Postprint: Academic Paper Recommendation Algorithm Based on Frequent Topic Set Preferences

Authors: Li Ran, Lin Hong

Date: 2018-05-24T00:00:00+00:00

Abstract

To address the item cold-start problem in academic paper recommendation, we propose a collaborative topic regression model based on preferences for frequent topic sets. This algorithm considers users' preferences for research hotspots when selecting academic papers, utilizes frequent topic sets to represent research hotspots, and models users' preferences for research hotspots as preferences for frequent topic sets. First, the paper-topic probability distribution matrix is mined through a Latent Dirichlet Allocation topic model, and topics with higher probabilities in papers are filtered; then, frequently occurring topic sets are mined to obtain a paper-frequent topic set matrix; finally, users' preferences for frequent topic sets are incorporated when predicting unknown ratings. Experiments on the CiteULike dataset demonstrate that compared with matrix factorization models and collaborative topic regression models, the proposed algorithm achieves improvements in three metrics: recall, precision, and RMSE.

Full Text

Academic Paper Recommendation Algorithm Based on Frequent Topic Sets Preference

Li Ran, Lin Hong
(College of Computer Science & Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: To address the item cold-start problem in academic paper recommendation, this paper proposes a collaborative topic regression model based on preference for frequent topic sets. The algorithm considers users' preferences for research hotspots when selecting academic papers, using frequent topic sets to represent these hotspots and expressing user preferences for research hotspots as preferences for frequent topic sets. First, the paper-topic probability distribution matrix is obtained through the Latent Dirichlet Allocation (LDA)

topic model, and topics with higher probabilities in each paper are filtered out. Then, frequently occurring topic sets are mined to obtain the paper-frequent topic set matrix. Finally, user preferences for frequent topic sets are incorporated when predicting unknown ratings. Experiments on the CiteULike dataset demonstrate that the proposed algorithm improves upon matrix factorization models and collaborative topic regression models in terms of recall, precision, and RMSE metrics.

Keywords: paper recommendation; topic model; frequent topic sets

0 Introduction

Academic paper recommendation represents an application domain of recommender systems. By leveraging characteristics of academic papers, conventional recommendation algorithms such as content-based and collaborative filtering methods, which are widely used in e-commerce, have achieved certain effectiveness in academic paper recommendation. In content-based paper recommendation, the TF-IDF method is commonly employed to represent documents as feature vectors with keywords as dimensions, and document similarity is calculated from these vectors to recommend papers based on users' reading history. However, TF-IDF can only 统计 word frequency information in documents, failing to capture statistical features within and between documents or determine semantic characteristics, thus only recommending articles with superficial content similarity.

With the emergence of topic models and their important role in text mining, researchers have begun applying topic models to recommender systems. Topic models can capture semantic information within documents, discovering latent topic features and enabling content-based recommendations through topic similarity between documents. Additionally, substantial research has focused on better applying the Probabilistic Matrix Factorization (PMF) model to paper recommendation. Wang et al. combined PMF with content-based recommendation, proposing the Collaborative Topic Regression (CTR) model, which enhances item latent factor feature vectors through LDA. The hierarchical Bayesian model of collaborative deep learning performs deep representation learning of content information while conducting collaborative filtering on the feedback matrix, significantly improving existing 技术水平. Lu et al. proposed an author-conference-time-topic model to construct user topic features, which, combined with paper topic features built by LDA, enhance user and item latent factor feature vectors in PMF respectively. Besides these efforts, other recommendation algorithms have concentrated on 挖掘 more types of information to enrich user and paper features. Following this research direction, this paper also explores how to better construct paper feature vectors.

When conducting research in a particular field, users first need to read core papers in that domain to understand the main research content and key technologies. Second, reading newly published papers is crucial for users to keep

pace with 学科 development and broaden their horizons. Meanwhile, users typically pay greater attention to papers containing hotspot topics. Core papers often imply they have been read by many people in that field. Therefore, when recommending core papers, using probabilistic matrix factorization to recommend papers read by other users in the same field allows other users' opinions to play an important role, yielding good recommendation results. However, for papers published recently that have not yet been read, probabilistic matrix factorization cannot function, presenting the item cold-start problem. Thus, it is necessary to analyze paper content to recommend them to interested users. The aforementioned CTR and related work partially alleviate PMF' s cold-start problem by combining content-based recommendation with probabilistic matrix factorization. However, CTR is insufficient in 挖掘 research hotspots, especially for newly published papers that rely basically on content-based recommendation without reflecting the value of research hotspots in papers.

