# Purchase Prediction Based on a Non-parametric Bayesian Method

Yezheng Liu
School of Management,
Hefei University of Technology
Hefei, China
liuyezheng@hfut.edu.cn

Tingting Zhu
School of Management,
Hefei University of Technology
Hefei, China
zhutingting-hfut@foxmail.com

Yuanchun Jiang
School of Management,
Hefei University of Technology
Hefei, China
ycjiang@hfut.edu.cn

## Abstract

*Predicting customer's next purchase is of paramount importance for online retailers. In this paper, we present a new purchase prediction method to predict customer behavior based on non-parametric Bayesian framework. The proposed method is inspired by topic modeling for text mining. Unlike the conventional methods, we regard customer's purchase as the result of motivations and automatically determine the number of user purchase motivations. Given customer's purchase history, we show that customer's next purchase can be predicted by non-parametric Bayesian model. We apply the model to real-world dataset from Amazon.com and prove it outperforms the traditional methods. Besides that, the proposed method can also determine the number of the motivations owned by users automatically, rendering it a promising approach with a good scalability.*

## 1. Introduction

An accurate prediction of what a customer will purchase is of paramount importance to successful online retailing [1]. Purchase prediction is the basis of product recommendation system. It contributes to determining the positions of the customer's search query results, optimizing product collections to be displayed on personalized landing pages, planning the inventory at the point of sales and warehouse, and making strategic decisions on the manufacturing processes [2].

Purchase prediction has received much attention on a long history in consumer research [3-5]. Most purchase prediction studies are based on user behavior data [6], social network information [7], or user-generated text information [8]. However, customer information is unavailable in some cases. In terms of the product level, such features are often absent. Even if the information about product is available, it is difficult to extract appropriate variable. Consequently, the typical data that online retailers can use to predict future customer behavior is the customer purchase history. In this paper, we utilize purchase history data to predict purchase behavior.

Many online retailers use collaborative filtering algorithms to predict customer future purchase behavior in purchase history data. The algorithms mostly depend on counts of the co-occurrence of products [9, 10]. There are some limits when applying collaborative filtering algorithm. In a small dataset, it results in information loss. In a large combinations of products, it makes the co-occurrence count matrix sparse and causes a few matches in the customer base [1]. To address these weaknesses, Jacobs et al. [1] proposed a novel purchase prediction algorithm using latent Dirichlet allocation (LDA). LDA is one of the most famous topic models in the text modeling literature. Traditionally, LDA describes a document by associating the words in the text to latent topics. In the purchase prediction environments, Jacobs et al. regarded a customer's purchase history as a document and used products as words. A customer's certain preference for products was represented as a topic which can be called the motivation. LDA has a better predictive performance compared with traditional methods, it also has a big problem: the number of motivations is set artificially. Consumers' latest purchases may contain unseen motivations. Prespecifying the number of topics inhibits the incorporation of such unseen motivations, leading to inaccurate predictions.

Inspired by Jacobs' work, we regard purchase behaviors as the results of specific motivations. Different from the previous work, we exploit the non-parametric Bayesian method to predict customer behavior. The method automatically determines the number of motivations. Besides, by defining a global prior, all customers can share the same motivation. The non-parametric Bayesian method is called hierarchical Dirichlet process (HDP) mixture model. HDP can be considered a sequence of weighted-motivations, each of

URI: https://hdl.handle.net/10125/59572
ISBN: 978-0-9981331-2-6

them can be shared with each customer. It employs a two-level generation process to predict customer's next purchase. The numerical studies are based on the real-world data collected from amazon.com. The results show that the proposed method can automatically find diverse motivations and accurately predict customer future purchase. The contributions of this paper are threefold.

1) To the best of our knowledge, this is the first paper to predict customer future purchase based on hierarchical Dirichlet process mixture model.

2) HDP can automatically determine the amount of user purchase motivations.

3) In the tasks of purchase prediction, experimental results show that the proposed HDP method outperforms benchmarks in terms of precision and recall.

The remainder of the paper is organized as follows. Section 2 surveys the related work on the probabilistic topic models and model-based approaches for purchase prediction. We will detail the proposed model and the inference process in section 3. The experimental results are shown in section 4. Section 5 gives the conclusions and the directions of future works.

## 2. Related Works

In this section, we will review related works from two perspectives: probabilistic topic models and model-based approaches for purchase prediction.

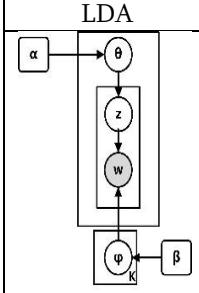
### 2.1. Probabilistic Topic Models

Probabilistic topic models are used to extract hidden semantic structure from large-scale text data. Generally, previous studies on topic modeling utilize matrix factorization or probabilistic graphical model to reveal hidden semantic in documents. For example, latent semantic analysis (LSA) is the earliest one that utilize singular value decomposition (SVD) to reveal the words relationships within documents [11]. By introducing hidden topic concept, the target of LSA is to transform original document-term matrix to a low-rank approximation matrix. Probabilistic latent semantic analysis (PLSA) [12] is another topic modeling method which is based on probabilistic statistics. It assumes that a document is a mixture distribution over topics, where a topic is a mixture distribution over words. By adding Dirichlet priors on document-topic distribution and topic-word distribution, LDA [13] extends original PLSA which is a more complete probabilistic generative model.

