[11], and computational biology[19]. Such models have an advantage that they support an infinite number of clusters, so it is possible to automatically find an appropriate number of clusters.

The HDP is represented by a layered construction of Dirichlet Process(DP). In this construction, the distribution drawn from an upper level DP is a base distribution of a lower level DP. By doing so, the HDP maintains sharing clusters across the groups. The Chinese restaurant Process(CRP)[15] is one metaphor used to explain the DP, and the Chinese restaurant franchise(CRF)[25] is one metaphor used to explain the HDP. The CRP can be described by a sequence of customers sitting down at tables in a restaurant. After a sequence of customers sit down, their configuration describes a distribution over partitions. In a CRP mixture model, each partition represented as a latent cluster and the data points corresponding to this partition are drawn from the same latent cluster. The CRF is originated from CRP by assuming that the restaurants, represented as CRPs, are sharing the same menu list. In each restaurant, the customer assignments are same the as CRP, but additionally each table chooses the dish eaten by customers sitting at the table. The selection of a dish is drawn from a global distribution and the global distribution is represented as assigning dishes by a menu-level CRP. In this setting, we can view that each dish corresponds to a latent cluster, and the shared menu allows clusters to share across the groups.

The CRP and CRF are exchangeable models, which means that the ordering of customers or dishes is not considered, and the same configurations over partitions have the same probability regardless of the order of customers and dishes. In some applications, exchangeability is reasonable enough, but in many it is not. For example, in the topic modeling task, a widely studied grouped clustering problem, for a sequential scientific corpus, a topic "spam filtering" should not be considered to model the documents written before 1990s. However, a model that assumes exchangeability of data cannot prohibit these documents to be modeled by that topic.

This thesis presents a new model, distance dependent CRF(ddCRF), in which the latent dishes are sequentially allocated by the distance dependent CRP(ddCRP). The ddCRP, a variant of CRP, considers the distances between customers in constructing the partitions over customers. Using the ddCRP, in a hierarchical construction, our model preserves and directly incorporate the sequence of groups to model the data. Here the distance can be a spatial dimension, time dimension, or other measures.

This thesis is organized in the following structure. Chapter 2 discusses related research on Topic modeling and Bayesian non-parametric models. Chapter 3 explains the relationship between the Chinese restaurant process and distance-dependent Chinese restaurant process. Chapter 4 explains the Chinese restaurant franchise, and proposes a new model, the distance-dependent Chinese restaurant franchise(ddCRF), for the non-exchangeable grouped clustering problems. Chapter 5 demonstrates the posterior inference procedure for the ddCRF. Section 6 shows the modeling performance of ddCRF by inferring the topics of four different time-varying corpora of conference proceedings. Finally, Section 7

discusses future work and conclusion.

Chapter 2. Related Work

This work can be positioned with respect to two related research areas: probabilistic topic modeling and Bayesian nonparametric mixture modeling.

2.1 Probabilistic Topic Models

Probabilistic topic models [4] such as the widely used latent dirichlet allocation (LDA) [3] discover topics that are highly represented in a corpus, and some extensions to LDA, [24, 2, 6, 23] consider the temporal aspect of the corpus as well. In [24], Wang et al. worked with asynchronous text streams to find common topics from documents with different timestamps. They found highly discriminative topics from asynchronous data and synchronized the documents according to topics, but they did not deal with the dynamics of the topics within the corpus. With dynamic topic models (DTM) [2], Blei and Lafferty analyzed how topics evolve over time in a sequential corpus, and they demonstrated how topics in the journal *Science* changed from 1881 to 1999. One limitation with DTM is that it only models the changes of word distributions within the topics and assumes the set of topics stays constant throughout the corpus, so it does not model how topics appear and disappear over time. The same limitation exists for the topic trend detection in [6]. Another limitation with DTM is it they captures the evolution of topics with discrete time. Continuous-DTM [22] extends DTM to capture the evolution of topics over a continuous time corpus. Our model also captures topics over continuous time data. Topics over Time (TOT) [23] by Wang and McCallum jointly models topics and timestamps to analyze when in the sequential data the topics occur, but the topics stay the same over time. Our model does not set a fixed number of topics because it uses a variant of Bayesian nonparametric prior, and infers an appropriate number of topics. One important advantage of TOT is that it does not make a Markov assumption of topics; one topic can appear at one point in time, disappear for a while, then reappear at a later time point. Our model is comparable with TOT but explicitly captures when a topic first appears and allows the topic to reappear at a later time. One interesting work in [13] is that they include the time and spatial dimension simultaneously to capture the spatiotemporal pattern of topics. We did not directly consider the spatial dimension in our work, but the distance incorporated in our model can be applied to all measurable criteri, including the spatial dimension.

Applications of the probabilistic topic model are not restricted to analyzing a document corpus. One of the advantages of the LDA framework is the easiness of the extension to new models. These variants are widely applied to diverse fields[18, 9, 7, 12].

2.2 Bayesian Nonparametric Models

Bayesian methods are most powerful when the prior adequately captures appropriate beliefs. However, in parametric models, we need to set the parameters a priori, and that inflexibility may result in suboptimal inferences. For example, when we do not know the optimal number of mixture components in a mixture model or the order of the polynomial, we may need to try various settings of the parameter to get good results. Bayesian nonparametric models provide a way of getting flexible priors. nonpara-

metric models can automatically infer an adequate model complexity from the data without explicitly comparing model, by taking an infinitely many number of parameters.

The LDA framework cannot escape from the limitation of the parametric models. Many researchers have compared their results with different numbers of topics to find the optimal number of topics. However, these comparisons are often too expensive, and it is inefficient to find the optimal number of topics for every new corpus. Teh et al[25]. study a Bayesian nonparametric prior, a hierarchical dirichlet process(HDP), which extend a bayesian nonparametric prior, dirichlet process(DP). By using HDP they show their result is comparable with LDA in terms of optimal number of topics and held-out likelihood. The HDP-LDA also extends to capture the dynamics of topics over time. A very recent work by Zhang et al. [26] proposes a bayesian nonparametric model, called EvoHDP, that models multiple time-varying corpora. Emerging concerns on the bayesian nonparametric prior entails various posterior inference techniques such as variational inferences[14], MCMC sampling methods[5].

There is another concern about the exchangeability of nonparametric Bayeisan models. Exchangeability is a reasonable assumption in some applications, but in many cases it may not be correct. Blei and Frazier developed distance dependent Chinese restaurant process[1], distributions over partitions that allows for non-exchangeability. Though their work cannot be called a bayesian nonparametric model, it captures the non-exchangeability of data and allows the model to automatically capture the model complexity itself. Their model is well suited for a sequential corpus such as NYT news corpus and NIPS conference corpus. Griffin and Steel [8] propose a framework for Bayesian nonparametric modeling with continuous covariates, called order-based dependent dirichlet process. In this work they propose the nonparametric distribution to depend on covariates through ordering the random variables building the weights. In this work, the ordering constrains are not limited to the time dimension but also suitable for the spatial dimension.

Chapter 3. Chinese Restaurant Process

In this chapter, we introduce the Chinese restaurant process, used as a prior for Bayesian nonparametric methods, and its extension to the distance-dependent Chinese restaurant process.

3.1 Chinese restaurant process.

The Chinese restaurant process(CRP) is a probability distribution on partitions. The distribution is obtained by a process by which N customers sit down in a Chinese restaurant which has an infinite number of tables. The basic process of CRP is described as a sequential process by which customers sit down at a randomly chosen table drawn from its probability distribution. After N customers have sat down, their configuration represents a random partition.

In the CRP, the probability of a subsequent customer sitting at a table is computed by preceding customers already sitting at the table. Let K be the number of tables occupied by any customer, z_i the table assignment of the ith customer and n_k the number of customers sitting at the table k. The probability of each table for the ith customer is specified as follows:

$$\begin{aligned} p(z_i = k|\ z_{1:(i-1)}, \gamma) &= \frac{n_k}{\gamma + i - 1} \\ p(z_i = K+1|\ z_{1:(i-1)}, \gamma) &= \frac{\gamma}{\gamma + i - 1} \end{aligned}$$

where γ is a parameter. A probability of a customer sitting at the new table is proportional to γ, and a probability of customer sitting at table k is proportional to the number of customers sitting at the table k.

