Learning to Rank Audience for Behavioral Targeting in Display Ads

Jian Tang[1*], Ning Liu[2], Jun Yan[2], Yelong Shen[3*], Shaodan Guo[4*],
Bin Gao[2], Shuicheng Yan[5], and Ming Zhang[1]

[1]School of EECS, Peking University, Beijing, China, {tangjian, mzhang}@net.pku.edu.cn

[2]Microsoft Research Asia, Beijing, China, {ningl, junyan, bingao}@microsoft.com

[3]Beihan University, Beijing, China, shengyelong@gmail.com

[4]Huazhong University of Science and Technology, Wuhan, China, guoshaodan@gmail.com

[5]National University of Singapore, Singapore, eleyans@nus.edu.sg

ABSTRACT

Behavioral targeting (BT), which aims to sell advertisers those behaviorally related user segments to deliver their advertisements, is facing a bottleneck in serving the rapid growth of long tail advertisers. Due to the small business nature of the tail advertisers, they generally expect to accurately reach a small group of audience, which is hard to be satisfied by classical BT solutions with large size user segments. In this paper, we propose a novel probabilistic generative model named Rank Latent Dirichlet Allocation (RANKLDA) to rank audience according to their ads click probabilities for the long tail advertisers to deliver their ads. Based on the basic assumption that *users who clicked the same group of ads will have a higher probability of sharing similar latent search topical interests*, RANKLDA combines topic discovery from users' search behaviors and learning to rank users from their ads click behaviors together. In computation, the topic learning could be enhanced by the supervised information of the rank learning and simultaneously, the rank learning could be better optimized by considering the discovered topics as features. This co-optimization scheme enhances each other iteratively. Experiments over the real click-through log of display ads in a public ad network show that the proposed RANKLDA model can effectively rank the audience for the tail advertisers.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – retrieval models

General Terms: Algorithms, Economics, Experimentation

Keywords

User ranking, behavioral targeting, display ads

* This work was done when the first, fourth and fifth author were visiting Microsoft Research Asia.

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

1. INTRODUCTION

Behavioral targeting (BT) [4, 5, 9, 11] is one of the mainstream forms in online advertising. It aims to deliver the right advertisements to the right audience according to their online behaviors such as searching and browsing histories. Display advertising [12], which is one of the major ads delivery channels nowadays, generally appears on the Internet in the form of graphical ads, including text, images, logos, flashes, videos, etc. Due to the well-known semantic gap and the limited textual information involved in display ads, behavioral targeting is playing an important role in targeting the display ads to the right audience.

The classical behavioral targeting solutions [11] pre-divide users into user segments, i.e. user categories, within each of which users are assumed to share similar interests. Each segment is labeled with keywords to indicate the interests of the users in the segment. For example, a segment labeled with "car buyer" indicates the users in this segment may want to buy cars. For advertisers, they buy one or more user segments for their ads delivery based on the keyword descriptions of the user segments. Recently, with the rapid growth of online Customer to Customer (C2C) business, there is an emerging market of the long tail advertisers. These long tail advertisers have a common characteristic: they have limited budget for their ads delivery. However, even a single user segment produced by classical behavioral targeting solutions, which has little consideration to the long tail advertisers, may have a large number of users and hence it is generally too large for these advertisers, who cannot afford to reach such a large number of targeted users. Thus if only by adopting classical behavioral targeting solutions, the long tail advertisers, who consist a huge and underexplored market in display ads, cannot be satisfied. This motivates us to study the problem of user ranking (audience ranking) according to their click probabilities over the given group of ads. Then those long tail advertisers may select arbitrary number of top ranked users according to their budget for ads delivery.

The classical ranking algorithms, which calculate the relevance between display ads and users and then ranking according to the relevance score, are not applicable for the user ranking problem since the display ads contain little textual information. Instead, we propose to learn to rank users from their ads click behaviors, and it only relies on user features. Specifically, we proposed a novel generative model RankLDA for learning to rank users according

to their ads click probabilities from their ads click behaviors. Based on the basic assumption that *users who clicked the same group of ads will have a higher probability of sharing similar latent search topical interests,* the RankLDA model combines the topic discovery from users' search behaviors and learning to rank users from their ads click behaviors together. By combining the two procedures together, on one hand, the topic discovery can benefit from the supervised information of rank learning, namely users' ads click behaviors; on the other hand, the discovered topics can serve as good features in the process of learning to rank users. Thus the two processes are mutually enhanced. Experiments with a real word search engine log with corresponding ads click through demonstrate the superiority of the RANKLDA model over the baselines both in audience ranking and topic discovery. Furthermore, by comparing the weights of the topics, we can understand what behaviors of users tend to lead to ads clicks in the target domain.