LDA is currently the most popular probabilistic topic model which can extract the hidden topics from a document. It assumes that a document contains diverse topics and is denoted by a multinomial topic distribution. Each word in a document is generated by a topic. Table 1 gives the probabilistic graphical model and the generative process of the LDA. Circle nodes on the graph are random variables and the shaded ones are observable variables. Prior distribution is represented by the rounded rectangle which are α and β in this model. The straight lines with arrows denote the dependency between random variables. The rectangular boxes represent repetitions, and the letters in the bottom right corner represent the number of repetitions. For a given set of documents, the model training process is to estimate the document-topic distribution θ and the topic-word distribution φ. The online variational Bayes (VB) method and Gibbs sampling can be used to estimate the LDA parameters [13, 14].

LDA can extract refined topics from documents, nevertheless, 1) the number of topics is set artificially. 2) the optimal number of motivations cannot be used for other datasets. To solve these problems, researchers have also developed many other topic models. Hierarchical Dirichlet process (HDP) mixture model is one of the most famous one [15]. With a global shared prior distribution, HDP can automatically determine the number of topics across different documents. Since customer's purchase records are a constantly changing set, it is apparently intractable to discover the number of motivations artificially. Thus, this paper utilizes HDP to determine the number of motivations based on customer's purchase history data.

**Table 1. The graph model representation of the LDA and the generative process**

| LDA | The Generative Process |
|---|---|
|  | 1.For each document d ∈ D:<br>Draw topic mixture proportion $\theta_d \sim Dirichlet(\alpha)$;<br>2.For each latent topic dimension k ∈ $[1, K]$:<br>Draw $\varphi_k \sim Dirichlet(\beta)$;<br>3.For each word $w_{di}$ in document d:<br>(i)Draw topic assignment $z_w \sim Multinomial(\theta_d)$;<br>(ii)Draw word w$\sim Multinomial(\varphi_{z_w})$; |

### 2.2. Model-based approaches for purchase prediction

Predicting customer purchase behavior provides vital information for online retailing. In recently, more and more purchase prediction studies are based on user behavior data, social network information or user-generated text information. For example, Li et al. proposed a new method to predict user purchase

behavior based on user behavior logs. They focus on predicting next-one-day purchase behavior [8]. Yuho et al. approached a problem and attempted to predict purchasing actions of Twitter users used social network information [7]. In terms of whether a specified user purchased a certain brand, Zhao et al. proposed a framework with a threshold-moving approach to predict sets of pairs (user id and brand id) according to their historical activity records [6].

However, the typical data that online retailers can utilize to predict future customer behavior is the customer purchase history. Predicting user purchase behavior by model-based methods has a long history [16-19]. Discrete choice model (DCM) [17, 20] is one of the most famous methods. It describes, explains, and predicts choices between two or more discrete alternatives, such as studying consumer demand and predicting customer's next purchase. Logistic regression [21] and probit regression [22] are two well-known basic discrete choice models. A collaborative filter is a deterministic algorithm for predicting customer future purchase behavior [23-25]. The algorithms mostly rely on counts of the co-occurrence of products in purchase history data. In recent years, researchers try to predict customer purchase based on topic modeling method. For example, Jacobs et al. [1] utilized latent Dirichlet allocation (LDA) to predict future customer purchase, which is a parametric Bayesian method. They compared its predictive performance with those of a collaborative filter and a discrete choice model. LDA provides a better predictive performance and outperforms the other methods.

## 3. Proposed Approach

In this section, we present our prediction method. HDP and LDA share the subsequent notation: the products are numbered $\mathrm{j} = 1, \dots\dots, \mathrm{J}$ which are from the J different products. For each customer $\mathrm{i} = 1, \dots\dots, \mathrm{I}$, the customer has $n_i$ product purchases. The vector $\boldsymbol{y_i} = [y_{i1}, \dots\dots, y_{in_i}]$ denotes the purchase history of customer i, the customer i's n-th purchase is represented by $y_{in} \in \{1, \dots\dots, J\}$. The purchase histories in $\boldsymbol{Y} = [\boldsymbol{y_1}, \dots\dots, \boldsymbol{y_I}]$ are combined. Every customer has various motivations. The purpose of the model is to predict what a customer next purchase in the future based on the motivations. Before elaborating our prediction method, we firstly review the Dirichlet process (DP) and Dirichlet process mixture model (DPMM). They are the theoretical basis of our model. Then we detail the proposed purchase prediction model construction and inference procedure.

