Collaborative Filtering Group Recommendation Based on Information Entropy and User Behavior Consistency: Postprint

Authors: Su Mengke, Yang Yupu

Date: 2018-10-11T00:00:00+00:00

Abstract

Within research on collaborative filtering recommendation algorithms that utilize only the input rating matrix as the sole algorithmic input, this work addresses the issue of how disparities arising from different data quality levels impact recommendation results. This includes the emphasis and attention devoted to data quality aspects, methods for characterizing quality differences, and approaches for grouped recommendation modeling tailored to user segments with varying data quality. We propose a data quality characterization that comprehensively evaluates data quality and groups users by considering two metrics: user behavior consistency and user information entropy. For users in different groups, more precise recommendation modeling can be conducted based on analysis of their historical behavior. Experimental results demonstrate that disparities in data quality indeed exert a significant influence on recommendation accuracy improvement, while also substantiating the necessity of user grouping for recommendation purposes. The experimental results further indicate that the comprehensive characterization employing both user behavior consistency and user information entropy metrics yields the most pronounced accuracy enhancement.

Full Text

Preamble

Collaborative Filtering Group Recommendation Based on Information Entropy and User Behavior Consistency

Su Mengke, Yang Yupu (Key Laboratory of System Control & Information Processing, Ministry of Education of China, Dept. of Automation, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: In collaborative filtering recommendation algorithms that use only the rating matrix as input, differences in data quality significantly impact recommendation results. This raises concerns about the importance of data quality, how to characterize quality differences, and how to model group recommendations for user groups with different data quality. This paper proposes a data quality characterization method that comprehensively evaluates data quality using two metrics—user behavior consistency and user information entropy—to group users. For different user groups, more accurate recommendation modeling can be performed based on analysis of their historical behavior. Experimental results demonstrate that data quality differences indeed have an important impact on recommendation accuracy improvement, while also proving the necessity of group recommendation. The results further show that the combined characterization using both user behavior consistency and user information entropy yields the most significant accuracy improvement.

Keywords: information entropy; noise characterization; data quality difference; user behavior consistency; collaborative filtering

---

0 Introduction

The rapid development of internet technology has accelerated the explosive growth of online resources, with vast amounts of information flooding the network daily. This information overload problem makes it increasingly costly for users to find information of interest through traditional search methods, while also making it difficult for them to effectively process the exponentially growing information. As an important technology for solving information overload, recommender systems have been widely applied on major platforms such as Amazon's e-commerce platform and various social networking sites. Among these, collaborative filtering-based recommendation algorithms have received extensive research and application since their emergence, primarily due to their simplicity and scalability. The classic user-based collaborative filtering algorithm finds a set of highly similar users for each target user by calculating similarity between users based on co-rated items. Model-based collaborative filtering algorithms, on the other hand, use rating data to mathematically model user rating patterns. Matrix factorization models map both users and items to a common low-dimensional latent space, attempting to explain ratings through the dot product of user and item vectors in this latent space. Later, with the development of the Netflix competition, matrix factorization and its improved models stood out due to their outstanding performance.

The advantage of model-based collaborative filtering algorithms lies in their ability to improve recommendation accuracy. The key to these algorithms is using training datasets to offline learn a prediction model, where both the algorithm and the data are two critically important factors affecting model accuracy. Traditional recommendation algorithms assume that the data quality in the rating dataset is uniform and that different users' behavioral data ac-

curately represent their true preferences, thus treating all users equally during modeling without classification. However, real-world recommender systems in e-commerce platforms often contain malicious users who provide feedback of little reference value, while some users' feedback is highly beneficial for making recommendations. In other words, different users' behavioral data exhibits quality differences—some users demonstrate consistent behavior with stable feedback, which not only reduces modeling difficulty but also enables more accurate recommendations, while other users show large inconsistencies in their behavior, making their data difficult to utilize and resulting in less accurate outcomes. Therefore, data quality and quantity greatly affect recommendation accuracy, constituting a noise problem in data that has not received widespread attention. Noise in recommender systems can generally be divided into two categories: (a) deliberate noise that appears for certain purposes, such as improving commercial interests or maliciously disrupting systems; and (b) natural noise where users rate too casually without truly expressing their opinions.