2. PRELIMINARIES&RELATED WORK

2.1 Behavioral Targeting in Display Ads

Behavioral targeting [4, 5, 9, 11] has attracted much research attention recently in online advertising community, especially for display advertisement. It does not rely on the contextual information of Web pages for ad delivery. Instead, it enables advertisers to target the advertisements to the right audience by leveraging users' recent online activities such as their search queries in search engines, page views in some special Websites. For example, if we observe a user often searches queries about cell phone or visits websites about cell phone, then this user may want to buy a cell phone recently and we may deliver the ads about cell phone to him.

Classical behavioral targeting methods [11] generally segment users into multiple user interest segments, within each of which users are assumed to share similar interests. Each segment is labeled with keywords and corresponds to a product domain. For example, suppose there is a user segment labeled as "car buyer", then the users are grouped into this segment because they have shown behaviors related with buying a car. Then based on the keyword descriptions of the user segments, the advertisers buy one or more user segments for their ads delivery. However, even a single user segment may have a large number of users, and those long-tail small advertisers cannot afford an entire user segment. Instead, they may only want to buy a small portion of users who are most interested in their ads. Thus this results in a user ranking problem, which ranks users according to their probabilities of interest in the target ads.

2.2 Learning to Rank

In this subsection, we briefly introduce the pairwise learning to rank methods, which will be used in our proposed algorithms and compared baselines. The general pairwise ranking criterion is to minimize the expected empirical risk, which is the average loss over the whole training pairs. In order to overcome the over fitting problem, regularization is incorporated. As a result, the general criterion for pairwise ranking methods is to minimize the following objective function:

$$min_{\boldsymbol{\eta}} L(\boldsymbol{\eta}, P) + C \parallel \boldsymbol{\eta} \parallel_2^2$$

where $L(\boldsymbol{\eta}, P)$ is the empirical risk over all the preference pairs in the training data P:

$$\mathrm{L}(\boldsymbol{\eta}, P) = \frac{1}{|P|} \sum\nolimits_{(u_i, u_j) \in P} \ell\big(f(\boldsymbol{\eta}, \boldsymbol{x_i}), f(\boldsymbol{\eta}, \boldsymbol{x_j})\big)$$

$\ell(\cdot,\cdot)$ is a loss function defined on a preference pair $(\boldsymbol{x_i}, \boldsymbol{x_j})$, $f(\boldsymbol{\eta}, \boldsymbol{x_i}) = \boldsymbol{\eta}^T \boldsymbol{x_i}$ is the prediction function, and $\boldsymbol{\eta}$ is the corresponding feature weight vector. C is a parameter to tradeoff between minimizing empirical risk and finding a simple model, $\parallel \boldsymbol{\eta} \parallel_2$ is the ℓ_2-norm of vector $\boldsymbol{\eta}$.

2.3 Topic Models

Topic models such as Latent Dirichlet Allocation (LDA) [1] are used to extract latent semantic topics within a corpus. Most of the existing models built on LDA are unsupervised methods, which all rely on the co-occurrence of words, and don't rely on any supervised information. Recently, some models that combine the unsupervised process in LDA and supervised information together are proposed [2, 6]. In [2], Blei *et al.* proposed a supervised topic model (sLDA), in which each document is paired with a response. In sLDA, for each document, the words are generated in the same manner as that in the classic LDA, and then a response variable is generated based on the topics of the document. The goal of sLDA is to jointly discover the latent topics in documents and infer latent topics predictive of the response. The most similar work with RANKLDA model is the relational topic model (RTM) [3], which models the generation of documents and the links between them. However, RTM is a pure unsupervised model and focuses on link prediction while RANKLDA integrates supervised information from users' ads click behaviors and focuses on ranking, which is different from existing variants of LDA model.