### 3.1. Dirichlet Process and Dirichlet Process Mixture Model

In essence, Dirichlet process (DP) is a famous random process utilized in no-parametric Bayesian method and is often regarded as a prior distribution in infinite mixture models [15, 26]. The metaphor of the Chinese restaurant process (CRP) can be used to describe the Dirichlet process [27] in Fig. 1. The metaphor is as follows. In a Chinese restaurant, it has an infinite number of tables. Customer 1 selects the first table to sit. The following customer either selects the same table as customer 1, or a new table. The rest of customers do the same thing. They select an occupied table with a probability. The probability is proportional to the number of customers in the occupied table. They also can select a new table with a probability proportional to the hyper-parameter $\gamma$. We adopt the metaphor for the purchase prediction environments. The customer is regarded as the Chinese restaurant, customer's purchase histories are regarded as customers, and a table represents a certain motivation.

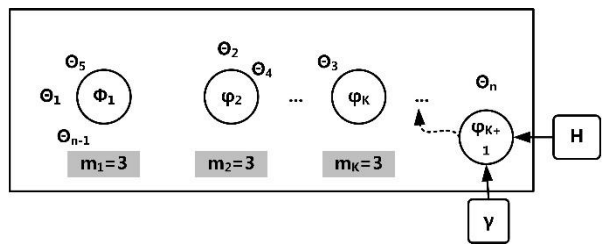



**Figure 1. Graphical model representation of a CRP**

The common application for DP is the Dirichlet process mixture model (DPMM) [15, 26]. The DP is treated as a nonparametric prior by DPMM. The number of clusters is automatically determined in DPMM. A user's purchase record can be regarded as a DPMM consisting of infinite motivations. Each product can be allocated to a certain motivation. Considering $y_{in}$ to be the n-th product of the customer i and

$$G \sim \mathrm{DP}(\gamma, \mathrm{H}) \tag{1}$$
$$\theta_{in} | G \sim \mathrm{G} \tag{2}$$
$$y_{in} | \theta_{in} \sim \mathrm{f}(\theta_{in}) \tag{3}$$

where $\mathrm{f}(\theta_{in})$ represents the distribution of products $y_{in}$ given, and the parameters of f are $\theta_{in}$. $G$ is distributed according to a DP with concentration parameter $\gamma$ and base probability measure H. DPMM is referred to as a DP mixture model, it is shown in Fig. 2.

This CRP process is stated as Eq. (4) and shown in Fig. 1.

$$\theta_{in}|\theta_{i1},\dots\dots,\theta_{i,n-1},\gamma,\mathrm{H}\sim\sum_{k=1}^{K}\frac{m_k}{n-1+\gamma}\delta_{\varphi_k}+\frac{\gamma}{n-1+\gamma}H \quad (4)$$

where the number of motivations is K, the parameter of $y_{in}$ is $\theta_{in}$, the number of the products which belong to motivation $\varphi_k$ is $m_k$.

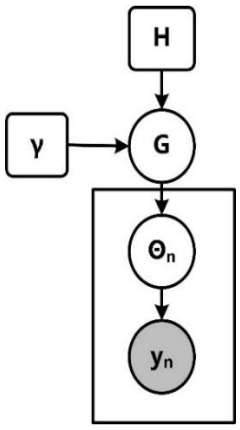



**Figure 2. Graphical model representation of the DP mixture model**

In DPMM, each $\theta_{in}$ gets a value from a motivation $\varphi_k$, and each product $y_{in}$ belongs to one of these motivations. When we identify the motivation to which product $y_{in}$ belongs, applying Bayes' theorem calculate the posterior:

$$\mathrm{p}\left(\theta_{in}|\theta_{i1},\dots\dots,\theta_{i,n-1},y_{in}\right)\propto p(y_{in}|\theta_{in})p(\theta_{in}|\theta_{i1},\dots\dots,\theta_{i,n-1}) \quad (5)$$

and the prior $p(\theta_{in}|\theta_{i1},\dots\dots,\theta_{i,n-1})$ can be obtained by Eq. (6).

$$p\left(\theta_{in}|\theta_{i1},\dots\dots,\theta_{i,n-1}\right)=\sum_{k=1}^{K}\frac{m_k}{n-1+\gamma}\delta_{\varphi_k}+\frac{\gamma}{n-1+\gamma}H \quad (6)$$

where K is the current number of motivations. Notably, we must make the distribution f and $H$ conjugated.

## 3.2. Hierarchical Dirichlet Process and Inference

A hierarchical Dirichlet process mixture model is a supplement for Dirichlet process. It is an approach to model customers of data and the relationship among these customers, each customer is associated with its own mixture model. Due to the motivation is overlapped in different users, the HDP is utilized to establish the purchase prediction model. HDP describes the relationship among different customers by shared motivations. The global probability measure $G_0$ is distributed as a Dirichlet process with concentration parameter $\gamma$ and base probability measure H. HDP also describes a set of local distribution $G_i$ which is given by a Dirichlet process with probability measure $G_0$ and a concentration parameter α. Each $G_i$ represents a customer. HDP can be simply denoted as

$$G_0\sim\mathrm{DP}(\gamma,\mathrm{H}) \quad (7)$$
$$G_i\sim\mathrm{DP}(\alpha,G_0) \quad (8)$$
$$\theta_{in}|G_i\sim G_i \quad (9)$$
$$y_{in}|\theta_{in}\sim\mathrm{f}(\theta_{in}) \quad (10)$$