3.2 Distance dependent CRP.

Blei and Frazier[1] introduced the distant dependent Chinese restaurant process (ddCRP) in which the probability of a customer sitting at a table is dependent on the distances between customers. They also introduce the customer-based CRP as an alternative representation of a table-based-CRP. The major difference between the customer-based-CRP and the table-based-CRP is illustrated in Figure 3.1. In the customer-based-CRP setup, a customer is explicitly assigned to a table, and customers consist partitions around tables. In the customer-based-CRP setup, however, a customer is not explicitly assigned to a table, instead a customer is assigned to another customer or not assigned to any other customer. Although there is not explicit an table representation in the customer-based-CRP, the link structure between customers implicitly exhibit a clustering property. By using a customer-based representation, CRP can be further extended to ddCRP in which the existence of assignment between two customers is proportional to how close they are to each other.

Let c_i denote the index of the customer who has an inlink from 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. Then the ddCRP draws

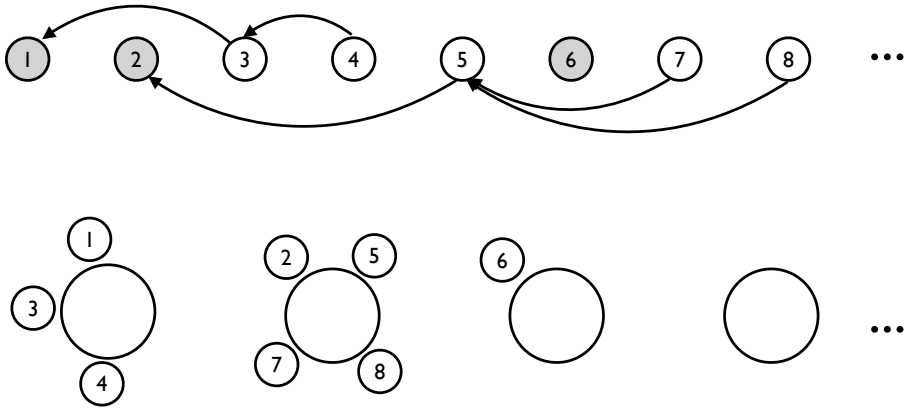


Figure 3.1: Comparison between two different representation of CRP(Customer based CRP vs. Table based CRP).

customer assignments c_i sequentially,

$$\begin{aligned} p(c_i = j|\ D, c_{1:(i-1)}, f, \gamma) &= \frac{f(d_{ij})}{\gamma + \sum_{j \neq i} f(d_{ij})} \\ p(c_i = i|\ D, c_{1:(i-1)}, f, \gamma) &= \frac{\gamma}{\gamma + \sum_{j \neq i} f(d_{ij})}. \end{aligned}$$

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 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 in which no customer can be assigned to a later customer.

Explicit table assignments do not occur in the customer based CRP, but the connected components implicitly exhibit a clustering property. Moreover a customer based CRP can be reverted to a table based CRP by summing over the distances within the same component of customers' link structure. Let K be an imaginary number of tables, which is the same as the number of connected components of customers, and z_i denote the index of the imaginary table of the ith customer. The probability of each table for the ith customer is specified as follows:

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

The partition probability over customers can be computed in the customer-based ddCRP simply as follows:

$$p(c_{1:N} \mid D, f, \gamma) = \prod_{i=1}^{N} \frac{\mathbf{1}[c_i = i]\gamma + \mathbf{1}[c_i \neq i] f(d_{ij})}{\gamma + \sum_{j \neq i} f(d_{ic_i})}.$$

However, in the table-based ddCRP, 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

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

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

Although this commitment to a sequential table-based ddCRP would lose the relative advantages of the customer-based ddCRP in the efficiency of sampling [1], we use the table-based distance dependent CRP in the rest of this paper. We chose to do so because in the hierarchical model presented in the next section, it is relatively easy to implement and compute the conditional probabilities. We discuss this in more detail in Sections 4.

3.3 ddCRP Mixture.

We ca use the CRP as a prior of a mixture model where the number of mixture components are unknown. In a ddCRP mixture, the observed data X arise as follows:

1. For each observation $x_i \in [1, N]$ draw assignment $z_i \sim$ dd-CRP(α, f, D).
2. For each observation $x_i \in [1, N]$
   (a) If $z_i \neq K + 1$ then set the parameter for the ith customer to $\theta_i = \theta_{z_i}$.
   Otherwise choose the parameter randomly, $\theta_i \sim G_0$
   (b) Draw the ith observation, $w_i = F(\theta_i)$

where K denotes the number of occupied tables by the preceding customers, G_0 is a base distribution over observations, and typically it is assumed a conjugate prior of a data distribution. The ith customer corresponds to a factor θ_i, and $F(\theta_i)$ denotes the distribution of the observation x_i given θ_i. For a corpus of documents observations are documents, G_0 is a Dirichlet distribution over the distribution of terms, and a distribution over terms is typically a multinomial distribution, therefore we can easily integrate out a factor θ_i by using conjugacy of the Dirichlet distribution and multinomial distribution.

Chapter 4. Distance Dependent Chinese Restaurant Franchise

4.1 Chinese Restaurant Franchise

A Chinese Restaurant Franchise (CRF) is a two-level hierarchical CRP. It is devised to model in which groups that each group is composed of data points drawn from a latent cluster. The number of clusters might be known or unknown, but we assume that the number of clusters is unknown and thus should be inferred. We note that all clusters can be shared across the groups.

A simple example of this type of problem can be found in the topic modeling task. In the general topic modeling task, it is common to view the words in a document as arising from a number of latent clusters, referred to as topics, where a topic is generally modeled as a multinomial distribution over words. In this setup, a document corresponds to a group, and is composed of multiple topics which correspond to the shared clusters.

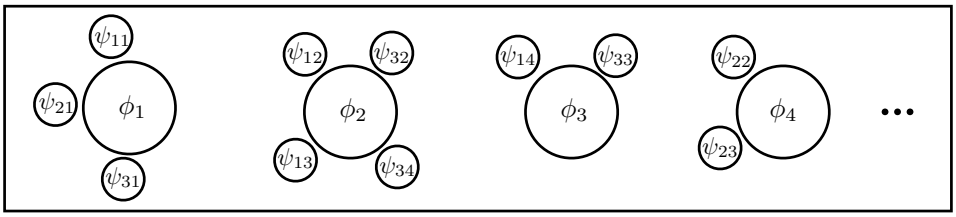
The CRF metaphor is explained by extending the basic CRP metaphor as follows. We have a restaurant franchise that has a shared menu across the restaurants. At each table of each restaurant one dish is ordered from the menu by the first customer who sits there, and it is shared among all customers sitting at the table. Multiple tables in multiple restaurants can serve the same dish.

The CRF is composed with two levels of CRP. First, the menu level CRP mediates the distribution of clusters shared among groups, and second, the customer level CRP contains N number of independent CRPs, where each CRP mediates the distribution over grouped data points. Figure 4.1 depicts this metaphor. In this figure, each rectangle in the second level CRP corresponds to a restaurant. The ith data point, a customer, in the jth restaurant is represented by the factor θ_{ji}, and a dish served for the tth table in the jth resturant is represented by ψ_{jt}.

