Combining Geographical Information of Users and Content of Items for Accurate Rating Prediction

Zhi Qiao
Institute of Computing Technology, Chinese Academy of Science
University of Chinese Academy of Sciences
Beijing, China
zhiqiao.ict@gmail.com

Peng Zhang
Institute of Information Engineering, Chinese Academy of Science
Beijing, China
zhangpeng@iie.ac.cn

Jing He
Victoria University
Melbourne, Australia
jing.he@vu.edu.au

Yanan Cao
Institute of Information Engineering, Chinese Academy of Science
Beijing, China
caoyanan@iie.ac.cn

Chuan Zhou
Institute of Information Engineering, Chinese Academy of Science
Beijing, China
zhouchuan@iie.ac.cn

Li Guo
Institute of Information Engineering, Chinese Academy of Science
Beijing, China
guoli@iie.ac.cn

ABSTRACT

Recommender systems have attracted attentions lately due to their wide and successful applications in online advertising. In this paper, we propose a bayesian generative model to describe the generative process of rating, which combines geographical information of users and content of items. The generative model consists of two interacting LDA models, where one LDA model for location-based user groups (user dimension) and the other for the topics of content of items(item dimension). A Gibbs sampling algorithm is proposed for parameter estimation. Experiments have shown our proposed method outperforms baseline methods.

Categories and Subject Descriptors H.2.8 [Database Management]: Database Applications - Data Mining
General Terms Theory, Algorithms, Performance
Keywords Recommendation Systems, Generative Model.

1. INTRODUCTION

Recommender systems have attracted more and more attentions. Recently due to the wide use of mobile devices and ubiquitous sensors, it is essential to incorporate geographical information in rating an item [1]. Many studies have been conducted to explore the benefit of combining the geographical information for recommendation. Some works took the geographical information as a general attribute dimension to extend feature vector [1, 2]. Resnick et al. [3] presented LCARS framework which models both of the geographical information and the content of items for rating prediction.

However, these works didn't simultaneously consider the impacts of users' geographical information and items' content (profile used to represent item) for rating prediction. In this paper, we propose a Bayesian generative model to combine geographical information of users and content of items for modelling rating. Fig 1 describes the modelling process. We suppose there exist many latent groups for users and some latent topics for items. Thus, value of a user rating a item is derived by computing the weighted rating value through integrating all possible latent group assignment of the user and topic assignment of the item.

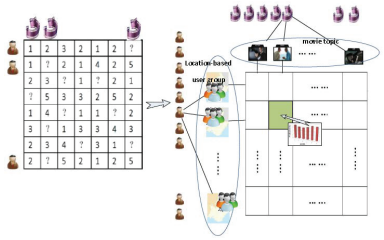


Figure 1: Rating matrix analysis

2. METHODOLOGY

The proposed model, Clscpr, is mainly a probabilistic mixture generative model. In Fig.2, the model presents that each rating is determined by the rating value distribution of user's group to item's topic. The users and items are symmetrically modelled in the Clscpr model, which learns both location-based user group and content-based item topic. In the Clscpr model, the ratings are generated by the following way:

1. Choose $K^l \times K^g$ distributions over rating $\Psi_{ij} \sim Dir(\xi)$.
2. Choose a distribution over K^l location-based user group for each user $\Omega_u \sim Dir(\omega)$.
3. Choose a distribution over K^g item topics for each item $\theta_v \sim Dir(\alpha)$.
4. For each user-item pair (uv):
(a) Choose a user group $z_l^u \sim Multinomial(\Omega_u)$.
(b) Choose a movie group $z_g^v \sim Multinomial(\theta_v)$.
(c) Choose a rating $r_{uv} \sim p(r_{uv}|z_l^u, z_g^v, \Psi_{z_l^u z_g^v})$ here is Multinomial distribution.

In the modelling procedure, we suppose that Ψ_{ij} is the distribution over the rating values $1...K^s$ when ith group rates jth topic. Additionally, l_u is the specific location of the uth user. $\phi_k^{l_i}$ determines generative probability of the location of the user i from the kth group. Here, we suppose

Copyright is held by the author/owner(s).