3. LEANING TO RANK AUDIENCE WITH RANKLDA

In this section, we describe our model for learning to rank audience. We first formally formulate the audience ranking problem and then introduce a novel model, which is named Rank Latent Dirichlet Allocation (RANKLDA), for ranking users according to their probabilities of interest in the ads of the target domain.

3.1 Problem Formulation

The audience ranking problem aims to rank users according to their ads click probabilities. Specifically, we formally defined the problem as below:

Given a group of ads in the target domain and a collection of users that viewed one of these ads at least once, each user u is represented with a feature vector $\boldsymbol{x}$*, which is extracted from his search behaviors, including search queries and the titles of the clicked URLs. Audience ranking aims to rank users according to a ranking score* $f(\boldsymbol{x}) = \boldsymbol{\eta}^T \boldsymbol{x}$*, where* $\boldsymbol{\eta}$ *is the feature weight vector.*

In order to learn to rank audience, we collect training data from their ads click behaviors. Given a group of ads, if two users u_i and u_j both viewed the ads, but u_i clicked the ads and u_j did not click the ads, then u_i is generally more interested in this group of ads than u_j, and (u_i, u_j) is treated as a user preference pair. Additionally, we associate each user preference pair with a confidence score c_{ij}, which is used to measure the confidence that u_i is more interested in this group of ads than u_j, and is defined as:

$$c_{ij} = \frac{click_i}{Impres_i} * \sigma(Impres_j)$$

where $click_i$ and $Impres_i$ are the number of clicks and impressions over this group of ads respectively, $\sigma(.)$ is the logistic function $\sigma(x) = 1/(1 + \exp(-x))$.

Let U be the set of all users, C the set of click users, then the training data *P* can be formally defined as:

$$P = \{(u_i, u_j, c_{ij}) | u_i \in C, u_j \in U \backslash C\}$$

3.2 RANKLDA

In this subsection, we describe the proposed RANKLDA model for user ranking, which is a probabilistic generative model that jointly models topic discovery from users' search behaviors and learning to rank users from their ads click behaviors. The Notations to be used is summarized in Table 1.

Table 1. Notations to be used in RANKLDA model

Notation	Description
K	Total number of topics
M	Total number of users
α	The prior parameter for the user topic Dirichlet
$\boldsymbol{\theta_i}$	Topic distribution of user *i*
w_{in}	The n_{th}word in user *i*'s profile
z_{in}	The topic of the n_{th}word in user *i*'s profile
N_i	The total number of words in user *i*'s profile
$\boldsymbol{w}_i$	The words of user *i*'s profile
$\boldsymbol{z}_i$	The topics of user *i*'s profile
$\{\boldsymbol{\beta_k}\}$	The topics
$\boldsymbol{\eta}$	The topic weight vector
y_{ij}	Response variable between user *i* and *j*
P	The user preference pair set

3.2.1 Definition of RANKLDA

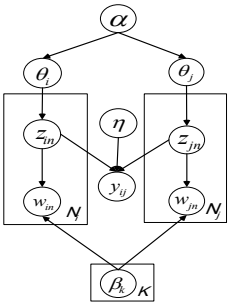

Figure 1. Graphical representation of RANKLDA

Our RANKLDA model combines topic discovery from users' search behaviors and learning to rank users from their ads click behaviors together. The intuition that combines the two processes together is that users who both clicked the same group of ads have a higher probability of sharing similar search topical interests, thus the topic discovery from users' search behaviors can benefit from their ads click behaviors; meanwhile, the learning of user ranking can also be better optimized by using the discovered topics as users' features. Specifically, the model consists of two parts: the modeling of topic discovery from users' search behaviors and the learning to rank users from their ads click behaviors. For the topic discovery, each user *i* with search profile $\boldsymbol{w}_i$ is generated as the manner in the classic LDA model [1]; for the user ranking, a response variable $y_{ij} = 1$ is generated for each user preference pair (i, j), i.e. $y_{ij} = 1$ if user *i* clicked the ads in the domain and user *j* did not. The graphical representation of RANKLDA is presented in Figure 1, and the detailed generative process is summarized as:

(1) For each user *i*,

a. Draw a topic proportion $\boldsymbol{\theta_i} \sim Dirichlet(\alpha)$,

b. For each word w_{in} in user *i*'s profile, n = $1,2,\cdots,N_i$,

Draw a topic assignment $z_{in} \sim Multinomial(\boldsymbol{\theta_i})$,

Draw a word $w_{in} \sim Multinomial(\boldsymbol{\beta}_{z_{in}})$,

(2) For each user preference pair with corresponding weight $(i, j, c_{ij}) \in P$, draw a response variable $y_{ij} = 1|\bar{\boldsymbol{z}}_i, \bar{\boldsymbol{z}}_j, \boldsymbol{\eta} \sim \text{Bernoulli}\left(\sigma(\rho_{ij})^{c_{ij}}\right)$.

where $\bar{\boldsymbol{z}}_i = 1/N_i \sum_{n=1}^{N_i} z_{in}$, $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic function ,and $\rho_{ij} = \boldsymbol{\eta}^T \bar{\boldsymbol{z}}_i - \boldsymbol{\eta}^T \bar{\boldsymbol{z}}_j$.

Given model parameters $\Theta = \{\alpha, \{\boldsymbol{\beta}_k\}_{k=1}^K, \boldsymbol{\eta}\}$, the joint probability of the observed and hidden variables is:

$$\begin{aligned} &P\left(\{\boldsymbol{w}_i, \boldsymbol{z}_i, \boldsymbol{\theta}_i\}_{i=1}^M, \{y_{ij}\}_{(i,j,c_{ij})\in P} | \Theta\right) \\ &= \prod_{i=1}^M P(\boldsymbol{\theta_i}|\alpha) \prod_{n=1}^{N_i} P(z_{in}|\boldsymbol{\theta_i}) P\left(w_{in}|\boldsymbol{\beta}_{z_{in}}\right) \\ &\times \prod_{(i,j,c_{ij})\in P} P(y_{ij} = 1|\bar{\boldsymbol{z}}_i, \bar{\boldsymbol{z}}_j, \boldsymbol{\eta})^{c_{ij}} \end{aligned} \quad (1)$$

3.2.2 Model Learning

The learning process for RANKLDA is to maximize the log likelihood of the observed variables $\log P\left(\{\boldsymbol{w}_i\}_{i=1}^M, \{y_{ij}\}_{(i,j)\in P} | \Theta\right)$. In order to maximize this objective function, we adopt Expectation Maximization (EM) algorithm, which is composed of two key steps: E-step and M-step. In the E-step, we aim to calculate the posterior distribution of the latent variables. In the M-step, we aim to maximize the expectation of the log complete likelihood over the distribution calculated in E-step.

However, as its counterpart in the conventional LDA learning [1], in the E-step, the posterior distribution of latent variables is computational intractable. Thus we resort to the variational inference method, which tries to minimize the KL-divergence between the approximated and the true posterior distribution of latent variables.

Specifically, we adopt the mean-field variational inference [8], which is generally used in the classical LDA model. The mean-field variational inference forms a factorized distribution over the latent variables, which is expressed as:

$$q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^N q(z_n|\phi_n), \quad (2)$$

where $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ are the free parameters, and $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\gamma}), z_n \sim Multinomial(\phi_n)$. So the objective is to find the optimum $\{\boldsymbol{\gamma}, \boldsymbol{\phi}\}$ such that $q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$ can be the best approximation of conditional distribution $P(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \Theta)$.

Based on the Jensen inequality [7], we can find a lower bound of the log likelihood. Besides, the difference between the log

likelihood and the lower bound is the KL divergence between the variational posterior probability and the true posterior probability, i.e.

$$\log \mathrm{P}\left(\{\boldsymbol{w}_i\}_{i=1}^{M}, \{y_{ij}\}_{(i,j)\in P}\right) = \mathrm{L}(\boldsymbol{\gamma}, \boldsymbol{\phi};\ \Theta) + D\left(\mathrm{q}(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) \middle|\middle| \mathrm{P}(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \Theta)\right), \tag{3}$$

where $\mathrm{L}(\gamma, \phi;\ \Theta)$ is the lower bound of the log likelihood and can be calculated as below:

$$\begin{aligned}\mathrm{L}(\boldsymbol{\gamma}, \boldsymbol{\phi};\ \Theta) &= E_q\left[\log\left(\{\boldsymbol{w}_i, \boldsymbol{z}_i, \boldsymbol{\theta}_i\}_{i=1}^{M}, \{y_{ij}\}_{(i,j)\in P}\middle|\Theta\right)\right] = \\ &\sum_{(i,j)\in P} E_q\left[\mathrm{c}_{ij} log P\left(y_{ij}=1\middle|\bar{\boldsymbol{z}}_i, \bar{\boldsymbol{z}}_j, \boldsymbol{\eta}\right)\right] + \sum_{i=1}^{M} E_q \log P(\boldsymbol{\theta}_i|\alpha) \\ &+ \sum_{i=1}^{M} E_q \log \mathrm{P}(\boldsymbol{z}_i|\boldsymbol{\theta}_i) + \sum_{i=1}^{M} E_q \log P\left(\boldsymbol{w}_i\middle|\boldsymbol{z}_i, \{\boldsymbol{\beta}_k\}_{k=1}^{K}\right) \\ &- \sum_{i=1}^{M} E_q \log \mathrm{q}(\boldsymbol{\theta}_i) - \sum_{i=1}^{M} E_q \log \mathrm{q}(\boldsymbol{z}_i)\end{aligned} \tag{4}$$

For the last five terms on the right-hand side of above equation, we can calculate them as the way introduced in standard LDA [1], so we only need to focus on calculating the first term, in which

$$E_q\left[\mathrm{c}_{ij} log P\left(y_{ij}=1\middle|\bar{\boldsymbol{z}}_i, \bar{\boldsymbol{z}}_j, \boldsymbol{\eta}\right)\right] = \mathrm{c}_{ij} E_q\left[log\sigma\left(\boldsymbol{\eta}^T \bar{\boldsymbol{z}}_i - \boldsymbol{\eta}^T\ \bar{\boldsymbol{z}}_j\ \right)\right] \tag{5}$$

To calculate the expectation, we adopt the method introduced in [10], which uses the variational method again with the following variational bound [8]:

$$\sigma(\mathrm{x}) \geq \sigma(\xi) \exp\left(\frac{x-\xi}{2} + \mathrm{g}(\xi)(x^2 - \xi^2)\right), \tag{6}$$

where ξ is the free variational parameter, and $\mathrm{g}(\xi) = (\frac{1}{2} - \sigma(\xi))/2\xi$. Then we have

$$\begin{aligned}E_q\left[log P\left(y_{ij}=1\middle|\bar{\boldsymbol{z}}_i, \bar{\boldsymbol{z}}_j, \boldsymbol{\eta}\right)\right] &\geq log\sigma(\xi) - \frac{\xi}{2} - \xi^2 g(\xi\ ) \\ &+ \frac{1}{2}\boldsymbol{\eta}^T E_q(\bar{\boldsymbol{z}}_i) - \frac{1}{2}\boldsymbol{\eta}^T E_q\left(\bar{\boldsymbol{z}}_j\right) + g(\xi) E_q\left({\rho_{ij}}^2\right).\end{aligned} \tag{7}$$

In order to minimize the KL-divergence between the variational distribution q and the true distribution p, we only need to maximize $\mathrm{L}(\boldsymbol{\gamma}, \boldsymbol{\phi};\ \Theta)$ with respect to $\left\{\{\xi_{ij}\}, \boldsymbol{\gamma}, \boldsymbol{\phi}\right\}$. We omit some derivations here and just list the final update equations:

$$\xi_{ij} = \left(\ E_q\left({\rho_{ij}}^2\right)\right)^{\frac{1}{2}}\ , \tag{8}$$

$$\gamma_{ik} = \alpha_k + \sum_{n=1}^{N_i} \phi_{i,n,k}\ . \tag{9}$$

For user i who clicked ads:

$$\begin{aligned}&\phi_{i,n,k} \propto \\ &\beta_{kv} \exp\left(\sum_j \mathrm{c}_{ij} \frac{1}{2N_i}\eta_k + \mathrm{c}_{ij} g\left(\xi_{ij}\right)\left(\frac{2\left(\boldsymbol{\eta}^T \sum_{s=1, s\neq n}^{N_i} \boldsymbol{\phi}_{-is}\right)\eta_k + (\boldsymbol{\eta}\circ\boldsymbol{\eta})_k}{{N_i}^2} - \right.\right. \\ &\left.\left.\frac{2(\boldsymbol{\eta}^T \sum_{t=1}^{N_j} \boldsymbol{\phi}_{jt})\eta_k}{N_i N_j}\right) + \Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right) \\ &s.t \sum_{k=1}^{K} \phi_{i,n,k} = 1\ .\end{aligned} \tag{10}$$