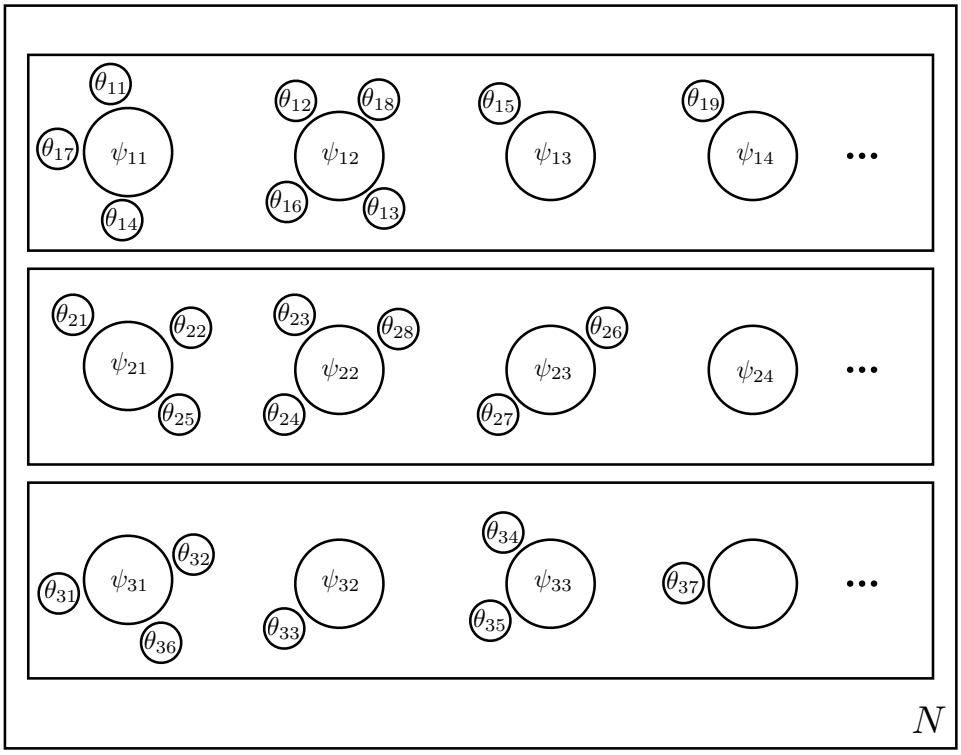
Dishes, ψ_{jt}, are also clustered by a first(menu) level CRP. That is, dishes correspond to customers in the first level CRP, and sit around a global menu table ϕ_k. The selection of dishes are purely decided by the other dishes already sitting around the menu. Note that the selection of dishes does not depend on the customer assignments in the second(customer) level CRP. By using this way of dish allocation, the global menu ϕ_k is shared across the restaurants.

It is worth re-emphasizing that each customer θ_{ji} eats one dish ψ_{jt} corresponding to a menu ϕ_k. We define additional notation as follows. We adopt n_{jt} denote a number of customers already sitting at the tth table in jth restaurant, m_{jk} to denote a number of tables eating the menu θ_k at restaurant j, and K to denote the available number of menus. In particular, we let t_{ji} be the index of the ψ_{jt} associated with θ_{ji}, and let k_{jt} be the index of ϕ_k associated with ψ_{jt}. The dots represent the marginal counts, so $m_{j\cdot}$ represents the number of tables in jth restaurant.

By using the metaphor described above we compute the conditional distribution of a customer θ_{ji} and a dish ψ_{jt}. First the conditional distribution of a customer θ_{ji} depends on the preceding customers, and can be computed in a straightforward way.



First(menu) level CRP



Second(customer) level CRP

Figure 4.1: Chinese Restaurant Francise.

$$\theta_{ji} \mid \theta_{j1}, ..., \theta_{j,i-1}, \alpha, G \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt}}{i-1+\alpha}\delta(\psi_{jt}) + \frac{\alpha}{i-1+\alpha}G \tag{4.1}$$

where α is a second level hyperparameter and δ is a point mass function centered at ψ_{jt}. New customer θ_{ji} is drawn from the right-hand side of mixture distribution. If θ_{ji} is drawn from the first summation then the customer sits at the tth table and we set θ_{ji} to ψ_{jt}, otherwise we set θ_{ji} to $\psi_{jm_{j\cdot}}$ drawn from a distribution G.

The G exists only in its role of distribution of ψ_{jt}, which is distributed according to the first level CRP, and again following the definition of CRP, the conditional distribution of ψ_{jt} is represented as follows:

$$\psi_{jt} \mid \psi_{11}, \psi_{12}, ..., \psi_{21}, ..., \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\delta(\phi_k) + \frac{\gamma}{m_{\cdot\cdot}+\gamma}H. \tag{4.2}$$

In the first level menu distribution, in the same way as the second level customer distribution a new dish ψ_{jt} is drawn from the right-hand side of mixture distribution with the first level hyperparameter γ. If ψ_{jt} is drawn from the first summation then the tth table is served the kth menu, and we set the ψ_{jt} to ϕ_k, and if ψ_{jt} is drawn from the second term, then the franchise serves a new menu ϕ_{K+1} drawn from H.

4.2 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.

- First, we can model first (menu) level CRP as 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.

- Second, we can model the (second) customer, level CRP as 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.

- Third, we can model both first and second level of CRP as ddCRP. By replacing both levels of CRP to 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 approachescan be applied in a straightforward way.

The conditional distribution of the ddCRF follows directly from the conditional distribution of the CRF in Equation 4.1 and Equation 4.2, only we need to consider the decay function and distance for

the ddCRF. 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, thus Equation 4.1 remains the same. However, the conditional distribution of the first level ψ_{jt} has to be computed by considering the distances with other Ψ, hence we have:

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

Measuring the distance between table jt and $j't'$ 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 bezero between tables at the same restaurant.

4.3 ddCRF Mixture Model

The observed data X arise as follows:

1. For each observation $x_{ji} \in [J_1, J_N]$ draw assignment $t_{ji} \sim \text{CRP}(\alpha, t_{j1:ji-1})$.
2. For each observation $x_{ji} \in [J_1, J_N]$
   (a) If $t_i \neq m_{j\cdot} + 1$ then set the parameter for the ith customer to $\theta_{ji} = \psi_{jt_{ji}}$.
   (b) Ohterwise draw assignment $k_{jt_{ji}} \sim \text{dd-CRP}(\gamma, f, D)$
      i. If $k_{jt_{ji}} \neq K + 1$ then set the parameter for the t_{ji}th table to $\psi_{jt_{ji}} = \phi_{k_{jt_{ji}}}$
      ii. Otherwise choose the parameter randomly, $\phi_{k_{jt_{ji}}} \sim H$
   (c) Draw the ith observation, $w_{ji} = F(\phi_{k_{jt_{ji}}})$

This procedure is same with the CRF except that the data is arising from a distribution with mixture component ϕ_k.

Chapter 5. Posterior Inference

In this chapter, we describe a Gibbs sampling method for the distance dependent Chinese restaurant mixture model. Calculating a exact posterior probability is intractable to due to the coupling between mixture probability and partition probability. Several posterior approximation techniques, MCMC, variational inference techniques, are introduced for the CRP mixture and its related Bayesian non parametric mixture models.

5.1 Posterior Sampling in the ddCRF

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 $\mathbf{t}$ and $\mathbf{z}$ based on the observed data set $\mathbf{x}$. 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 probability of data x_{ji} given all other variables is computed as:

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

If we use a dirichlet distribution H and a multinomial distribution $p(\cdot|\phi_k)$ this equation can be further simplified as follows:

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

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.

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

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

$$
\begin{aligned}
p(x_{ji} \mid \mathbf{x}^{-ji}, \mathbf{t}^{-ji}, t_{ji} = t^{new}, \mathbf{k}) &= \sum_{k=1} \frac{\sum_{j't' \neq jt, k_{j't'}=k} f(d_{j't',jt})}{\sum_{j't' \neq jt} f(d_{j't',jt}) + \gamma} \\
&\quad \times p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}^{-ji}, t_{ji} = t^{new}, \mathbf{k}, t^{new} = k) \\
&\quad + \frac{\gamma}{\sum_{j't' \neq jt} f(d_{j't',jt}) + \gamma} \\
&\quad \times p(x_{ji}|\mathbf{x}^{-ji}, \mathbf{t}^{-ji}, t_{ji} = t^{new}, \mathbf{k}, t^{new} = k^{new})
\end{aligned} \tag{5.1}
$$

If t_{ji} is drawn from the second term then we have to draw $k_{jt^{new}}$ for the new table from following distribution.

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

We use the same computation as in Equation 5.1(omitting the common denominator), but we present this again to clarify the sampling procedure.

5.2 Sampling Hyperparameters

To improve our model, we place a prior for first and second level hyperparameter γ and α, as used in the original CRF. Sampling α is done in the same way as [25], thus we discuss here how to sample γ. We note that the probability of γ is conditionally independent of $\mathbf{x}$ and $\mathbf{t}$ given $\mathbf{k}$. From this fact, the probability of γ is

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

where $p(\gamma)$ is the prior on the parameter. As we derived in section 3.2

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

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 [16]. This method evaluates the probabilities on a finite set of points, and approximates the inverse cdf $p(\gamma|\mathbf{k})$ using these points, and samples from the approximated inverse cdf.