Recent related research has simply transformed the data quality problem into a partial denoising issue. For example, Chirita et al. proposed an evaluation metric for identifying malicious users, Bilge et al. used K-means clustering to identify malicious users and water army accounts, and Cao et al. approached the problem from a semi-supervised shilling attack detection perspective, using a small amount of data to pre-train a Bayesian classifier that then adapts to all data to obtain a final classifier. While these methods address recommendation accuracy problems caused by noise, they share the common drawback of requiring complex model training and parameter tuning, and noise is only one manifestation of data quality issues.

Liu Jiangdong et al. proposed using the concept of information entropy, comprehensively considering user information entropy and rating timeliness to filter partial users and improve recommendation accuracy. Yu Penghua considered data quality and quantity issues from a data perspective, grouping rating data and analyzing the impact of data quality and quantity differences on recommendation results. Zhang Jia et al. used user information entropy to express users' rating distributions and determine rating tendency degrees, utilizing information entropy to eliminate users with obvious tendency differences when determining a user's nearest neighbors in traditional user-based collaborative filtering, thereby improving final recommendation accuracy. Gao Cuihua et al. comprehensively considered information entropy and fuzzy clustering, using information entropy to measure uncertainty in membership degrees and proposing a collaborative filtering algorithm with information entropy-weighted fuzzy clustering that improved recommendation accuracy while simplifying algorithm complexity. Kluver et al. proposed that information entropy could characterize user rating quality, while Bellogín et al. proposed a new rating quality evaluation through analysis of user historical behavior. To fully demonstrate the impact of data quality on recommendation results, this paper comprehensively considers information entropy and user behavior consistency to jointly and comprehensively characterize data of different quality, and proposes a group collabora-

tive filtering recommendation algorithm based on different quality and quantity data subsets. This paper analyzes user quality by considering two extreme rating scenarios—casual and concentrated—and groups users into different quality data subsets for group recommendation, further achieving personalized modeling for different users and improving recommendation accuracy.

1.1 Basic Model BMF

To analyze the relationship between rating quality differences and recommendation accuracy, we examine the input to matrix factorization models as shown in Table 1. The table displays all rating sets for m users on n items, where if user u has rated item i, the corresponding rating is shown. The rating matrix is generally sparse, with unrated movies represented by 0 in the matrix. The key step in recommendation is using existing rating data to mathematically model the underlying rating rules for users and items and predict unknown ratings.

The Biased Matrix Factorization (BMF) model is an improved version of the latent factor model that became the most popular recommendation algorithm due to its outstanding performance in the 2009 Netflix competition. The BMF algorithm finds a low-dimensional latent factor space that maps original users and items into this space. Item i is mapped to vector q_i, where vector components represent the extent to which item i contains these basic factors. User u is mapped to vector p_u, where vector components represent the degree of user preference for these factors. After low-dimensional mapping, the preference degree of user u for item i can be represented by the dot product of user and item vectors.

Based on matrix factorization, the recommendation algorithm uses the shown matrix as algorithm input, mines latent factors of users and items for mathematical modeling, and predicts unknown ratings. This paper selects the BiasedMF (BMF) algorithm as the foundation and proposes improvements incorporating user information entropy and behavior consistency. The basic BMF algorithm is as follows:

Equation (1) represents the composition of rating bias terms, where represents the overall average rating, b_u represents user bias, and b_i represents item bias. Equation (2) shows that the predicted rating consists of the model prediction result plus bias terms. Training the model makes the predicted rating infinitely close to the actual rating, transforming into the optimization problem shown in Equation (3), where λ represents the regularization term to improve model generalization ability. The commonly used stochastic gradient descent method is selected to find optimal P and Q. After model training, predictions can be made for users' unrated movies, and recommendations are made according to ratings from high to low.

1.2 Noise Problem

Most related researchers collect the original rating matrix, divide it into training and test sets according to some strategy, and then perform modeling and

testing. Since the rating matrix is the sole input to collaborative filtering algorithms, quality differences in the rating matrix greatly affect the final algorithm results. How to characterize different users' rating quality and perform group recommendation is the concern of this paper.