For user j who did not click ads:

$$\begin{aligned}&\phi_{j,n,k} \propto \\ &\beta_{kv} \exp\left(\sum_i \mathrm{c}_{ij} \frac{-1}{2N_j}\eta_k + \mathrm{c}_{ij} g\left(\xi_{ij}\right)\left(\frac{2\left(\boldsymbol{\eta}^T \sum_{s=1, s\neq n}^{N_i} \boldsymbol{\phi}_{-is}\right)\eta_k + (\boldsymbol{\eta}\circ\boldsymbol{\eta})_k}{{N_j}^2} - \right.\right. \\ &\left.\left.\frac{2(\boldsymbol{\eta}^T \sum_{t=1}^{N_i} \boldsymbol{\phi}_{jt})\eta_k}{N_i N_j}\right) + \Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right)\right), \\ &s.t \sum_{k=1}^{K} \phi_{j,n,k} = 1\ ,\end{aligned} \tag{11}$$

where $\boldsymbol{\eta} \circ \boldsymbol{\eta}$ means the element-wise product between $\boldsymbol{\eta}$ and $\boldsymbol{\eta}$.

For the learning of RANKLDA, the variational EM algorithm is adopted, the detail updating equations are summarized as below:

1. **E-step:** update the parameters $\left\{\{\xi_{ij}\}, \boldsymbol{\gamma}, \boldsymbol{\phi}\right\}$ according to equations (8), (9), (10) and (11).
2. **M-step:** update the parameters $\Theta = \{\alpha, \{\boldsymbol{\beta}_k\}_{k=1}^{K}, \boldsymbol{\eta}\}$,

Update equation for $\boldsymbol{\eta}$:

$$\begin{aligned}\boldsymbol{\eta} = -\frac{1}{4}\left(\sum_{(i,j)\in P} \mathrm{c}_{ij} g\left(\xi_{ij}\right) E_q\left[\left(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}}_j\right)\left(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}}_j\right)^T\right]\right)^{-1} \\ \times \sum_{(i,j)\in P} \mathrm{c}_{ij} \left(E_q\left(\bar{\boldsymbol{z}}_i\right) - E_q\left(\bar{\boldsymbol{z}}_j\right)\right)\end{aligned} \tag{12}$$

The updated equations for the parameters $\alpha, \{\boldsymbol{\beta}_k\}_{k=1}^{K}$ are the same as in the classic LDA [1].

3.2.3 RANKLDA for Audience Ranking

Given a set of users together with their search behaviors, how should we rank these users with the estimated model Θ? Firstly, for each user with profile $\mathbf{w}$, we need to estimate the topic proportion of the user. Thus we approximate the posterior distribution of latent variables $\mathrm{P}(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \Theta)$ with factorized distribution $\mathrm{q}(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$ and by minimizing the KL divergence between the two distributions we can estimate the optimum $\boldsymbol{\gamma}, \boldsymbol{\phi}$. Then the ranking score $\mathrm{f}(\mathbf{w})$ for user with profile $\mathbf{w}$ is calculated as:

$$\mathrm{f}(\mathbf{w}) = \boldsymbol{\eta}^T E_q(\bar{\mathbf{z}}) = \frac{1}{N}\boldsymbol{\eta}^T \sum_{n=1}^{N} \phi_n \tag{13}$$

4. EXPERIMENT

In this section, we elaborate on the experiments for audience ranking. We first introduce the dataset used, and then describe the evaluation metrics. For the performance of user ranking, we compare RANKLDA with pairwise learning to rank methods with BOW features and topics discovered by LDA respectively.

4.1 Dataset

Our dataset is an integration of two data sources. One is the query log of a commercial search engine, which records the search behaviors of search engine users. The other data source is the advertisement click-through log in corresponding ad network using the display ads as the ads delivery channel. The two datasets have overlapped user IDs for us to identify the unique users. To avoid the privacy issues, only the abstract user IDs are used and no other user information such as demographic and geographic are used. In other words, the dataset contains both users' ad click information in a display ads' ad network and users' search behaviors in corresponding search engine of the ad network. The advertisements we used for research purpose are a group of ads in the "auto insurance" domain. There are 43