Chapter 6. Experiments

This section describes the experiments to evaluate the performance of the distance dependent Chinese restaurant franchise on four different text datasets and demonstrates how the ddCRF performs better than the original CRF. The data sets for the experiments include conferences, SIGIR, SIGMOD, SIGGRAPH abstracts, collected through the ACM digital library,[1] and the NIPS article data set[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. We modeled these datasets by using the ddCRF to capture the topic changes within a conference though time.

6.1 Dataset

Here are the details of the four datasets we used in the experiments.

- NIPS: We use 1,740 NIPS articles published between 1988-1999. This corpus consists of 6,455 unique terms, around 450K observed words and an average of 133 articles per year. We randomly sampled 20% of words for each article.
- SIGIR: We use 1,838 SIGIR paper abstracts published between 1978-2010. This corpus consists of 1,998 unique terms, around 106K observed words, and an average 56 of articles per year.
- SIGMOD: We use 2,311 SIGMOD paper abstracts published between 1978-2010. This corpus consists of 2,745 unique terms, around 165K observed words, and an average of 56 articles per year.
- SIGGRAPH: We use 783 SIGGRAPH paper abstracts published between 1974-1991. This corpus consists of 2,381 unique terms, around 53K observed words and an average of 44 articles per year.

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. Figure 6.1 shows the number of articles for each year. We trained the original-CRF mixture model, ddCRF mixture model, and the basic latent Dirichlet allocation(LDA) 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. The hyperparameters were given vague gamma priors with scale parameter of 1 and shape parameter of 1. It is possible to sample the Dirichlet parameter, but in that case, the number of topics increases too much[21] and it is not efficient in practice, so we just leave it as a constant.

6.2 Comparison to Original CRF

First, we compare ddCRF with CRF using two widely used metrics, held-out likelihood and complexity. Topic models are typically evaluated by either measuring the performance on some secondary

[1] http://portal.acm.org/
[2] http://www.cs.utoronto.ca/∼sroweis/nips

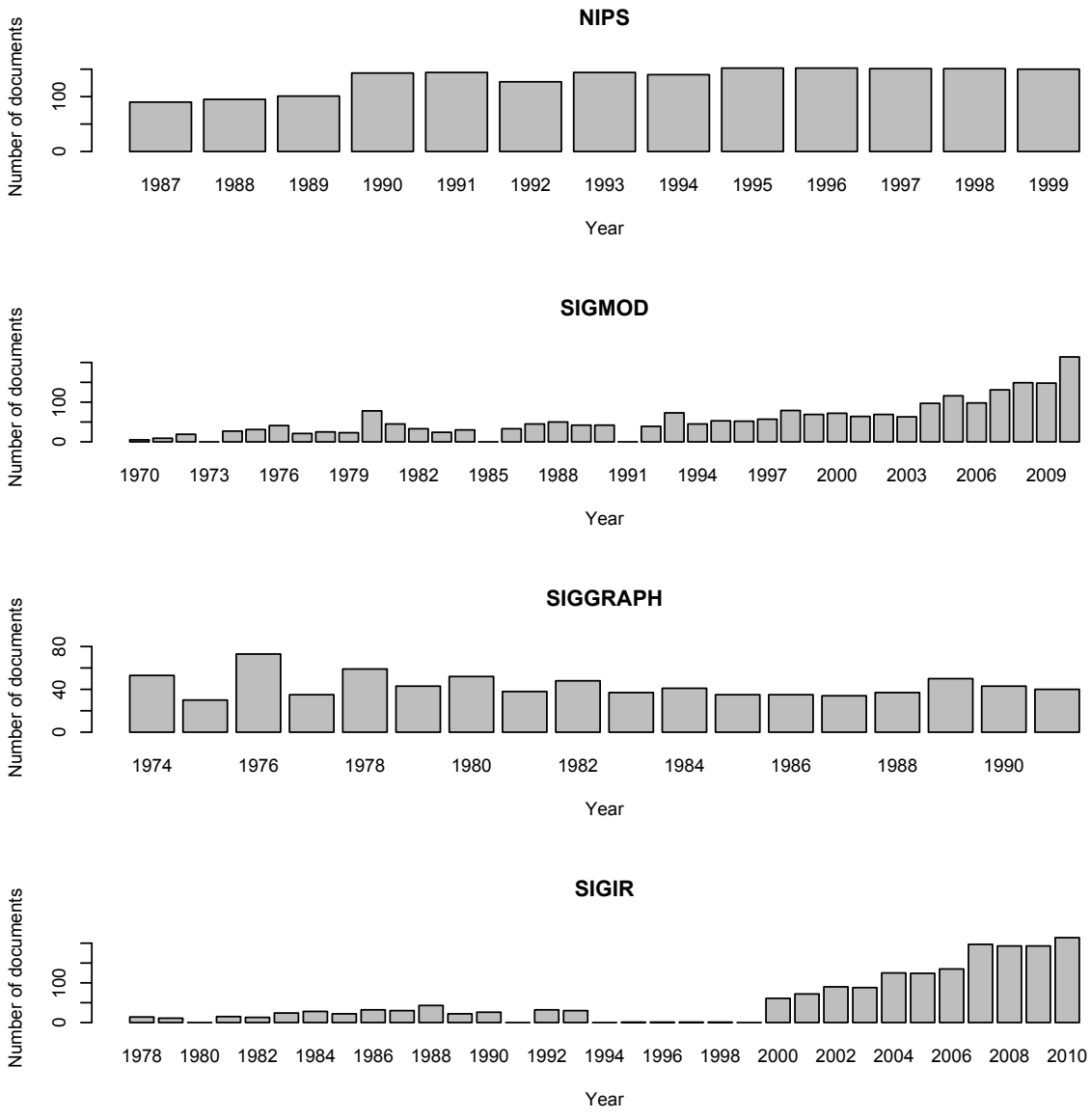


Figure 6.1: Number of articles by year

task, such as document classification or information retrieval, or by estimating the probability of unseen held-out documents given some training documents. A better model will give rise to a higher probability of held-out documents on average.

Held-out Likelihood: Held-out likelihood is widely used in a topic modeling community to compare how well the trained model explains the held-out data(cf. [17] [20] [25]).

$$\text{Held-out Likelihood} = log\ p(W|M_{train}),$$

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

Figure 6.2 shows the held-out likelihood of the datasets. For all datasets ddCRF shows better held-out likelihood than the CRF regardless of decay parameter and decay function. The results exhibit that the held-out likelihood gets lower when the decay parameter increases on average.

Complexity: bayesian nonparametric and its related methods are often used for alternatives 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 held-out likelihood, the less complex model is preferred. To capture the complexity of models, we compute a complexity of each model defined by Blei and et. al[1]. 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 the complexity captures how many topics are used to explain each document and sum it through entire corpus. Lower complexity indicates that the model uses fewer of number of topics to represent the corpus, and higher complexity indicates that the model decomposes the data into many dimensions and might be over-fitting to the data.

Figure 6.3 shows the complexity of four different datasets. The average complexities of ddCRF are better than the CRF except the SIGMOD dataset.