Fig. 3 represents the graphical model of the HDP. The HDP can be constructed using the metaphor of the Chinese restaurant franchise [15]. The Chinese Restaurant Franchise (CRF) is the predictive process for a hierarchical partitioning of grouped data. It is a generalization of the Chinese Restaurant Process. The CRF can specify a nonparametric distribution: each customer of data is a draw from a mixture model, where the mixture motivations are shared among different customers. The local layer of the model is consisted of some DPMMs, each of them is made using the products of a certain user. Different from the traditional DPMM, the DPMM in HDP can select motivations from the higher layer. The higher layer refers to a global set of motivations. Therefore, the motivation can be shared with everyone. We relate the overview of the CRF to the purchase prediction problem. Considering the parameter $\theta_{in}$ of $y_{in}$ and obeying the following equation:

$$\theta_{in}|\theta_{i1},\dots\dots,\theta_{i,n-1},\alpha,G_0\sim\sum_{t=1}^{m_{i^*}}\frac{n_{it^*}}{n-1+\alpha}\delta_{\psi_{it}}+\frac{\alpha}{n-1+\alpha}G_0 \quad (11)$$

where $\psi_{it}=\varphi_k$ denotes user $i$'s medium t belonging to motivation k, $m_{i^*}$ denotes the number of mediums, and $n_{it^*}$ denotes the number of products belonging to medium t in user $i$.

Notably, each $\psi_{it}$ is related with one motivation $\varphi_k$, the conditional probability of the medium t in user $i$ being allocated to the motivation can be written as

$$\psi_{it}|\psi_{11},\psi_{12},\dots,\psi_{21},\psi_{22},\dots,\psi_{i,t-1},\gamma,\mathrm{H}\sim\sum_{k=1}^{K}\frac{m_{*k}}{m_{**}-1+\gamma}\delta_{\varphi_k}+\frac{\gamma}{m_{**}-1+\gamma}H \quad (12)$$

where $m_{*k}$ represents the number of mediums that are contained by motivation $\varphi_k$, and $m_{**}$ denotes the number of mediums. Fig. 4 shows the metaphor of the Chinese restaurant franchise. According to Eq. (12), the probability that a product belongs to motivations in

medium t is proportional to $n_{it^*}$. More products in medium t, the probability that a new product selecting the medium t increases. Similarly, the probability that a medium chooses motivation $\varphi_k$ is proportional to $m_{*k}$.

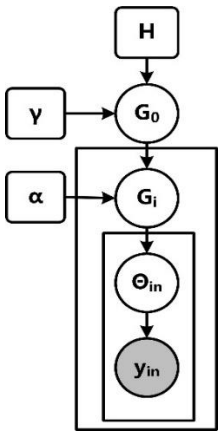



**Figure 3. Graphical model representation of the HDP model**

The goal of HDP is to find parameter$\theta_{in}$. The $\theta_{in}$ denotes the motivation $\varphi_k$ to which product $y_{in}$ is allocated. We make H has density h(·). The likelihood $p(y_{in}|\theta_{in}) = f_k^{-y_{in}}(y_{in})$ represents the conditional density of $y_{in}$ belonging to mixture motivation k. $f_k^{-y_{in}}(y_{in})$ that belonging to motivation k except $y_{in}$ is given by Eq. (13)

$$f_k^{-y_{in}}(y_{in} = j) = \frac{\int f(y_{in}|\varphi_k)\prod_{i'n' \neq in, z_{i'n'}=k} f\left(y_{i'n'}\middle|\varphi_k\right)h(\varphi_k)d\varphi_k}{\int \prod_{i'n' \neq in, z_{i'n'}=k} f(y_{i'n'}|\varphi_k)h(\varphi_k)d\varphi_k} \tag{13}$$

$$= \begin{cases} \frac{N_{k,j}+\beta}{N_{k,*}+J\beta} & if\ k\ exists \\ \frac{1}{J} & if\ k\ is\ new \end{cases} \tag{14}$$

The meaning of Eq. (13) is straightforward. The denominator represents the summation of probabilities apart from the product $y_{in}$ belongs to motivation $\varphi_k$. The numerator represents the totality of probabilities after product $y_{in}$ is allocated. Since f is conjugate to the base probability measure H, the mixture motivation parameter $\varphi_k$ is integrated to yield the likelihood. Where $N_{k,j}$ represents the number of product type j allocated to motivation k, $N_{k,*}$ is the entire number of products that belong to motivation k in all users.

Instead of calculating $\theta_{in}$ and $\psi_{it}$ directly, we compute probabilities of index variables $t_{in}$ and $k_{it}$. In general, $\theta_{in}$ and $\psi_{it}$ can be reconstructed from the related variables and the $\varphi_k$. This representation enables the Markov chain Monte Carlo sampling procedure more efficient [28]. Notice that the $\theta_{in}$ and the $\psi_{it}$ exchangeability properties are inherited by the $t_{in}$ and the $k_{it}$; the conditional distribution in (11) and (12) can be expressed by $t_{in}$ and $k_{it}$. The state space is composed of values of **t** and **k**. The number of $k_{it}$ is not fixed which is represented explicitly by the algorithm. We can think of the actual state space that is composed of an infinite number of $k_{it}$.