different ads in this group. Our dataset is preprocessed in three steps. First, we extract all the users that viewed any ads in this group at least once range from 20th to 26th of September 2010 and extract the number of clicks and impressions for each user from display ads click through log. Second, we extract the search behaviors of these users between 13th to 19th of September 2010, including search queries and the clicked URLs, from a corresponding commercial search engine query log. We filter out users that did not have any behaviors during this period. Finally, we extract the titles of all the webpages that have been clicked by these users. Through this way we get a total number of 1,935,534 unique users with 5,294,484 ad impressions in display ads and 1,912 ad clicks. Thus the average click through rate (CTR) for this group of ads is 0.036% and the ratio of user who clicked ads is 0.090%. With respect to users' search behaviors, the number of average queries and clicked URLs for each user is 6.6 and 6.1 respectively.

4.2 Evaluation Metric

The evaluation aims to measure the effect of improving average user CTR for advertisers who only buy a small part of users that are most likely to click the ads. We evaluate the performance of the improvement of users' average CTR. Firstly, based on the ranking result, we calculate the average CTR of users who are ranked in the top 20%, denoted by CTR@20%, i.e.,

$$\text{CTR@20\%} = \frac{\sum_{u \in \{u | user\ u\ ranked\ in\ top\ 20\%\}} CTR_u}{|u \in \{u | user\ u\ ranked\ in\ top\ 20\%\}|}$$

where CTR_u is the CTR of user u, which is calculated by:

$$CTR_u = \frac{Click_u + \alpha}{Impres_u + \beta}$$

and $Click_u, Impres_u$ are the number of clicks and impressions for user u respectively, and α, β are the average number of users' clicks and impressions respectively, which are the smoothing coefficients that are defined globally for all users as done in [5]. In our dataset, $\alpha = 9.88 \times 10^{-4}$ and $\beta = 2.74$. The definition of CTR@40%, CTR@60%,..,CTR@100% can be defined similarly.

In order to calculate the CTR improvement after user ranking, we compare CTR@20% with the average CTR of the whole users, i.e., CTR@100%. We define:

$$\text{Impr@20\%} = \frac{CTR@20\% - CTR@100\%}{CTR@100\%}$$

By observing the value of Impr@20%, we can see how much average user CTR can be improved if the advertisers only buy the segment of users who are ranked in top 20%.

4.3 Baselines

The pairwise learning to ranking method introduced in Section 2.2 can be used to rank users directly with any kinds of user features. We provide two kinds of user features: "bag-of-words" (BOW) and topic discovered by classic LDA with the corresponding topic weight as feature value. We summarize our baselines as below:

(1) *BOW+Pairwise ranking*: The pairwise learning to rank method with "bag-of-words" features.

(2) *Topic+Pairwise ranking*: The pairwise learning to rank method with topics discovered by classic LDA as features (with the corresponding topic weight as feature value).

4.4 Results

4.4.1 Audience Ranking

We perform five-fold cross validation over all the users in the dataset with our RANKLDA model. Figure 2 shows the average results. We can see that after user ranking, users ranked in higher positions have higher ads CTR, which proves that user ranking is effective in ranking users according to their ads click probabilities and hence can help find better small user segments.

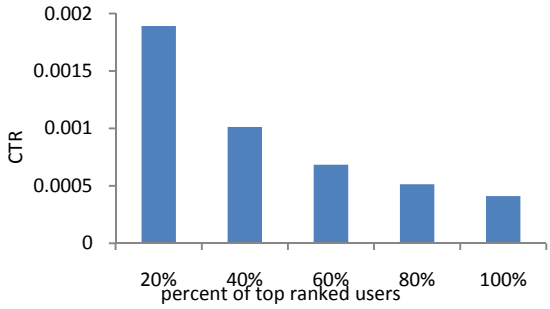

Figure 2: Average user CTR with different percent of top ranked users

Table 2 lists the result of audience ranking with Impr@20% metric. We can see that no matter how many topics we choose, the performance of RANKLDA or Topic+Pairwise ranking method is consistently better than that of BOW+Pairwise ranking. This proves that topics can serve as better features than "BOW". Furthermore, the performance of RANKLDA is much better than the *Topic+Pairwise ranking* method. As mentioned previously, RANKLDA combines the topic discovery from users' search behaviors and learning to rank users from their ads click behaviors together, and the two processes are mutually enhanced.