Given these problems, this paper proposes a collaborative topic regression model based on frequent topic set preferences. When predicting unknown ratings, papers containing frequent topic sets are given a certain degree of weight, as frequently occurring topic sets typically represent research hotspots in academia, thereby highlighting the value of academic papers containing research hotspots. The model first processes the corpus to obtain papers' probability distributions over topics, then mines frequently occurring topic sets, and finally incorporates the influence of frequent topic sets into the collaborative topic regression model.

1.2 Frequent Itemset Mining

Frequent itemsets refer to collections of items that frequently appear together, where "frequent" is measured by a predefined threshold (minimum support). The support of an itemset is defined as the proportion of records in the dataset that contain that itemset.

The Apriori algorithm is a classic algorithm for mining association rules and frequent itemsets, using a level-wise search iteration method to generate frequent itemsets—that is, obtaining k-item frequent sets from (k-1)-item frequent sets. It consists of two steps. First, self-joining obtains candidate sets: the candidate set in the first round consists of items in the dataset, while candidate sets in other rounds are obtained through self-joining of frequent sets from the previous round. Second, candidate sets are pruned by removing items in the candidate set whose support is less than the minimum support and whose subsets contain non-frequent sets. This process ultimately yields frequent k-itemsets.

This paper applies the Apriori algorithm to the document-topic distribution generated by the LDA model to mine frequently co-occurring topic sets and obtain the distribution of each frequent topic set across papers.

2.1 Frequent Topic Set Preference

The basic idea of the LDA topic model indicates that each paper has one or more topics, meaning each paper corresponds to a topic set. Therefore, from the paper-topic probability distribution matrix obtained by LDA, as shown in Figure 1: see original paper, by filtering out topics with higher probability values in each paper, the matrix can be represented in the form shown in Figure 1: see original paper, where the set of topics corresponding to value 1 in each row represents the topic set contained in that paper. Obviously, papers in the same research direction contain identical or similar topic sets, and the more frequently a particular topic set appears across different papers, the higher its attention in that research direction.

Based on the above analysis, frequent topic sets—i.e., topic sets that frequently appear in the same academic paper—reflect research hotspots in a particular field to some extent. Papers containing research hotspots often have greater value for users. Distinguishing papers by their "hotness" can recommend more valuable papers to users. Therefore, when constructing paper feature vectors, the influence of frequent topic sets should be considered. Especially for papers with fewer readings, algorithms like CTR rely only on papers' latent topics to recommend them to users who have read similar papers. Incorporating user preferences for frequent topic sets on this basis can further improve recommendation effectiveness.

2.2 Algorithm Model Representation

Probabilistic matrix factorization is a classic recommendation model widely used in academic paper recommendation. It factorizes the user historical rating matrix into user and paper feature matrices in the topic space. This algorithm has high scalability and can incorporate information such as similarity and social networks to constrain user and paper feature matrices, thereby improving recommendation effectiveness. CTR is based on this model and incorporates paper content information, defining the predicted rating $\hat{R}_{ij}$ of user i for paper j as follows:

To incorporate the global influence factor of frequent topic sets into the CTR model, this paper redefines the predicted rating of users for papers as:

$$\hat{R}_{ij} = u_i^T v_j + p^T I_j$$

where u_i and v_j represent the feature vectors of user i and paper j, respectively. θ_j is the probability distribution vector of paper j over topics obtained through the LDA topic model. The user feature vector is still defined as following a Gaussian distribution with mean 0, as shown in equation (3). The paper feature vector is defined as in equation (2), where ϵ_j follows a Gaussian distribution with mean 0, balancing the influence of user rating records and paper content on paper feature vectors.

According to the analysis in the previous section, frequent topic sets have a certain influence on users' paper selection. Therefore, this paper proposes a collaborative topic regression model based on frequent topic set preferences, incorporating the global influence factor p of frequent topic sets into CTR to improve recommendation effectiveness. The model schematic is shown in [Figure 2: see original paper].

I_j represents whether paper j contains frequent topic sets. The t-th value of I_j is 1, indicating that paper j contains the t-th frequent topic set. The generation process of vector I_j is as follows: a) Obtain the paper-topic probability distribution matrix P through the LDA topic model; b) Filter out topics with higher probability values in each paper to obtain the topic set contained in each paper; c) Use the Apriori algorithm to mine frequently occurring topic sets and simultaneously generate the paper-frequent topic set matrix T.