Sampling **t**. Based on the remainder of the variables, we utilize exchangeability to compute the conditional distribution of $t_{in}$. For computation, we treat $t_{in}$ as the last variable in (11) and (12). To compute the conditional posterior for $t_{in}$, we combine the conditional prior distribution for $t_{in}$ with the likelihood of $y_{in}$.

Using (11), the prior probability is proportional to $n_{it^*}^{-in}$ when $t_{in}$ is taking on a used medium t, while the probability is proportional to α when taking on a new medium. Due to $y_{in}$, the likelihood is $f_k^{-y_{in}}(y_{in})$ which is given $t_{in} = t$ for the previously used t. For $t_{in} = \mathrm{t}^{new}$, the likelihood can be computed by integrating out the possible values of $k_{it^{new}}$ using (12):

$$p(y_{in}|\boldsymbol{t}^{-in}, t_{in} = t^{new}, \boldsymbol{k}) = \sum_{k=1}^{K} \frac{m_{*k}}{m_{**}-1+\gamma} f_k^{-y_{in}}(y_{in}) + \frac{\gamma}{m_{**}-1+\gamma} f_{\mathrm{k}^{new}}^{-y_{in}}(y_{in}) \tag{15}$$

where $f_{\mathrm{k}^{new}}^{-y_{in}}(y_{in}) = \int f(y_{in}|\varphi)h(\varphi)d\varphi$ is simply the prior density of $y_{in}$. The conditional distribution of $t_{in}$ is then

$$p(t_{in} = t|\boldsymbol{t}^{-in}, \boldsymbol{k}) \propto \begin{cases} n_{it^*}^{-in} f_k^{-y_{in}}(y_{in}) & if\ t\ previously\ used \\ \alpha p(y_{in}|\boldsymbol{t}^{-in}, t_{in} = t^{new}, \boldsymbol{k}) & if\ t = t^{new} \end{cases} \tag{16}$$

If the sampled value of $t_{in}$ is $t^{new}$, we obtain a sample of $k_{it^{new}}$ by sampling from (15):

$$p\left(k_{it^{new}} = k|\boldsymbol{k}^{-it^{new}}, \boldsymbol{t}\right) \propto \begin{cases} m_{*k} f_k^{-y_{in}}(y_{in}) & if\ k\ previously\ used \\ \gamma f_{\mathrm{k}^{new}}^{-y_{in}}(y_{in}) & if\ k = k^{new} \end{cases} \tag{17}$$

If updating $t_{in}$, the probability will be zero that some medium $t$ will be unoccupied in the future. Because $t_{in}$ is proportional to $n_{it^*}$. Thus, we can delete the $k_{it}$. If deleting $k_{it}$, there are some components k becomes unassigned, then, we will delete this mixture motivation.

Sampling **k**. Since $k_{it}$ is related to all variables that connected by medium $t$, changing $k_{it}$ results in changes of the motivation membership of all data. Thus,

the conditional probability of $k_{it}$ is given by $f_k^{-y_{in}}(y_{in})$ ,the specific formulation is as follows:

$$p(k_{it^{new}} = k|\boldsymbol{k}^{-it}, \boldsymbol{t}) \propto \begin{cases} m_{*k}^{-it} f_k^{-y_{in}}(y_{in}) & if\ k\ previously\ used \\ \gamma f_{k^{new}}^{-y_{in}}(y_{in}) & if\ k = k^{new} \end{cases} \quad (18)$$

To predict user next purchase, it is required to compute predictive distributions for each customer. In this paper, the predictive distribution is conditioned with the whole HDP model. In specific, the predictive distribution for next purchase of the customer i is conditioned by model parameters $\alpha, \beta, \gamma, \mathbf{t}, \mathbf{k}$ . The formulations are as follows:

$$\begin{aligned} & p(\widetilde{y_{in}} = j|\boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{k}, \boldsymbol{Y}) \\ & = \sum_{t=1}^{m_{i^*}} p(t_{in} = t|\boldsymbol{t}^{-in}, \boldsymbol{k}, \boldsymbol{\alpha})\, p(\widetilde{y_{in}} = j|\boldsymbol{t}^{-in}, t_{in} = t, \psi_{it} = k, \beta) \\ & + p\big(t_{in} = t^{new}|\boldsymbol{t}^{-in}, \boldsymbol{k}, \boldsymbol{\alpha}\big) * \\ & (\sum_{k=1}^{K} p\big(\widetilde{y_{in}} = j|\boldsymbol{t}^{-in}, t_{in} = t^{new}, \psi_{it^{new}} = k, \beta, \gamma\big) + \\ & p\big(k_{it^{new}} = k^{new}|\boldsymbol{k}^{-it^{new}}, \boldsymbol{t}^{-in}, t_{in} = t^{new}, \gamma\big) * \\ & p\big(\widetilde{y_{in}} = j|\boldsymbol{t}^{-in}, t_{in} = t^{new}, \psi_{it^{new}} = k^{new}\big)) \\ & = \sum_{t=1}^{m_{i^*}} \frac{n_{it^*}}{n-1+\alpha} \frac{N_{\psi_{it}=k,j} + \beta}{N_{\psi_{it}=k,*} + J\beta} + \\ & \frac{\alpha}{n-1+\alpha} (\textstyle\sum_{k=1}^{K} \frac{m_{*k}}{m_{**}-1+\gamma} \frac{N_{k,j}+\beta}{N_{k,*}+J\beta} + \frac{\gamma}{m_{**}-1+\gamma} * \frac{1}{J}) \end{aligned} \quad (19)$$