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577342 .

that each group has parameters: Expection μ and Variance Σ, and user's location conforms to Gaussian distribution for each group, which can be denoted as $l_i \sim N(\mu_j, \Sigma_j)$ where l_i represents the location of the ith user and (μ_j, Σ_j) represents the parameters of the jth group. w_v is the word set of the content of the vth item. $\Theta_k^{g_j}$ determines the generative probability of the content of the jth item generated by the kth topic. Here, we suppose the generative probability is constituted by integrating all of generative probability of words of item's content under the assigned topic of the item.

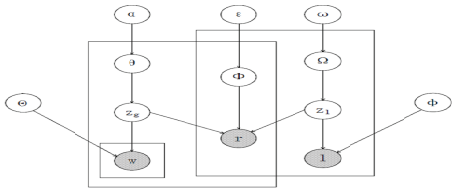


Figure 2: The Graphical model of Clscpr

Now our goal is to model the observed rating based on prior parameters. Hence, we put everything together, and then we obtain the joint distribution for the Clscpr model as Equation 1.

$$\begin{aligned} & P(r_{ij}|\alpha, \varepsilon, \omega) \\ &= \int\int P(r_{ij}|\vec{\Phi_{ij}}) P(\vec{\Phi_{ij}}|\varepsilon) P(\vec{\theta_i}|\alpha) P(\vec{\Omega_j}|\omega) \\ &= \sum_{p=1}^{K^g} \sum_{q=1}^{K^l} P(r_{ij}|z_g^i = p, z_l^j = q, \vec{\Phi_{ij}}) P(z_g^i = p|\vec{\theta_i}) \\ & P(z_l^j = q|\vec{\Omega_j}) P(\vec{\Phi_{ij}}|\varepsilon) P(\vec{\theta_i}|\alpha) P(\vec{\Omega_j}|\omega) \end{aligned} \tag{1}$$

Model Inference. We use collapsed Gibbs sampling to obtain the samples of hidden variable assignments and to estimate unknown parameters $\{\Phi, \Theta, \Psi, \theta, \Omega\}$ in the model. Specifically, we employ a two-step Gibbs sampling procedure. Due to space constraints, we show only the derived Gibbs sampling formulas, and omit the detailed derivation process. We first sample the coin z_l according to the posterior probability:

$$\begin{aligned} & P(z_i^l = k|z_{-i}^l, R, L, \Omega, \Phi, \Theta) = \\ & \prod_{j=1}^{k^g} \frac{n_{l_k,-i}^{r_{ij}} + \beta_{r_{ij}}}{\sum_{s=1}^{k^s}(n_{l_k,-i}^s + \beta_{r_{ij}})} \times \frac{n_k^{-i} + \omega_k}{\sum_{t=1}^{k^s}(n_t^{-i} + \omega_t) - 1} \times \phi_k^{l_i} \end{aligned} \tag{2}$$

$$\phi_k^{l_i} = \frac{1}{2\pi\sqrt{|\Sigma_k|}} exp(-\frac{1}{2}(l_i - \mu_k)^T \Sigma^{-1}(l_i - \mu_k)) \tag{3}$$

We then sample the z_g according to the posterior probability:

$$\begin{aligned} & P(z_j^g = k|z_{-j}^g, R, L, \Omega, \Phi, \Theta) = \\ & \prod_{i=1}^{k^l} \frac{n_{g_k,-j}^{r_{ij}} + \beta_{r_{ij}}}{\sum_{s=1}^{k^s}(n_{g_k,-j}^s + \beta_{r_{ij}})} \times \frac{n_k^{-j} + \alpha_k}{\sum_{t=1}^{k^s}(n_t^{-j} + \alpha_t) - 1} \times \Theta_k^{g_j} \end{aligned} \tag{4}$$

$$\Theta_k^{g_j} = \prod_{i=1}^{n^w} p(w_{ji}|z_j^g = k) \tag{5}$$

After a sufficient number of sampling iterations, we can obtain group assignment of users and topic assignment of items. Hence, $\phi_k^{l_i}$ can be computed according to equation 3 after group parameter of each group is attained, and $\Theta_k^{g_j}$ can be computed according to equation 5, where $P(w_{ji} = c|z_j^g = k) = N_k^c/N_k$ where N_k represents the size of item set of the kth topic and N_k^c represents the size of item set where each item member has topic k and contains word c in its content.

