Late-Breaking Developments in the Field of Artificial Intelligence
Papers Presented at the Twenty-Seventh AAAI Conference on Artificial Intelligence

Supervised Topic Model with Consideration of User and Item

Sheng Wang[§], Fangtao Li[†] and Ming Zhang[§]
[§]School of Electronics Engineering and Computer Science, Peking University
[†]Department of Computer Science and Technology, Tsinghua University

Abstract

In this paper, we propose a new supervised topic model by incorporating the user and the item information. The proposed model can simultaneously utilize the textual topic and user-item factors for label prediction. We conduct prediction experiment with a public review dataset. The results demonstrate the advantages of our model. It shows clear improvement compared with traditional supervised topic model and recommendation method.

1. Introduction

Probabilistic topic model (Blei, Ng, and Jordan 2003; Hofmann 1999; Blei 2012), which aims to analyze the words of the original texts to discover themes, has become a poplar topic in recent years. Most of previous topic models focused on the textual content of the document without any labels (Rosen-Zvi et al. 2004). Recently, supervised topic models (Blei and McAuliffe 2007; Zhu, Ahmed, and Xing 2009; Ramage et al. 2009) are proposed to analyze the textual content and label, i.e. review rating, in topic level. However, these supervised models only consider the textual content, but seldom utilize the author of the document and the item expressed in the document. Taking review rating prediction as an example (Titov and McDonald 2008; Wang, Lu, and Zhai 2010; Li et al. 2011), different users may use different expressions for different products or items, even with same rating score. Therefore, it is interesting to incorporate the user and item information into textual topic models for rating prediction.

From another point of view, the research community of recommender system has also investigated rating prediction (Resnick and Varian 1997). They assume that the users who have similar rating behaviors tend to give similar ratings for similar items. They mainly employ the previous user-item relationship (i.e. rating behavior) for rating prediction. For example, the matrix factorization technique (Hu, Koren, and Volinsky 2008) is widely used to find latent behavior patterns for recommendation. However, besides the behavior information, the expression from the user to the target item is also an important factor when we recommend items to the user. This is especially important with the development of Internet, where more and more user-generated-data are expressed in microblogs, forums, news sites et al. Therefore, it is also interesting to incorporate the textual evaluations into user-item based recommender system.

Based on the above observations, we think it is interesting to provide a unified model, which can simultaneously consider the textual and user-item information. In this paper, we propose a new supervised topic model with consideration of user and item information. This new model can be considered as a combination of traditional supervised topic model and recommendation techniques. From the discussion above, we hope this model can benefit both textual analysis and recommender system.

2. Proposed Model

Before introducing our new model, we first briefly review the traditional supervised topic model and recommender system. One of the most widely used supervised topic models is sLDA (Blei and McAuliffe 2007). sLDA predicts the document rating by modeling the topics of textual content, which can be denoted as $r = f(\theta)$, where r is the rating label, θ is the extracted topic from the document. For recommender system, one of the typical techniques is matrix factorization (MF). Based on the user-item rating matrix, MF predicts the rating by modeling user and item information, which can be denoted as $r = f(u, v)$, where u is the user latent factor, and v is the item latent factor. PMF (Salakhutdinov and Mnih 2008) is the probabilistic version of MF. In this paper, we aim to propose a new supervised probabilistic model, which can simultaneously model the textual topic and user-item latent factors for rating prediction. It can be denoted as $r = f(\theta, u, v)$. Following this idea, we will briefly introduce our new model.

The graph representation of our probabilistic model is shown in Figure 1. We can see that it can be considered as a combination of sLDA and PMF. The generative process of our model is as follows:

1. For each user i, draw user latent factor $\mathbf{u}_i \sim \mathcal{N}(0, \lambda_{\mathbf{u}}^{-1} I_k)$
2. For each item j, draw item latent factor $\mathbf{v}_j \sim \mathcal{N}(0, \lambda_{\mathbf{v}}^{-1} I_k)$
3. For each document d, expressed by user i for item j
   (a) Draw topic proportions $\theta_d \mid \alpha \sim \text{Dir}(\alpha)$.
   (b) For each word
      i. Draw topic assignment $z_{dn} \mid \theta_d \sim \text{Mult}(\theta_d)$.
      ii. Draw word $w_{dn} \mid z_{dn}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{dn}})$.

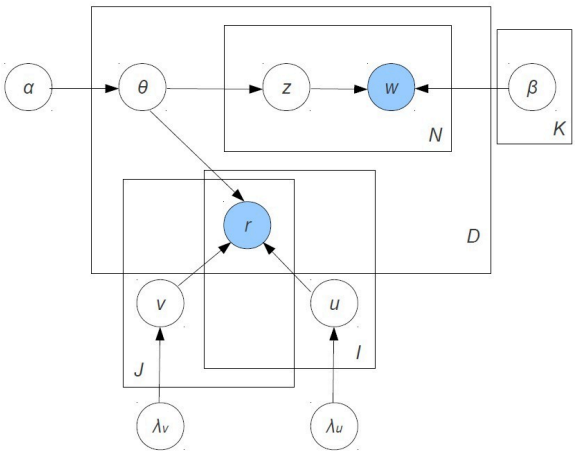

Figure 1: Graph Representation

4. For each user-item-document triple (i, j, d), draw the rating $r_{ij} \sim \mathcal{N}(< \mathbf{u}_i, \mathbf{v}_j, \theta_d >, \sigma)$

In our model, we use the tensor outer products among reviews topic proportion, user latent factor and item latent factor to predict rating,

$$< \mathbf{u}_i, \mathbf{v}_j, \theta_d >= \sum_{f=1}^{K} \mathbf{u}_{if} \cdot \mathbf{v}_{jf} \cdot \theta_{df} \quad (1)$$

where K is the dimension of latent factor.