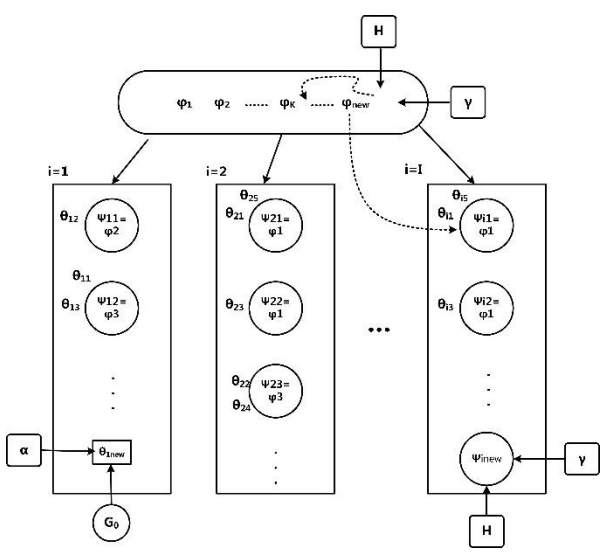



**Figure 4. The graphical model of Chinese restaurant franchise. Top rectangle box is the global shared menu, others are the restaurants**

# 4. Experimental Results

In this section, we apply the prediction method in real-world data from amazon.com, which is one of the biggest successful stores on the Internet. We firstly introduce the data and elaborate baseline models for comparison. Then the evaluation method of precision is introduced. Finally, we summarize the experimental results.

## 4.1. Data

Due to the large amount of data from amazon.com, we choose only one type data randomly. The data is movies and TV as experimental data. Initially, the data contains 123960 unique user IDs and 50050 unique product IDs. We remove some products and randomly select 10,000 users as the experimental data. After data preprocessing, the data contains 205606 product purchases of 2805 unique products which is generated by 10000 distinct customers.

Purchase data is split into two parts for evaluations: 80% of them are used as training data and the rest 20% are the test data. In our method, the number of motivations is 12 when we use the Markov chain Monte Carlo sampling scheme.

## 4.2. Baseline Models for Comparison

We present two benchmark methods to which we will compare the predictive performance of our proposed method. The first benchmark method is LDA and the second one is the collaborative filtering algorithm.

(1) LDA: In this model, alpha is set to 50/K and beta is set to 0.01 for all experiments. The predictive distribution for a new purchase $\widetilde{y_{in}}$ can be shown to equal[29]

$$\begin{aligned} & p(\widetilde{y_{in}} = j|\alpha, \beta, \boldsymbol{k}, \boldsymbol{Y}) \\ & = \sum_{k=1}^{K} p(\widetilde{y_{in}} = j|\widetilde{z_{in}} = k, \beta, \boldsymbol{k}, \boldsymbol{Y})\, p(\widetilde{z_{in}} = k|\alpha, \boldsymbol{k}^{-in}, \boldsymbol{Y}) \\ & = \textstyle\sum_{k=1}^{K} \theta_{ik}\varphi_{kj} = \sum_{k=1}^{K} \frac{\alpha + c_{ik}}{\sum_{k'}^{K} c_{ik'} + I\alpha} * \frac{N_{k,j}+\beta}{N_{k,*}+J\beta} \end{aligned} \quad (20)$$

where $c_{ik}$ is the number of purchases result from motivation k that is made by user i.

(2) Collaborative filtering: Collaborative filtering is one of the most famous algorithm that used in purchase prediction. The methods rely on co-occurrence of products purchased by users. In this experimental setting, we set the number of the neighbors to 1 and the algorithm is denoted by CF.

## 4.3. Evaluation Method

Similar to Cassar's work [30], we use precision and recall to evaluate the performance for all methods. The mathematical formulations of the two indicators are as follows:

$$\text{precision} = \frac{|A \cap B|}{|B|} * 100\% \quad (21)$$

$$\text{recall} = \frac{|A \cap B|}{|A|} * 100\% \quad (22)$$

where A represents the products provided by the purchasing predict method and the number is set to vary from $\{1,5,10,15,20,50\}$; B is the set of relevant products provided by the test data.

### 4.4. Results of Prediction

In this part we report on the predictive performance of the methods considered in this paper. Before showing the results, we firstly determine the number of the K for LDA based on heuristic algorithm. To find the optimal value, we set K from 1 to 16 and compare the hit number of the different settings. The hit number is the number of $|A \cap B|$ when $|A|$ is set to 50. The results are shown in Fig. 5.