we can estimate the parameters $\Omega_i, \theta_i, \Phi_{ij}^r$ as follow:

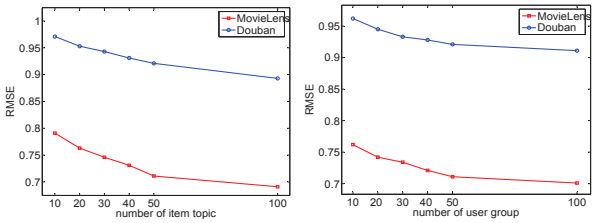
$$\Omega_i = \frac{n^i + \omega_i}{\sum_{j=1}^{k^l}(n^j + \omega_i)}; \theta_i = \frac{n^i + \alpha_i}{\sum_{j=1}^{k^g}(n^j + \alpha_i)}; \Phi_{ij}^r = \frac{n_{ij}^r + \varepsilon_r}{\sum_{t=1}^{k^s}(n_{ij}^t + \varepsilon_r)} \tag{6}$$

Model Application. A rating task in our recommendation system takes four arguments ($u <$ *user* $>$, $l_u <$ *location of user* $>$, $v <$ *item* $>$, $c_v <$ *content of item* $>$). After we obtain parameters $(\Omega, \theta, \Phi, \phi, \Theta)$, we can calculate the predicted value of r_{ij} as follow equation.

$$\tilde{r_{ij}} = \sum_{p=1}^{k^l} \sum_{q=1}^{k^g} \sum_{t=1}^{k^s} \Phi_{ij}^t \theta_p^i \Theta_p^{g_i} \Omega_q^j \phi_q^{l_j} \tag{7}$$

3. EXPERIMENTS

The experiments are based on two data sets: Douban and MovieLens. In order to examine the performance of the methods in the experiments, we adopt two below methods: The Root Mean Squared Error (RMSE) and The Mean Absolute Error (MAE). To test the effectiveness of our method, we compare our method with Pure content-based predictor (PCP), User-based Collaborative Filter (UCF) and Matrix Factory (MF).



(a) K^g (b) K^l

Figure 3: Latent variables impacts.

Datasets	Algorithm	PCP	UCF	MF	Clscpr
MovieLens	RMSE	1.153	1.095	0.812	**0.711**
	MAE	1.055	0.998	0.763	**0.689**
Douban	RMSE	1.323	1.215	1.012	**0.921**
	MAE	1.112	1.069	0.963	**0.889**

Table 1: Precision comparison of different models

Table 1 demonstrates the experimental results under two standard evaluation metrics RMSE and MAE. The experimental results show that out model has higher accuracy than other baseline methods on both data sets. Actually, our proposed model needs two additional predefined parameters, which are K^l representing the number of the groups and K^g representing the number of the topics. Thus, we implement the model by varying the values on both parameters separately, use RMSE to measure the precision of model and then we get the results in Fig. 3. We can find that the error rate decreases as parameter increasing on both situations, and there are small changes beginning from 50.

Acknowledgments. This work was supported by the NSFC (No. 61370025) and the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences (No.XDA06030200).

4. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of- the-art and possible extensions. *TKDE*, 3:734–749, 2005.

[2] J. Levandoski and et al. Lars: A location-aware recommender system. In *ICDE2012*, pages 450–461. IEEE, 2012.

[3] H. Yin and et al. Lcars: A location-content-aware recommender system. In *KDD2013*, pages 221–229. ACM, 2013.