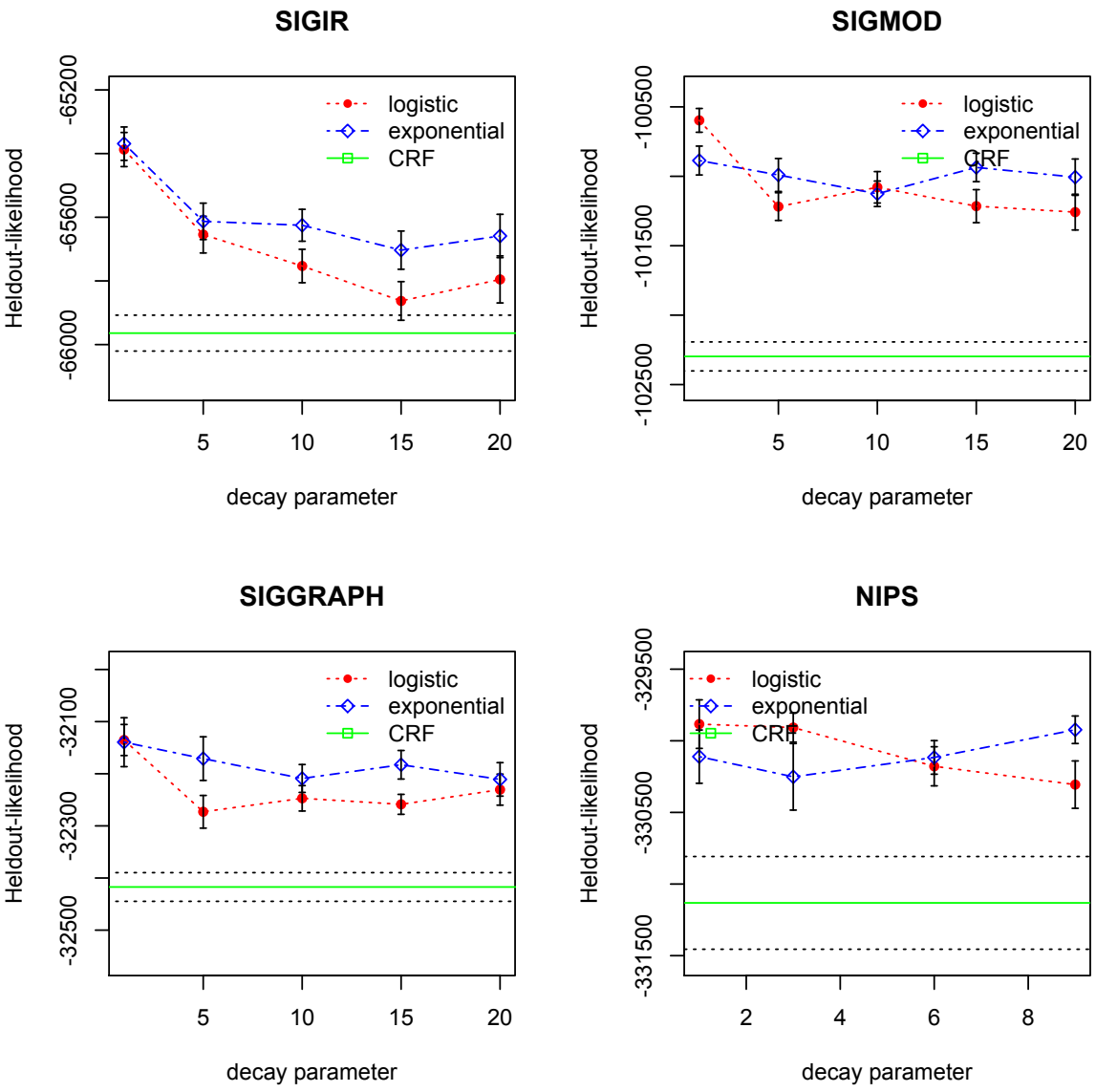


Figure 6.2: Heldout-Likelihood. Higher is better

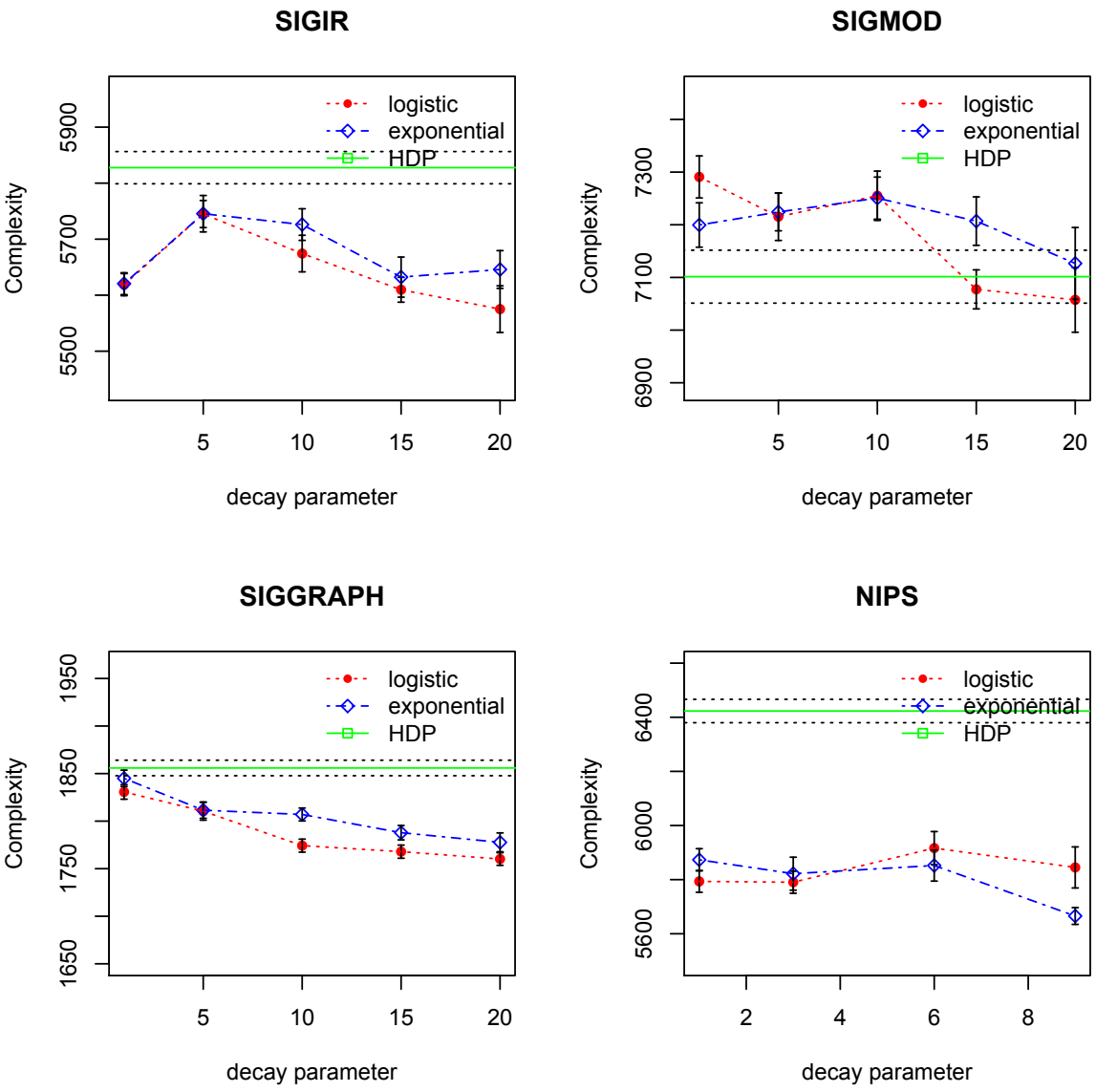


Figure 6.3: Model Complexity. Lower is better

Appearance 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 table to be assigned to a topic that first appears at a later point in the dataset. Therefore the posterior topic assignment explicitly represents when the topic first appears. In contrast, the disappearance of topics cannot be explicitly captured from posterior assignment. With certain decay function and parameter, however, we can guess when the topic disappeared by computing a decay function with the time at which the topic last appeared. For example, with the logistic decay function of five decay parameter, if a certain topic as not be allocated in the last seven years, we can conclude that topic has disappeared.

To measure the trends of topics 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 total number of terms at time slice t.