4.4.2 Topic Discovery

In Table 3, we compare the topics that discovered by *Topic+Pairwise ranking* and the RANKLDA model respectively. For both models, we have a weighting vector associated with the features, i.e., the topics. The weight of each topic reflects its importance in influencing users' ads click behaviors, and thus we rank the discovered topics according to their weights in descending order, and finally we choose the topics ranked in the top three positions. The discovered topics with both models in "auto insurance" domain are listed in Table 3. The topics discovered by *Topic+Pairwise ranking* are quite general, such as "car", "insurance", and "travel". On the contrary, RANKLDA integrates supervised information which are users' ads click behaviors and the topics are much more specific. For example, the topic "car insurance" is a self-contained topic discovered by RANKLDA, while the baseline method treats it as two separate topics: "car" and "insurance". The discovered topics tell us that people who pay attention to car insurance, travel, and health & kids are most likely to click the ads in "auto insurance" domain.

Table 2: The audience ranking results with Impr@20% Metric

	Impr@20%			
Model	**Topic 20**	**Topic 40**	**Topic 60**	**Topic 80**
BOW+Pairwise ranking	1.015±0.210			
Topic+Pairwise ranking	1.266±0.253	1.426±0.154	1.337±0.345	1.088±0.249
RANKLDA	**3.599±0.118**	**3.637 ± 0.176**	**3.626 ± 0.079**	**3.640±0.090**

Table 3: The discovered topics that lead to ads clicks in "auto insurance" domain

Topic+Pairwise ranking			**RANKLDA**		
#Topic : Car	#Topic : Healthy	#Topic : Travel	#Topic : CarInsurance	#Topic : Travel	#Topic : Healthy&Kids
chevy	houston	airline	car	los	health
car	hospital	vegas	insurance	city	medical
ford	insurance	travel	auto	travel	care
truck	center	southwest	america	coolest	child
wheel	medical	vaction	banking	town	houston
bmw	health	hotel	account	iowa	story
racing	care	flight	bmw	texas	toy
auto	saint	cheap	drive	lake	boy

5. CONCLUSIONS

In this paper, we investigated the problem of learning to rank audience for behavioral targeting in display ads. By ranking audience according to their probabilities of interest over the given ads, those small long tail advertisers can select to buy an appropriate portion of top ranked users for their ads delivery. We proposed a novel audience rank model referred to as Rank Latent Dirichlet Allocation (RANKLDA). RANKLDA well integrates the process of topic discovery from users' search behaviors and also learning to rank audience from their ads click behaviors. By the joint process, the two tasks are mutually boosted. We performed a large scale evaluation study based on a commercial search engine log and also the corresponding display ads click through log. The experimental results well demonstrated that the RANKLDA model outperforms the baselines both in audience ranking and topic discovery.

6. REFERENCES

[1] D. Blei and J. Lafferty. Latent dirichlet allocation. In *Journal of Machine Learning Research,* pages 993-1022 , 2003.

[2] D. Blei and J. Lafferty. Supervised topic models. In *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems,* 2007.

[3] J.Chang and D. Blei. Relation topic models for document networks. *Artificial Intelligence and Statistics,* 2009.

[4] T. Chen, J. Yan, G.Xue, and Z. Cheng. Transfer learning for behavioral targeting. In *Proceedings of the 19th International Conference on World Wide Web,* pages 1077-1078, 2010.

[5] Y. Chen, D. Pavlov, and J. F.Canny. Large-Scale Behavioral Targeting. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 209-218, 2009.

[6] X. Gu, S. Yang, and H. Li. Named entity mining from click-through data using weakly supervised latent dirichlet allocation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 267-274, 2009.

[7] M. Jordan, editor. Learning in Graphical Models. *MIT Press,* Cambridge, MA, 1999.

[8] T. Jaakkola. Variational methods for inference and estimation in graphical methods. *PhD thesis,* MIT. 1997.

[9] N. Liu, J. Yan, D. Shen, D. Chen, Z. Chen, and Y. Li. Learning to rank audience for behavioral targeting. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719-720, 2010.

[10] Y. Liu, A. Niculescu-Mizil, and W. Grys. Topic-Link LDA: joint models for topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning,* 2009.

[11] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web,* pages 261-270, 2009.

[12] Wikipedia. http://en.wikipedia.org/wiki/Display_advertising