$\hat{R}_{ij}$ represents the predicted rating, with u_i and v_j defined the same as in equations (3) and (2). $g(\cdot)$ is the logistic function that maps predicted ratings to [0,1]. p is the influence factor vector of frequent topic sets, where p_t represents the influence value of the t-th frequent topic set on user paper ratings, and $|p|$ is the dimension of frequent topic sets. $I_j^T I_j$ represents the number of frequent topic sets contained in paper j, i.e., the number of 1s in vector I_j. When paper j does not contain any frequent topic sets, the influence value of frequent topic sets is defined as the average of the influence values of all frequent topic sets, and vector I_j represents a unit vector. Moreover, vector p is assumed to follow a Gaussian distribution with mean 0, like vectors u_i and v_j:

$$p \sim N(0, \sigma_p^2 I)$$

The loss function can then be derived as defined in equation (7). R_{ij} is the true rating of user i for paper j; I_{ij} is an indicator function that returns 1 if user i has interacted with paper j and 0 otherwise; λ_u, λ_v, and λ_p are regularization parameters for u_i, v_j, and p, respectively.

By applying stochastic gradient descent to vectors u_i, v_j, and p as shown in equation (8), we can solve for the values of user latent topic vectors, paper latent topic vectors, and frequent topic set influence factor vectors that minimize the loss function, thereby predicting unknown ratings through equation (1).

3.1 Experimental Setup

This paper adopts the dataset provided by the CiteULike website (http://www.citeulike.org/faq/data.adp). The dataset includes 16,980 papers and 5,551 users from 2004 to 2010. Each user has a personal paper library recording browsed papers, containing a total of 204,986 user-paper browsing records. During the experiment, the LDA topic model algorithm and Apriori algorithm are applied sequentially to the 16,980 papers to mine frequently occurring topic sets. Each paper is then represented

as a vector with frequent topic sets as dimensions, yielding matrices P and T as known parameters for predicting unknown ratings.

The user-paper browsing records are divided into training and test sets at a ratio of 80% to 20% for the following experiments: a) Analyze the impact of the number of frequent topic sets and parameter p on the collaborative topic regression model based on frequent topic set preferences to determine reasonable parameter values; b) Compare the recommendation effectiveness of the proposed model with PMF and CTR.

3.2 Evaluation Metrics

In rating prediction systems, Root Mean Squared Error (RMSE) is commonly adopted as a metric, where smaller RMSE indicates higher recommendation accuracy. The RMSE formula is as follows, where $Test$ represents the test set:

$$RMSE = \sqrt{\frac{\sum_{(i,j)\in Test}(R_{ij} - \hat{R}_{ij})^2}{|Test|}}$$

Additionally, since the purpose of recommender systems is to recommend papers that users might be interested in, after predicting user ratings for papers, we sort the predicted ratings and select papers with high scores that users have not yet interacted with for recommendation. Recall and precision are used to measure recommendation effectiveness. Assuming the top m papers with highest predicted ratings are recommended to users, for a specific user, the recall and precision are defined as:

$$recall@m = \frac{TP}{TP + FN}$$

$$precision@m = \frac{TP}{TP + FP}$$

where TP is the number of papers in the recommendation list that the user likes, FN is the number of papers the user likes but are not recommended, and FP is the number of papers in the recommendation list that the user dislikes. The recommendation algorithm' s recall is defined as the average recall across all users, and precision is defined as the average precision across all users.

Moreover, since recall and precision can be contradictory, the F-measure is often used to comprehensively consider both. F-measure is the weighted harmonic mean of recall and precision. Specifically, when $\alpha = 1$, it becomes the most common F1 metric. This paper adopts F1 to measure recommendation effectiveness.

3.3 Experimental Results

3.3.1 Impact of Frequent Topic Set Count

In the stage of mining frequently occurring topic sets, different minimum support values yield different numbers of frequent topic sets, reflecting the distribution of research hotspots in the current paper collection. Setting the topic count of the LDA model to 200 and minimum support to 0.0014, 0.00125, 0.00118, 0.0012, and 0.00105 respectively, the numbers of frequently occurring topic sets satisfying these minimum supports are 54, 81, 97, 118, and 159. presents the model' s average recall at different recommendation list lengths k as the number of frequent topic sets varies. The RMSE trend is shown in [Figure 3: see original paper]. Other parameters in the experiment are set as $\lambda_u = 0.01$, $\lambda_v = 0.01$, $\lambda_p = 1$.

As the number of frequent topic sets increases, RMSE first decreases and recall correspondingly increases, improving algorithm performance. However, when the number of frequent topic sets exceeds a certain level, recommendation effectiveness decreases. Experimental results indicate that during frequent topic set mining, setting a reasonable minimum support to obtain frequent topic sets corresponding to research hotspots enables the proposed algorithm to achieve optimal recommendation effectiveness.

3.3.2 Impact of Regularization Parameter

In equation (7), a smaller parameter λ_p means a larger proportion of user preference for frequent topic sets in predicting unknown ratings. To investigate the impact of frequent topic sets on ratings, the settings of regularization parameters λ_u and λ_v remain the same as in the previous section, while different λ_p values are selected to measure algorithm performance. Experimental results show that when $\lambda_p = 1$, the proposed algorithm achieves optimal recall and relatively small RMSE values.