To address the problem that different users have varying degrees of noisy data, this paper characterizes user rating quality differences from a data noise perspective, comprehensively considering user information entropy and user history to characterize rating quality and grouping users into different quality subsets. According to literature [8], ratings from users with low credibility exhibit excessive concentration, such as one-sided ratings or ratings only for specific targets. This paper introduces information entropy to characterize user rating quality. Information entropy can characterize the uncertainty of random variable values, depicting a random variable's probability distribution. The information entropy value of a random variable is proportional to the confusion degree of its distribution. The information entropy definition for a variable is shown in Equation (4).

From Equation (4), we know that information entropy is closely related to variable probability distribution. User information entropy can reflect the richness of user ratings. If user information entropy is too low, it indicates that user ratings are overly concentrated. However, if we only filter out users with overly concentrated ratings, those with rich ratings are not necessarily reliable. Therefore, user rating quality cannot be characterized by concentration alone. According to Bellogín's proposal based on user historical behavior data, user behavior consistency is introduced to measure rating quality from another perspective. User behavior consistency reflects whether user ratings are coherent over time, allowing users to be divided into groups with higher and lower behavior consistency degrees. This paper comprehensively considers user information entropy and user behavior consistency to analyze the impact of different data quality on recommendation results.

2.1.1 Characterizing User Rating Quality with Information Entropy

How to characterize the quality differences in rating matrices and divide different users into quality subsets? This paper characterizes data quality from a noise perspective and proposes comprehensively analyzing user rating quality using both information entropy and historical behavior. Based on the definition in Equation (4), let R_u represent all rating information for user u. For user u, define the probability of a rating value as the proportion of times that rating appears to the total number of ratings, as shown in Equation (5). User information entropy is defined as C_1(u) in Equation (6).

User information entropy reflects the possibility of a user exhibiting water army characteristics. According to the formula definition, low user information entropy indicates a higher possibility of water army behavior. Figure 2 and Figure 3 show the distribution and quantity statistics of information entropy for 6,040 users, indicating that most rating quality is reliable.

2.1.2 Characterizing User Rating Quality Based on User History

Based on users' historical rating behavior, Figure 1 shows rating sets for two users on movies with similar genre distributions. Users' preferences are reflected by rating levels. The user on the left in Figure 1 prefers thriller movies over romance and ethics movies, while the user on the right has more scattered behavior data that is difficult to mine for patterns. In the long term, we can believe that the left user's ratings are more stable than the right user's, because the left user gives similar ratings to similar movie types.

Assuming coherent user behavior is manifested by small rating deviations for the same or similar types, we define user behavior stability using rating deviations in similar item spaces. For any user, divide rated items into different feature spaces. Let R_f(u) represent the set of items rated by user u in a specific feature space, and $\bar{r}$_f(u) represent the average rating of user u in this feature. Let R(u) represent all rating sets of user u. Then user behavior consistency C_2(u) can be characterized by the deviation of user ratings across various feature spaces, as shown in Equation (8).

σ_f^2(u) represents a user's specific rating variance in feature space f. σ^2(u) represents the overall rating variance, essentially a weighted average of the user's specific rating variances. Its value is proportional to user behavior consistency. If user behavior is consistent over time, these users are easier to model.

2.2 User Grouping and Group Recommendation

This paper adopts sequential user grouping for the proposed method. Through analysis of user rating data, each user obtains corresponding C_1(u) and C_2(u) metrics. Regarding the information entropy metric, since information entropy is defined for user rating data with one-sided characteristics, we first directly filter users with low information entropy by removing them from the original data. Then, based on C_2(u) values, users are clustered into difficult users and easy users to analyze how quality differences in different groups affect recommendation accuracy.

3.1 Experimental Dataset

To test the impact of data quality differences on final recommendation accuracy and verify the effectiveness of quality metrics and the necessity of group recommendation, we use the famous MovieLens1M (ml-1m) public real dataset for experimental evaluation. This dataset includes rating data from over 6,000 users, with each user rating more than 20 movies, totaling 1,000,209 ratings for 3,900 movies. The rating values reflect users' preference degrees for movies, with higher values indicating higher evaluations. This dataset is of manageable size yet rich in information, with quality differences in user data, making it suitable for this verification.