We train LDA by using training data, then we predict customer's next purchase based on the results of training. After that, we compare the prediction set with the test data. We use the hit number as the evaluation. The result can be seen from Fig. 5, as the K increases, the hit number also increases. However, when K is larger than 13, the hit number becomes stable. We set K to 13 as the final parameter. In the spirit of our K selection criterion for LDA, we instead select the smallest value of K that corresponds to a local maximum in the range of the value. The number of motivations in LDA is close to the number of motivations in HDP which automatically determine the amount of user purchase motivations.

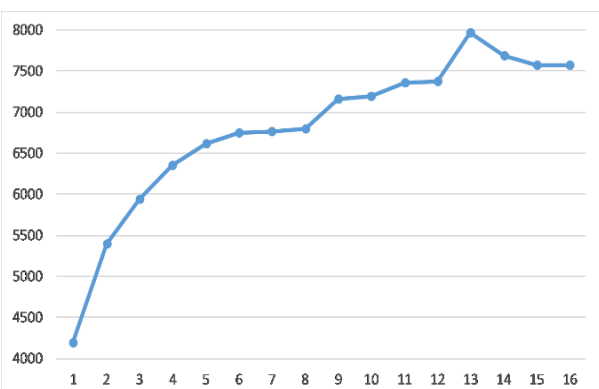



**Figure 5. Predictive performance for the sample test data with different values of K**

To assess the predictive performance by the proposed method, we evaluate its precision and recall for the test data, see (21) and (22). For precision evaluation, the higher precision value, the more relevant items that the methods returned, and vice versa. Fig. 6 shows the precision results for each method, obtained across all customers in the test data. The following conclusions can be drawn from Fig. 6:

(1) Comparisons of predictive performance: HDP consistently displays obvious advantages under varying number of predictions set and achieves the highest precision, with the highest precision reaching 0.15. The predictive method utilized in the study features desirable predictive accuracy.

(2) Comparisons of the length of the prediction set size: the precision of the HDP reaches the peak when the prediction set size total to 50; the precision of the predictive algorithms all reaches the peak when the prediction set size come to 50. This shows that the accuracy of the HDP in the study improves as the length of the prediction set size increases.

(3) The precision of CF is higher than that of LDA when prediction set size from 5 to 20. When prediction set is 5, our method is slightly stronger, with a precision close to that of CF.

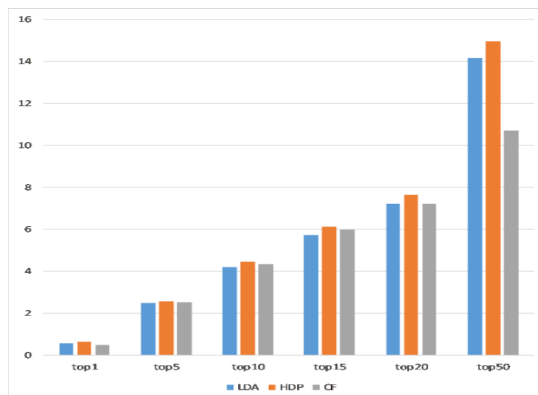



**Figure 6. Precision in the test data**

For recall evaluation, if the recall value is high, it means that most of the returned items of the method is relevant. Low recall means that the returned items are most of irrelevant. Fig. 7 presents the recall for each method, obtained across all customers in the hold-out data. The following conclusions can be drawn from Fig. 7:

(1) Comparisons of predictive performance: HDP consistently displays obvious advantages under varying number of predictions set and achieves the highest recall, with the highest recall reaching 3.7%. The predictive method utilized in the study features desirable predictive accuracy.

(2) Comparisons of the length of the prediction set size: the recall of the HDP reaches the peak when the prediction set size total to 1; the recall of the predictive algorithms all reaches the peak when the prediction set size come to 1. This shows that the accuracy of the HDP in the study improves as the length of the prediction set size decreases.

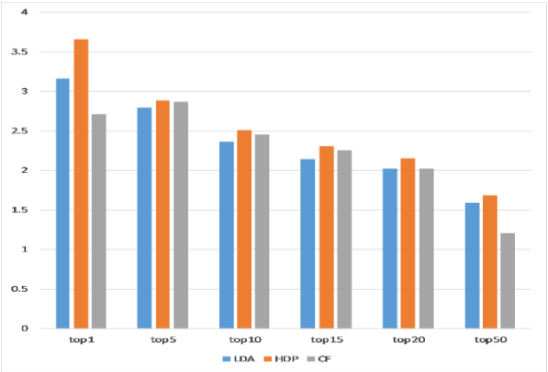



**Figure 7. Recall in the test data**

(3) The recall of CF is higher than that of LDA when prediction set size from 5 to 20. When prediction set is 5, our method is slightly stronger, with a recall close to that of CF.

## 5. Conclusions and Future Work

An accurate prediction of what a customer will purchase is of paramount importance. This paper explored to predict user purchase behavior based on purchase history by using HDP mixture model. Unlike the conventional LDA and CF methods, this method links different customers by using the property of sharing motivations in the HDP and automatically determine the amount of user purchase motivations. Therefore, the proposed model provides a better predictive performance between the real-world dataset from amazon.com. Furthermore, the experiments show the HDP outperforms other traditional methods and improves performance.