3.3.3 Recommendation Algorithm Comparison

The proposed model extends the original PMF model and borrows ideas from CTR. Comparing it directly with PMF and CTR models demonstrates the improvement of the proposed model in terms of recall, precision, and RMSE. Therefore, these two models are selected as comparison objects in our experiments.

Through experiments, optimal parameter settings for the three models are obtained: the feature space dimension is 200 for all three models; in PMF and CTR, $\lambda_u = \lambda_v = 0.01$, while in the proposed model, $\lambda_u = \lambda_v = 0.01$, $\lambda_p = 1$. Based on this, with recommendation list length k set to {200, 150, 100, 50, 10} respectively, the effectiveness of the three models in recall, precision, and RMSE is compared. presents detailed experimental results, and [Figure 4: see original paper] shows the performance comparison of the three models.

Experimental results demonstrate that at different recommendation list lengths, the proposed model' s recall, precision, and F1 are significantly superior to PMF and CTR, with RMSE values also reduced. Moreover, the paper-frequent topic set matrix T can be computed offline, so the proposed model achieves improved recommendation effectiveness with relatively small time overhead.

4 Conclusion

This paper considers the influence of frequent topic sets on users' paper selection and proposes a collaborative topic regression model based on frequent topic set preferences, aiming to help users find more valuable academic papers. Experiments on real datasets prove that compared with PMF and CTR models, the proposed model achieves certain improvements in recall and precision.

Due to users' personalized needs, the influence values of frequent topic sets may differ for different users. Therefore, constructing user-sensitive frequent topic set influence vectors is the focus of future research.

References

[1] Ramos J. Using TF-IDF to determine word relevance in document queries [C]// Proc of the 1st Instructional Conference on Machine Learning. 2003.

[2] Lops P, Gemmis MD, Semeraro G. Content-based recommender systems: state of the art and trends [M]. New York: Springer, 2011: 73-105.

[3] Philip S, Shola PB, Ovye A. Application of content-based approach in research paper recommendation system for a digital library [J]. International Journal of Advanced Computer Science & Applications, 2014, 5 (10): 37-40.

[4] Krestel R, Fankhauser P, Nejdl W. Latent Dirichlet allocation for tag recommendation [C]// Proc of the 3th ACM Conference on Recommender Systems. New York: ACM Press, 2009: 61-68.

[5] Huang Zeming. Research on recommended system of scholar paper based on topic model [D]. Dalian: Dalian Maritime University, 2013.

[6] Salakhutdinov R, Mnih A. Probabilistic matrix factorization [C]// Proc of the 20th International Conference on Neural Information Processing Systems. USA: Curran Associates Inc. 2007: 1257-1264.

[7] Wang Chong, Blei DM. Collaborative topic modeling for recommending scientific articles [C]// Proc of ACM SIGKDD: International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 448-456.

[8] Blei D M, Ng AY, Jordan M I. Latent Dirichlet allocation [J]. J Machine Learning Research Archive, 2003, 3: 993-1022.

[9] Wang Hao, Wang Naiyan, Yeung DY. Collaborative deep learning for recommender systems [C]// Proc of International Conference on Knowledge Discovery

and Data Mining. New York: ACM Press, 2015: 1235–1244.

[10] Lu Meilian, Zhang Zhenglin, Liu Zhichao. MFWT: a hybrid model for academic paper recommender [J]. Journal of Beijing University of Posts and Telecommunications, 2016, 39 (4): 24-29.

[11] Li Jingming, Cheng Jiaxing, Zhang Wei, et al. The research on construction of library website personalized recommendation model based on improved collaborative filtering algorithm [J]. Journal of Changchun Normal University, 2016. 35 (2): 43-48.

[12] Zhang Libin. Research on the application of collaborative filtering technology in the personalized recommendation service for academic resources of university libraries [J]. Hebei Library Journal of Science & Technology, 2017. 30 (4): 85-88.

[13] Wang Hao, Li Wujun. Relational collaborative topic regression for recommender Systems [J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27 (5): 1343-1355.

[14] Wang Hao, Chen Binyi, Li Wujun. Collaborative topic regression with social regularization for tag recommendation [C]// Proc of International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2013: 2719-2725.

[15] Agrawal R, Imielinski T, Swami AN. Mining association rules between sets of items in large databases, SIGMOD Conference [C]// Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press. 1993: 207-216.

[16] Wu Liaoyuan, Jiang Jun, Wang Gang. Study of scientific paper recommendation method based on unified probabilistic matrix factorization in scientific social networks [J]. Computer Science, 2016, 43 (9): 219-223.

Note: Figure translations are in progress. See original paper for figures.