We choose three topics from the SIGIR dataset based on the training results of ddCRF and choose the most similar topics based on the results of original-CRF, similarity measured in terms of JS divergence. Figure 6.4 shows the intensity of those five topics over time. The topics drawn in red line are identified by the ddCRF and the topics drawn in blue line are identified by the original-CRF. The topics in the first roware about "spam filtering", the topics in the second row are about "collaborative filtering", and the topics in the third row topic is about "file structure and record". We found similar topics from both models by computing JS divergence between topics. However there is no topic similar to the "file structure and record" topic in the result of original-CRF: the most similar topic's value of JS divergence is about 0.6 but that topic seems quite dissimilar to this topic. The "spam filtering" topic has appeared with the rapid growth of the usage of e-mail. The topic from ddCRF captures the emergence of this topic around 2000, but the similar topic from original-CRF seems to be a mix of a topic related to spam filtering and another topic related to news and events. If we consider that the spam filtering topic should have emerged in the SIGIR conference around 2000, then the original-CRF did not capture the phenomenon well. We can see the same problem in the "collaborative filtering" topic. It gains a lot of attention after 2000, but the original-CRF also identified a similar topic over the entire dataset. The last topic at third row shows the disappearance of a topic, which last appeared in 1988 and disappeared after that. As time increases, the value of decay function also goes to zero, therefore we can conclude that there is zero probability of that topic appearing in 2010.

6.3 Comparison to LDA

We also comopare our result to LDA with various number of topics. The number of topics found by the ddCRF is larger than the CRF in all cases. To show that ddCRF outperforms LDA with the same number of topics, we trained LDA with various number of topics. The results indicate that our model outperforms LDA, and the performance is not related to the number of topics. Figure 6.6 displays held-out likelihoods with four datasets, and we use the best held-out likelihood of the ddCRF results to compare with the results of LDA.