3.2 Evaluation Metrics

Different evaluation metrics apply to different research contexts. Since this pa-

per focuses on characterizing data quality differences, we use the Root Mean Square Error (RMSE) metric from rating prediction to reflect the impact of different quality data groups on recommendation accuracy and analyze the impact caused by different quality differences. RMSE evaluates recommendation accuracy by calculating the difference between predicted user ratings and actual user ratings after predicting unknown ratings using a model trained on the training set. RMSE is widely adopted due to its intuitive representation of recommendation accuracy. RMSE is inversely proportional to recommendation accuracy, as shown in Equations (9) and (10).

3.3.1 Characterizing User Rating Quality

Taking users in the dataset as a baseline, for each user's rating data, we calculate the corresponding user information entropy and user behavior consistency. The information entropy C_1(u) is shown in Figure 2 and Figure 3, and the consistency C_2(u) is shown in Figure 4 and Figure 5.

3.3.2 User Grouping Recommendation

First, users are filtered based on C_1(u) values by selecting different thresholds. For user rating data below the threshold, we directly delete these users. Experiments show that when C_1(u) = 0.5, the accuracy improvement is maximal. On this basis, users are divided into difficult users and easy users according to C_2(u) values (with threshold 8$×$10^-2). At this point, the difficult user group contains 509,623 ratings, and the easy user group contains 490,586 ratings, each approaching 50% of the original dataset to exclude the impact of data quantity on recommendation results. Both groups use 5-fold cross-validation to split into training and test sets, denoted as S_d^{train}, S_d^{test}, S_e^{train}, S_e^{test}. Let S_1 represent modeling recommendations on the difficult user subset, S_2 represent modeling recommendations on the easy user subset, S_3 represent modeling recommendations on the original data (comparison baseline), and S_4 represent modeling recommendations on the user group after information entropy filtering of the original data. Figure 6 shows the RMSE errors under different quality groupings, where the horizontal axis represents different thresholds. As C_1(u) increases, the rating matrix becomes increasingly sparse, thereby affecting recommendation accuracy.

3.3.3 Comparative Experiment

The improved model-based collaborative filtering algorithm (ABMF) proposed in this paper, which integrates user information entropy and user behavior consistency, is compared with the basic BMF algorithm, reducing the RMSE metric by 1.1%. The UEITMF method proposed in literature [16] integrates information entropy and timeliness, considering real-time dynamic characteristics, but shows insignificant accuracy improvement compared to the original method and increases algorithm complexity. Literature [17] only considers data quality from a data perspective, single-handedly considering changes in user behavior over time, also achieving certain recommendation accuracy improvement. Entropy-based-CF [18] integrates information entropy on the basis of user-based collab-

orative filtering, with high logicality and understandability. Table 2 shows the RMSE comparison of different algorithms.

3.4 Experimental Results Analysis

Comparing S_4 and S_3, without considering user behavior consistency and only performing information entropy filtering on users, although when C_1(u) = 0.5, recommendation accuracy improves from 0.8614 to 0.8594 with a weak effect, the proposal of this quality metric should have obvious effectiveness for analyzing behaviors like brushing orders in e-commerce systems under big data scenarios. After filtering users with extremely poor quality, the remaining users are grouped according to C_2(u) values. As shown in Figure 6, S_1 represents modeling on the difficult user subset and testing on the difficult user subset, S_2 represents modeling on the easy user subset and testing on the easy user subset. S_1 improves recommendation accuracy by 0.2% relative to S_4, S_2 improves by 1.1% relative to S_4, and S_2 changes the recommendation metric by 3.2% relative to S_1. The changes in recommendation accuracy demonstrate that group recommendation is necessary, and S_1, S_2, S_3, S_4 represent data quality from low to high, with RMSE clearly showing a decreasing trend, meaning recommendation accuracy shows an improving trend. Experimental results demonstrate that the proposed comprehensive analysis of user rating quality using both information entropy and user behavior consistency is very effective, with significant differences in modeling recommendation accuracy among different quality groups.