The proposed method is only one of the directions that mine hidden purchase motivation and many other study directions can be explored by other ways. One possibility is to study new methods of giving each customer a weighting coefficient. Another extension of the proposed model is to identify and analyze the meaning of the motivations which can help making strategic decisions on the manufacturing processes.

## 6. Acknowledgements

The authors would like to thank the mini-track chair and the anonymous reviewers for their insightful comments and suggestions which have helped in improving the quality of the paper. This work is supported by the Major Program of the National Natural Science Foundation of China (71490725), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (71521001), the National Natural Science Foundation of China (71722010, 91546114, 91746302, 71501057), The National Key Research and Development Program of China (2017YFB0803303).

## 7. References

[1] B.J. Jacobs, B. Donkers, and D. Fok, "Model based purchase predictions for large assortments", Marketing Science, 2016, pp. 389-404.

[2] A. Martínez, C. Schmuck, et al., "A Machine Learning Framework for Customer Purchase Prediction in the Non-Contractual Setting", European Journal of Operational Research, 2018.

[3] A. Sahni, "Incorporating perceptions of financial control in purchase prediction: An empirical examination of the theory of planned behavior", ACR North American Advances, 1994.

[4] E. Kim, W. Kim, and Y. Lee, "Combination of multiple classifiers for the customer's purchase behavior prediction", Decision Support Systems, 2003, pp. 167-175.

[5] Y. Zhang, et al., "Large Scale Purchase Prediction with Historical User Actions on B2C Online Retail Platform", Computer Science, 2014.

[6] Y. Zhao, L. Yao, and Y. Zhang, "Purchase prediction using Tmall-specific features", Concurrency and Computation: Practice and Experience, 2016, pp. 3879-3894.

[7] Y. Tsuboi, A. Jatowt, and K. Tanaka, "Product Purchase Prediction Based on Time Series Data Analysis in Social Media", in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on, 2015.

[8] D. Li, et al., "A Method of Purchase Prediction Based on User Behavior Log", in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, 2015.

[9] D. Jannach, et al., "Recommender systems: an introduction", Cambridge University Press, 2010.

[10] D.R. Liu, C.H. Lai, and W.J. Lee, "A hybrid of sequential rules and collaborative filtering for product recommendation", Information Sciences, 2009, pp. 3505-3519.

[11] S. Deerwester, et al., "Indexing by latent semantic analysis", Journal of the American society for information science, 1990, pp. 391.

[12] T. Hofmann, "Probabilistic latent semantic indexing", in ACM SIGIR Forum, 2017.

[13] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation", Journal of machine Learning research, 2003, pp. 993-1022.

[14] M. Hoffman, F.R. Bach, and D.M. Blei, "Online learning for latent Dirichlet allocation", in advances in neural information processing systems, 2010.

[15] Y.W. Teh, et al., "Hierarchical Dirichlet Processes", Publications of the American Statistical Association, 2006, pp. 1566-1581.

[16] P.M. Guadagni and J.D. Little, "A logit model of brand choice calibrated on scanner data", Marketing science, 1983, pp. 203-238.

[17] D. Mcfadden, "The choice theory approach to market research", Marketing science, 1986, pp. 275-297.

[18] U. Wagner and A. Taudes, "A multivariate polya model of brand choice and purchase incidence", Marketing Science, 1986, pp. 219-244.

[19] P.S. Fader and B.G. Hardie, "Modeling consumer choice among SKUs", Journal of marketing Research, 1996, pp. 442-452.

[20] G.S. Maddala, "Limited-Dependent and Qualitative Variables in Econometrics", Cambridge England Cambridge University Press, 1983, pp. 80-81.

[21] D.W. Hosmer Jr, S. Lemeshow, and R.X. Sturdivant, "Sturdivant, Applied logistic regression", Vol. 398, 2013.

[22] L. Cappellari and S.P. Jenkins, "Multivariate probit regression using simulated maximum likelihood", The Stata Journal, 2003, pp. 278-294.

[23] Z. Fu and L.F. Zhou, "A Purchase Prediction Based on Collaborative Filtering Algorithm", in Advanced Materials Research, Trans Tech Publ, 2014.

[24] B. Sarwar, et al., "Item-based collaborative filtering recommendation algorithms", in International Conference on World Wide Web, 2001.

[25] A.L. Deng, Y.Y. Zhu, and B.L. Shi, "A Collaborative Filtering Recommendation Algorithm Based on Item Rating Prediction", Journal of Software, 2003, pp. 54-65.

[26] H. Föllmer, "Dirichlet processes", Springer Berlin Heidelberg, 1981, pp. 476-478.

[27] D.J. Aldous, "Exchangeability and related topics", Springer Berlin Heidelberg, 1985, pp. 1-198.

[28] R.M. Neal, "Markov chain sampling methods for Dirichlet process mixture models", Journal of computational and graphical statistics, 2000, pp. 249-265.

[29] B. Jacobs, B. Donkers, and D. Fok, "Model-based Purchase Predictions for Large Assortments", Erim Report, 2016.

[30] G. Cassar, et al., "A hybrid semantic matchmaker for IoT services", in Green Computing and Communications (GreenCom), 2012 IEEE International Conference on, 2012.