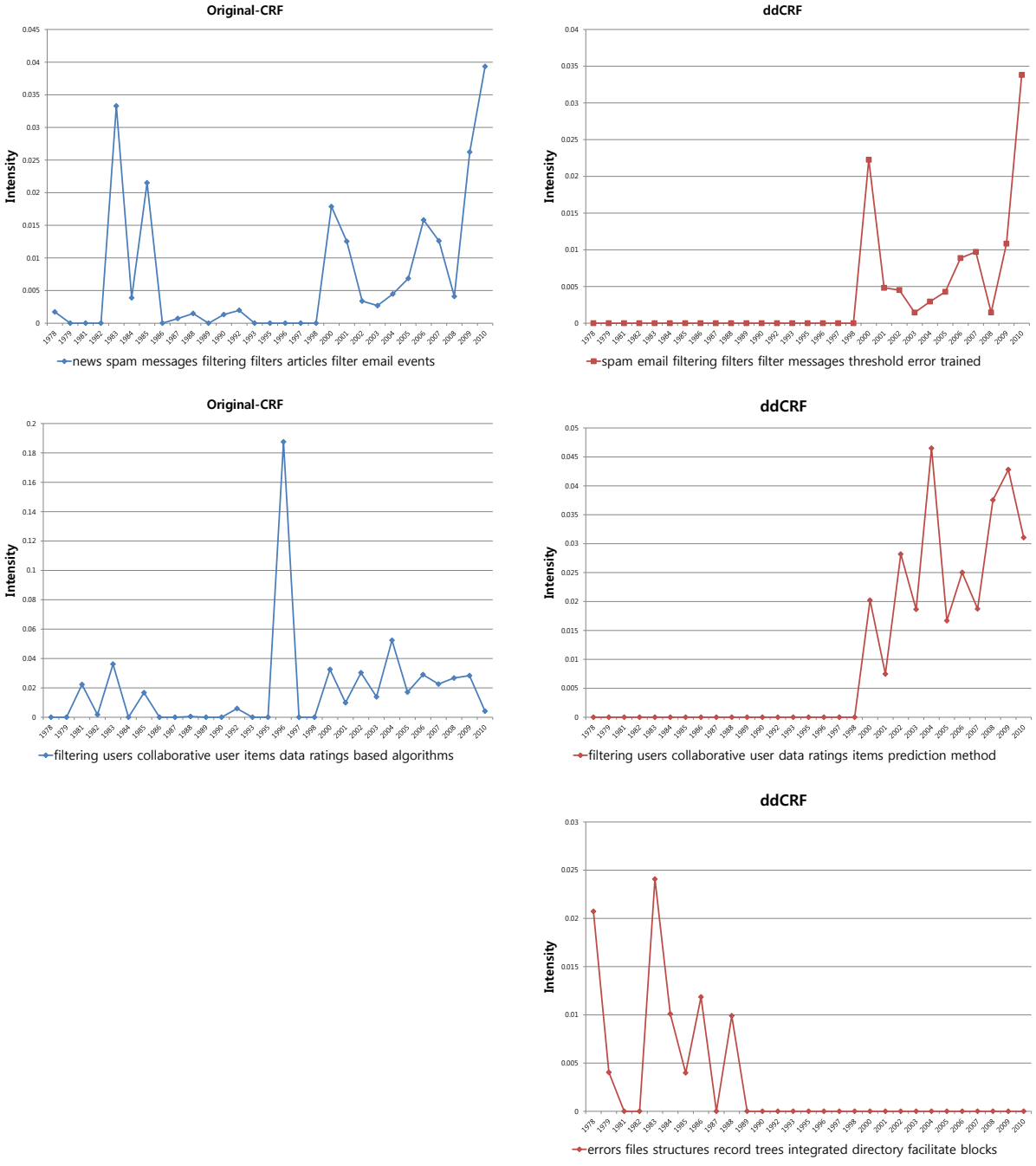


Figure 6.4: Topic proportion over time. Identified from SIGIR

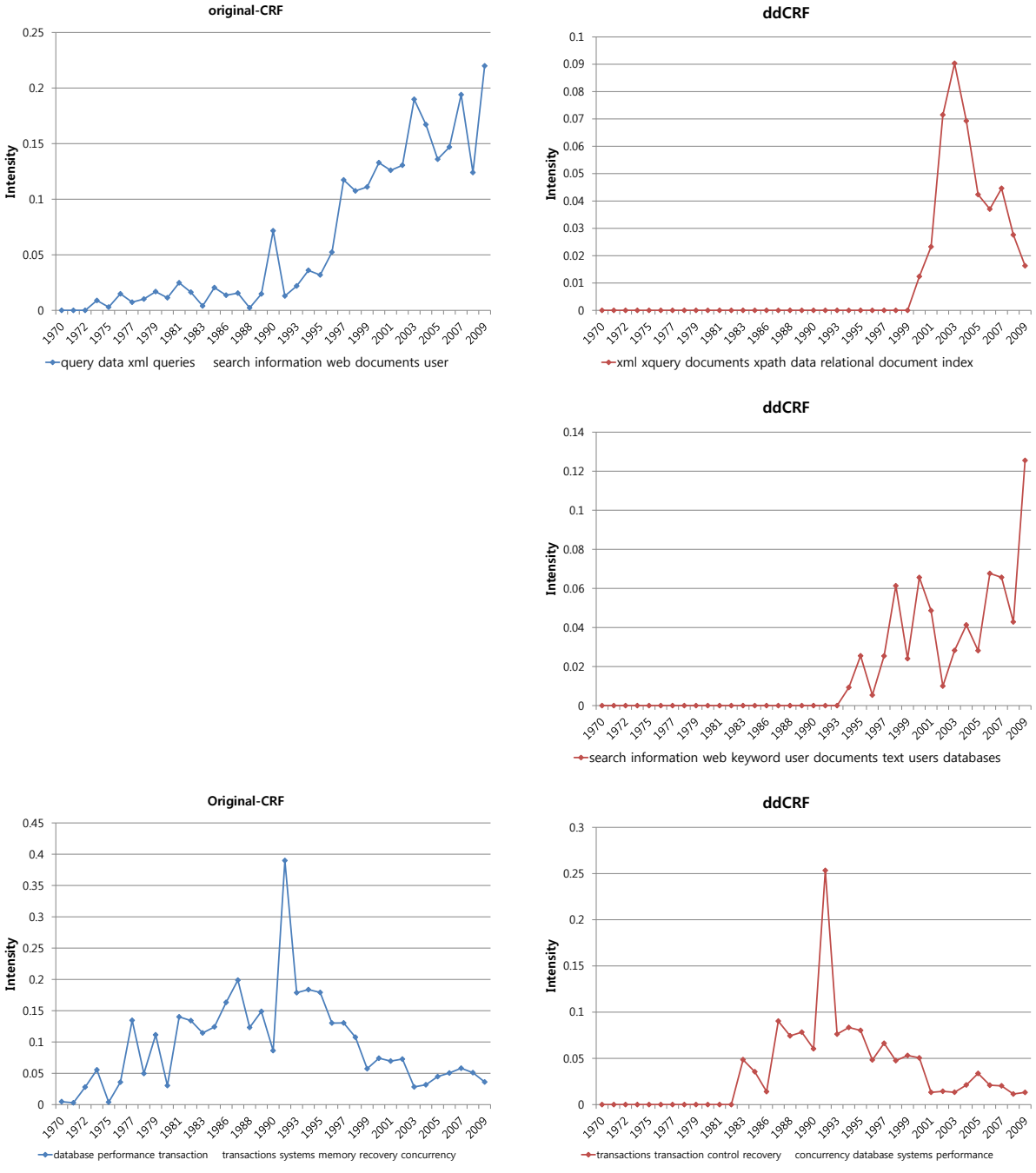


Figure 6.5: Topic proportion over time. Identified from SIGMOD

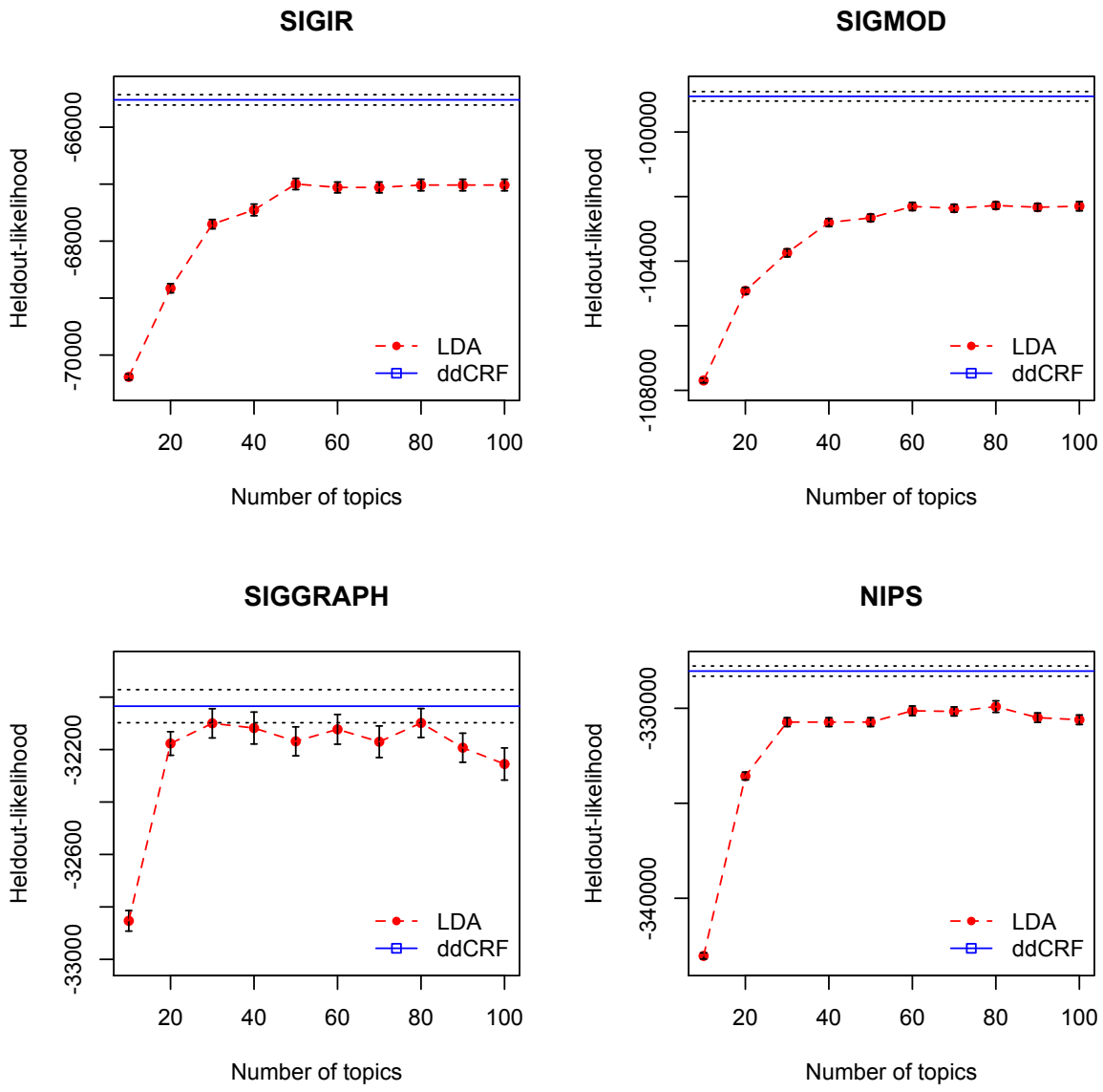


Figure 6.6: Held-out likelihood comparison to LDA. Higher is better

Chapter 7. Concluding Remarks

Throughout this thesis, we propose the distance dependent Chinese restaurant franchise in which the latent variables are sequentially allocated by the distance dependent CRP. A Gibbs sampling scheme is proposed to infer the model. Results on four conference datasets show the topic emergence, disappearance, and evolution within a time-varying corpus. The ddCRF can also give a more accurate predictive model.

There are many was that the work described here can be extended. One direction is to use more efficient models. We have computed the conditional probability of dishes by summing over the previously allocated dishes, but this procedure is expensive. Perhaps the most promising extension to the methods presented here is to incorporate a model of how words drift within a cluster.

References

[1] D Blei and P Frazier. Distance dependent chinese restaurant processes. In *Proceedings of the 26th international conference on Machine learning*, ICML '10, Jan 2010.

[2] D Blei and J Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

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

[4] David Blei and John Lafferty. Topic models. *Text Mining: Theory and Applications*, pages 71–93, 2009.

[5] David M. Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144, 2005.

[6] L Bolelli, Ş Ertekin, and C Giles. Topic and trend detection in text collections using latent dirichlet allocation. *Advances in Information Retrieval*, 2009.

[7] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 233–240, New York, NY, USA, 2007. ACM.

[8] J.E. Griffin and M.F.J. Steel. Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, 101:179–194, March 2006.

[9] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, volume 17, pages 537–544, 2005.

[10] Matthew D. Hoffman, David M. Blei, and Perry R. Cook. Finding latent sources in recorded music with a shift-invariant hdp. In *in International Conference on Digital Audio Effects (DAFx) (under review*, 2009.

[11] Derek Hao Hu, Xian-Xing Zhang, Jie Yin, Vincent Wenchen Zheng, and Qiang Yang. Abnormal activity recognition based on hdp-hmm models. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1715–1720, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[12] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA, 2009. ACM.

[13] Q Mei, C Liu, H Su, and C Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. *Proceedings of the 15th International Conference on World Wide Web*, 2006.

[14] Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS*, 9(2):249–265, 2000.

[15] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

[16] Christian Ritter and Martin A. 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.

[17] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[18] R Socher, S Gershman, A Perotte, and P Sederberg. A bayesian analysis of dynamics in free recall. *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[19] KA Sohn and EP Xing. A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *Annals*, 3(2):791–821, 2009.

[20] Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why Priors Matter. In *Proceedings of NIPS*, 2009.

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

[22] Chong Wang, David Blei, and David Heckerman. Continuous Time Dynamic Topic Models. In *Proceedings of ICML*, 2008.

[23] X Wang and A McCallum. Topics over time: a non-markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[24] X Wang, K Zhang, X Jin, and D Shen. Mining common topics from multiple asynchronous text streams. *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009.

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

[26] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. *Proceedings of the sixteenth ACM SIGKDD international conference on Knowledge discovery and Data Mining*, 2010.