4 Conclusion

This paper proposes research on improving model-based recommendation algorithm accuracy from a novel perspective of data quality. It comprehensively considers rating concentration and casualness, characterizing user rating quality differences through user information entropy and behavior stability respectively, and proposes an improved method of grouping users into different quality subsets and modeling separately. Experimental results fully demonstrate that data quality differences have an important impact on improving recommendation accuracy. Therefore, in the current big data development environment, data quality issues should receive widespread attention. Data quality issues become more prominent as data scale increases, with larger datasets showing more significant data quality differences. How to use these two metrics of information entropy and user behavior consistency to better analyze large datasets and reasonably group users is the next research direction. This paper achieves rating quality grouping for users; how to set reasonable classification standards for different items and define corresponding quality metrics to classify data from different perspectives are also future considerations.

References

[1] Chu Wei, Park S T. Personalized recommendation on dynamic contents using predictive bilinear models [C]// Proc of the 18th International Conference on World Wide Web. New York: ACM Press, 2009: 691-700.

[2] Shen Jian, Yang Yupu. Collaborative filtering recommendation algorithm based on two stages of similarity learning [J]. Application Research of Computers, 2013, 30 (3): 715-719.

[3] Wu Jinlong. Collaborative filtering algorithm in the Netflix Prize [D]. Beijing: Beijing University, 2010.

[4] Robert B, Yehuda K, CHRIS V. Modeling relationships at multiple scales to improve accuracy of large recommender systems [C]// Proc of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2007: 95-104.

[5] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems [J]. Computer, 2009, 42 (8): 30-37.

[6] Meng Xiangwu, Liu Shudong, Zhang Yujie, et al. Research on social recommender systems [J]. Journal of Software, 2015, 26 (6): 1356-1372.

[7] Gunes I, Kaleli C, Bilge A, et a1. Shilling attacks against recommender systems: a comprehensive survey [J]. Artificial Intelligence Review, 2014, 42 (4): 767-799.

[8] Kluver D, Nguyen T, Ekstrand M, et a1. How many bits per rating [C]// Proc of the 6th ACM Conference on Recommender systems. Dublin, Ireland: ACM Press, 2012: 99-106.

[9] Chirita P A, Nejdl W, Zamfir C. Preventing shilling attacks in online recommender systems [C]// Proc of the 7th Annual ACM International Workshop on Web Information and Data Management. New York: ACM Press, 2005: 67-74.

[10] Bilge A, ZdemiRZ, Polat H. A novel shilling attack detection method [J]. Procedia Computer Science, 2014, 31: 165-174.

[11] Cao Jie, Wu Zhiang, Mao Bo, et al. Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system [J]. World Wide Web, 2013, 16 (5//6): 729-748.

[12] Bellogin A, Said A, Pdevries A. The magic barrier of recommender systems no magic just ratings [C]// Proc of User Modeling, Adaption, and Personalization. Aalborg, Denmark: Springer International Publishing, 2014.

[13] Hofmann T. Latent semantic models for collaborative filtering [J]. ACM Trans on Information Systems, 2004, 22 (1): 89-115.

[14] Zhang Xuesheng. Research on collaborative filtering recommendation algorithm for data sparse [D]. Hefei: University of Science and Technology of China, 2011.

[15] Pang Yanwei, Ma Zhao, Pan Jing, et al. Robust sparse tensor decomposition by probabilistic latent semantic analysis [C]// Proc of the 6th International Conference on Image and Graphics. Washington DC: IEEE Computer Society, 2011: 893-896.

[16] Liu Jiangdong, Liang Gang, Feng Cheng, et al. Collaborative filtering recommendation based on information entropy and timeliness [J]. Journal of Computer Applications 2016, 36 (9): 2531-2534.

[17] Yu Penghua. Research on several problems of recommendation system for data quantity and quality [D]. Hangzhou: Zhejiang University, 2016.

[18] Zhang Jia, Lin Yaojin, Lin Menglei, et al. Collaborative filtering algorithm based on information entropy [J]. Journal of Shandong University: Engineering Edition, 2016, 46 (2): 43-50.

[19] Gao Cuifang, Huang Shanwei, Shen Wanqiang, et al. Improved algorithm for collaborative clustering based on information entropy weighting [J]. Application Research of Computers, 2015, 32 (4): 1016-1018.

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