The log likelihood function of U,V, θ is,

$$\mathcal{L} = -\frac{\lambda_{\mathbf{u}}}{2} \sum_i \mathbf{u}_i^T \mathbf{u}_i - \frac{\lambda_{\mathbf{v}}}{2} \sum_i \mathbf{v}_i^T \mathbf{v}_i +$$

$$\sum_d \sum_n \log \sum_k \theta_{dk} \beta_{k,w_{dn}} - \frac{1}{2} \sum_{i,j} (r_{ij} - < \mathbf{u}_i, \mathbf{v}_j, \theta_d >)^2$$

Since computing the full posterior of $\mathbf{u}_i$, $\mathbf{v}_j$ and θ_d is intractable, we develop an EM-style algorithm. For more detailed analysis, please refer to (Hu, Koren, and Volinsky 2008; Wang and Blei 2011; Blei, Ng, and Jordan 2003), or see our later version of this paper. Based on the model, the new rating is predicted according to the corresponding $\mathbf{u}_i$, $\mathbf{v}_j$ and θ_d with: $r_{ij} =< \mathbf{u}_i, \mathbf{v}_j, \theta_d >$

3. Experiments

Same as the previous studies (Blei and McAuliffe 2007; Li et al. 2011), we employ the public available movie review data set (Pang and Lee 2005). A user may write a review for the target movie with a specified rating score. Our task is to predict this rating. The data set contains 15501 reviews with 458 users and 4543 target movies. The rating scale ranges from 1 to 4.

Since our task is a rating task, beside accuracy, we also employ the following evaluation metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE): $MAE = \frac{\sum_{(i)} |p_i - r_i|}{n}$, $RMSE = \sqrt{\frac{1}{n} \sum_{(i)} (p_i - r_i)^2}$ where r is the true rating, p is the predicted rating. A smaller value of MAE or RMSE indicates a more accurate prediction.

The experimental results are shown in Table 1. We compare our model with traditional supervised topic model sLDA (Blei and McAuliffe 2007), recommendation method PMF (Salakhutdinov and Mnih 2008). We also train a supervised classifier with topic model LDA (Blei, Ng, and Jordan 2003) as our baseline. From the table, we can see that our proposed model can truly benefit the traditional textual topic analysis and user-item recommendation techniques. On the one hand, our model achieves better results than sLDA and LDA with classifier, which demonstrates the importance of user-item impact for general textual topic analysis. On the other hand, our model also shows improvement compared to recommendation method PMF. We think it is necessary to incorporate the textual opinion into traditional recommendation method, if we consider the review is the opinion expression towards the item by the user.

	Accuracy	RMSE	MAE
LDA + classifier	0.3359	1.2625	0.9421
sLDA	0.4046	1.0032	0.7263
PMF	0.3768	1.2393	0.8950
Our Model	0.4324	1.0023	0.7035

Table 1: Experimental results

4. Discussions

In this paper, we propose a new supervised topic model, which can simultaneously model the textual topics and user-item latent factors for rating prediction. We should emphasize that our model is much different from the traditional content based collaborative filtering (CBCF) (Melville, Mooney, and Nagarajan 2002; Agarwal and Chen 2010; Wang and Blei 2011). Their content mainly refers to the description of user or item. While, here, the content means the textual expression or opinion to the item from the user. In the later work, we think it is interesting to test our model in the recommendation system with user-generated-content. For example, we plan to crawl a set of microblogs, including various users and expressed products (items). We want to predict whether it is worth to recommend some products to the target user by incorporating both textual and user-item information with our proposed model.

5. Acknowledgments

This work is supported by the Natioanl Undergraduate Experimental Program of Basic Subject for Computer Science Education, National Natural Science Foundation of China (NSFC Grant No. 61272343), as well as the Doctoral Program of Higher Education of China (FSSP Grant No. 20120001110112).

References

Agarwal, D., and Chen, B.-C. 2010. flda: matrix factorization through latent dirichlet allocation. In *Proceedings of the*

third ACM international conference on Web search and data mining, 91–100. ACM.

Blei, D. M., and McAuliffe, J. D. 2007. Supervised topic models. In *NIPS*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57. ACM.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, 263–272. IEEE.

Li, F.; Liu, N.; Jin, H.; Zhao, K.; Yang, Q.; and Zhu, X. 2011. Incorporating reviewer and product information for review rating prediction. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, 1820–1825. AAAI Press.

Melville, P.; Mooney, R. J.; and Nagarajan, R. 2002. Content-boosted collaborative filtering for improved recommendations. In *AAAI*, 187–192. Menlo Park, CA, USA: American Association for Artificial Intelligence.

Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 115–124. Association for Computational Linguistics.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 248–256. Association for Computational Linguistics.

Resnick, P., and Varian, H. R. 1997. Recommender systems. *Communications of the ACM* 40(3):56–58.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 487–494. AUAI Press.

Salakhutdinov, R., and Mnih, A. 2008. Probabilistic matrix factorization. *Advances in neural information processing systems* 20:1257–1264.

Titov, I., and McDonald, R. 2008. A joint model of text and aspect ratings for sentiment summarization. *Urbana* 51:61801.

Wang, C., and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 448–456. ACM.

Wang, H.; Lu, Y.; and Zhai, C. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 783–792. ACM.

Zhu, J.; Ahmed, A.; and Xing, E. P. 2009. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1257–1264. ACM.