Chapter A. Topics inferred by ddCRF

Table A.1: Top 10 topics identified from SIGIR

1	2	3	4	5
retrieval	information	retrieval	search	classification
query	retrieval	documents	query	learning
terms	systems	performance	queries	data
document	paper	evaluation	results	ranking
information	data	relevance	web	method
documents	based	systems	user	training
term	user	measures	users	algorithms
results	search	relevant	information	algorithm
performance	model	results	engine	text
using	techniques	test	engines	methods
6	7	8	9	10
clustering	model	web	text	index
document	models	pages	news	time
documents	language	page	topic	compression
algorithm	probabilistic	algorithm	summarization	inverted
paper	document	link	sentences	large
space	retrieval	search	detection	performance
data	modeling	graph	based	methods
information	distribution	ranking	summaries	text
based	probability	pagerank	documents	file
method	framework	links	approach	efficient

Table A.2: Top 10 topics identified from SIGMOD

1	2	3	4	5
data	query	performance	data	data
database	queries	data	database	spatial
model	data	database	applications	sampling
language	database	index	systems	method
paper	algorithms	storage	information	time
systems	paper	tree	management	approximate
base	results	file	web	objects
relational	problem	paper	application	paper
information	processing	memory	new	queries
management	optimization	algorithms	server	similarity
6	7	8	9	10
transactions	data	rules	data	search
transaction	mining	database	schema	information
control	clustering	rule	sources	web
recovery	large	dependencies	query	keyword
concurrency	algorithm	recursive	integration	user
database	algorithms	functional	mappings	documents
systems	rules	algebra	mapping	text
performance	clusters	logic	source	users
protocols	analysis	relational	schemas	databases
paper	analysis	relations	queries	structured

Table A.3: Top 10 topics identified from SIGGRAPH

1	2	3	4	5
graphics	algorithm	model	parallel	graphics
display	surface	constraints	performance	systems
user	surfaces	motion	architecture	cam
data	method	animation	pixel	cad
design	new	control	processor	industry
interactive	image	models	second	panel
systems	algorithms	dynamic	buffer	japanese
used	objects	modeling	memory	development
paper	images	objects	display	research
use	texture	simulation	frame	solid
6	7	8	9	10
points	objects	sampling	data	business
surface	language	control	map	session
three-dimensional	graphical	curves	maps	systems
method	structure	curve	base	decision
used	object	points	terrain	people
number	picture	amp	mapping	user-oriented
line	data	fit	information	new
shape	based	new	cartographic	computers
lines	pattern	splines	network	years
sections	tool	noise	spatial	impact

Table A.4: Top 10 topics identified from NIPS

1	2	3	4	5
network	model	model	circuit	memory
learning	cells	data	analog	network
neural	neurons	models	chip	dynamics
training	input	distribution	figure	state
networks	cell	gaussian	neural	networks
set	visual	probability	input	time
input	figure	mixture	vlsi	neural
using	response	algorithm	output	neurons
used	activity	bayesian	voltage	states
function	time	parameters	current	equations
6	7	8	9	10
speech	learning	image	image	functions
recognition	state	object	images	function
word	reinforcement	images	recognition	theorem
words	policy	field	object	threshold
hmm	control	objects	visual	let
character	value	vision	features	networks
training	optimal	figure	face	bounds
characters	action	receptive	feature	neural
speaker	actions	flow	objects	bound
phoneme	function	surface	human	size

Summary

Distance Dependent Chinese Restaurant Franchise

확률론적 주제 모델은 분류되지 않은 대규모의 문서집합에 내재하는 여러 주제들을 자동으로 분석할 수 있게 해주는 방법을 제공한다. 이와 같은 모델들은 텍스트 분석, 인지과학, 계산 생물학 등의 여러 분야에서 데이터 내에 잠재된 의미있는 패턴들을 찾아내는데 적용되어 왔다. 대규모 데이터의 분석에 대한 요구와 여러 연구자들의 노력으로 인해 주제 모델의 응용과 이종 모델들의 개발은 주제 모델이 더욱 많은 분야에 적용되는데 기여를 하고 있다. 본 연구는 거리 의존관계를 이용한 변종 비모수적 베이지안 확률 모델을 제안한다. 우리는 이 모델의 개발을 위해 우선 중국인 식당 프로세스(CRP)라고 불리는 비모수적 베이지안 모델의 사전 확률 분포에 대하여 알아 본 후, 이를 계층적 구조로 확장시킨 중국인 식당 프렌차이즈(CRF)라고 불리는 비모수적 베이지안 주제 모델에 대해 알아 본다. 최종적으로 중국인 식당 프렌차이즈 모델에 거리 의존관계를 포함시킨 모델(ddCRF)을 제안한다. 사후 확률의 추론을 위해서는 통계 추론을 위해 널리 사용되고 있는 마르코프 체인 몬테 카를로 방식을 제안한다. 본 연구는 총 4가지 학회들의 논문 발행물들을 사용하여 모델의 성능 평가를 진행한다. 성능 평가 결과 본 모델은 기존에 널리 사용된 모델인 잠재 디리클레 할당모형(LDA)와 중국인 식당 프렌차이즈(HDP)보다 가능성(Held-out Likelihood)와 복잡도(Complexity) 측면에서 뛰어난 결과를 보여준다. 또한 제안된 모델을 사용한 결과는 시간대 별로 구성된 문서집합에서 주제들을 분석 할 때 가까운 시간대에 작성된 문서들이 서로 비슷한 주제들을 가진다는 직관을 반영한다, 또한 시간대에 따른 주제의 발생과 소멸에 대해서도 기존의 모델들에 비해 향상된 결과를 보여준다. 본 모델의 또다른 장점은 기존에 사용되어 왔던 모수적 주제 모델들에 비해서 주제의 숫자를 선행적으로 결정해 주어야 할 필요가 없다는 점이다.

감 사 의 글

2년이라는 짧지 않은 석사 생활 동안 많은 도움을 주신 여러분들께 감사의 말을 전합니다. 우선 석사 기간동안 함께 동거동락한 우리 연구실 사람들, 현종이형, 영민씨, 요한이, 선준씨, 준희, 수인이에게 감사의 말을 전합니다. 연구실 사람들과 함께 지새운 수많은 낮과 밤을 잊지 못할 것 같습니다. 또 제가 하고싶은 연구를 마음껏 할 수 있게 지원해준 오혜연 교수님, 2년간 나의 투정과 짜증들을 모두 받아내 준 민정이, 제가 가고싶은 길로 갈 수 있도록 최선을 다해 지원하고 지지해주신 부모님께 감사의 말을 전합니다. 지나간 시간이 항상 순탄하기만 했던것은 아니지만 어려운 순간마다 항상 함께 있어준 사람들이 있어서 행복하다고 기억될 수 있는 석사 생활을 할 수 있었던 것 같습니다. 석사 생활을 통해 성취한 것보다는 부족한 것을 더 많이 깨닫고 졸업하지만, 이러한 부족함이 앞으로의 인생을 살아가는데 있어서 더욱 큰 도움이 될 것이라 믿습니다.

마지막으로, 언제나 저의 든든한 지원군이자 버팀목이었던 하늘에 계신 아버지께 이 논문을 바칩니다.

이 력 서

이 름 : 김 동 우

생 년 월 일 : 1984년 10월 8일

주 소 : 경북 포항시 남구 연일읍 생지리 삼도한솔타운 102동 406호

E-mail 주 소 : dw.kim@kaist.ac.kr

학 력

2000. 3. – 2003. 2. 포항고등학교

2003. 3. – 2009. 2. 성균관대학교 정보통신학부 (B.S.)

2009. 3. – 2011. 2. 한국과학기술원 전산학과 (M.S.)

경 력

2008. 1. – 2008. 2. 한국전자통신연구원 인턴연구원

학회활동

1. **Dongwoo Kim**, Alice Oh, *Topic Chains for Understanding a News Corpus*, 12th International Conference on Intelligent Text Processing and Computational Linguistics(CICLING 2011), Tokyo, Japan, 2011.
2. Il-Chul Moon, **Dongwoo Kim**, Yohan Jo, Alice Oh, *Learning Influence Propagation of Personal Blogs with Content and Network Analyses*, Workshop on Finding Synergies Between Texts and Networks at the IEEE International Conference On Social Computing(SocialCom 2010), Minneapolis, USA, 2010.
3. Il-Chul Moon, **Dongwoo Kim**, Yohan Jo, and Alice H. Oh, *Learning Influence Propagation on Personal Blogs*, The 30th Annual International Sunbelt Social Network Conference, Trento, Italy, Jun 29-Jul 4, 2010
4. **Dongwoo Kim**, Yohan Jo, Il-Chul Moon, and Alice Oh. *Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users*, Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems (CHI 2010), Atlanta, USA, 2010.