Enhancing Collaborative Filtering Recommendation by User Interest Probability

Jing Yu
Search And Recommendation Dept .Datagrand Company Product Technology Center
Shanghai, China
yujing@datagrand.com

Jingjing Shi
Search And Recommendation Dept .Datagrand Company Product Technology Center
Shanghai, China
shijingjing@datagrand.com

Yunwen Chen
Search And Recommendation Dept .Datagrand Company Product Technology Center
Shanghai, China
chenyunwen@datagrand.com

Wenhai Liu
Search And Recommendation Dept .Datagrand Company Product Technology Center
Shanghai, China
liuwenhai@datagrand.com

Kai Liu
Search And Recommendation Dept .Datagrand Company Product Technology Center
Shanghai, China
liukai@datagrand.com

Zhijun Xie
Search And Recommendation Dept .Datagrand Company Product Technology Center
Shanghai, China
xiezhijun@datagrand.com

2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD) | 978-1-6654-1515-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIBD51990.2021.9459028

***Abstract*—Traditional collaborative filtering recommendation is one of the most commonly used algorithms in current recommendation systems, and it is also the mainstream algorithm used in the e-commerce industry. The basic principle of collaborative filtering is that users with similar interests will have similar interest bias in the future. However, the traditional collaborative filtering algorithm faces some of the following problems in the e-commerce industry: (1) lacking of confidence based on user interest bias similarity; (2) without consideration of the time factor; (3) no correlation between behaviors. Therefore, the collaborative filtering model based on user interest probability (PUCF) proposed in this paper to solves the above problems. Firstly, Wilson confidence interval is used to solve the problem of confidence, the behavioral time decay factor is obtained through the normalized time. And then considering the problem of conversion between behaviors. Besides the algorithm assigns different calculation weights to the existence of progressive behavior relationships, and to a certain extent considers the influence of the internal connections between behaviors on user interest. Through looking for the optimal parameters and making comparative experiments, it shows that the effect of collaborative filtering model based on user interest probability (PUCF) proposed in this paper is better than other collaborative filtering methods.**

Keywords—collaborative filtering, Wilson confidence interval, time attenuation factor, user interest probability

I. INTRODUCTION

With the rapid development of the e-commerce industry and the infiltration of e-commerce culture such as 618 and Double Eleven, it is difficult for people to find products that meet their own interests among the hundreds of millions of products quickly. This difficulty is called information overload in academia. In order to solve the problem, the recommendation system is invented.

However, it cannot be ignored that while the e-commerce field has been developing rapidly, it is also facing many difficult problems. First of all, with the amount of data increasing, the computing power of traditional collaborative filtering is facing unprecedented challenges. In response to the above problem, PUCF uses the user's interest in categories to calculate the similarity between users, and the complexity of computation has been greatly reduced.

Secondly, when calculating probability, it often involves the concept of division in statistics. It is also a 10% probability that a certain category is clicked, but due to inconsistent denominators, its confidence ability is also very different. In order to solve the above problems, the article uses Wilson's confidence to make the result more convincing.

Then, for the same user, the time point of its existing behavior also has a certain impact on the subsequent behavior. For most users, the closer the behavior occurs, the better effective the subsequent predictions have. Therefore, through consideration of behavior time, this paper introduces the time attenuation factor of the normalized calculation, so as to effectively add the factors of the behavior time point to the model proposed in this paper.

Finally, and most critically, there are many behaviors in e-commerce industry, such as clicks, favorites, pre-sales, purchases, etc. The traditional collaborative filtering algorithm only considers whether to click, and ignores the interaction between user behaviors. This paper introduces the transformation relationship in PUCF, which increases the accuracy of collaborative filtering to a certain extent.

This article will focus on the PUCF model in detail, the specific chapters are arranged as follows. The II chapter will introduce related work about PUCF. The chapter III mainly introduces processes of PUCF. The IV chapter finds the most effective model parameters through experiments and compares them with the traditional collaborative filtering algorithms. Finally, in chapter V , we summarize and look forward to the PUCF model.

978-0-7381-3170-2/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: MIT-World Peace University. Downloaded on March 30,2023 at 04:38:24 UTC from IEEE Xplore. Restrictions apply.

II. Related Work

As we all know, collaborative filtering algorithm is the most widely used personalized recommendation system. Goldberg [2] et al. first proposed the concept of collaborative filtering in 1992. The proposal of the algorithm caused a great sensation in the academic and industrial circles of recommender systems. So far, improved versions of collaborative filtering algorithms have sprung up [3-4]. The core algorithm of traditional collaborative filtering lies in the calculation method of similarity between users (or items). Next, we introduce the commonly used similarity calculation methods.

Cosine similarity constructs a vector based on the user's rating of the item, and calculates the inner product of the two vectors as the similarity between two users. Through cosine similarity, we can quantify the similarity between users. The higher the cosine similarity, the more similar the two users. Pearson correlation-based similarity[5-6] uses correlation to replace cosine similarity, which introduces the concept of negative correlation, and the value range is between [-1,1], The above-mentioned two users whose cosine similarity degree is 0 may actually be negatively correlated. Through the Pearson similarity calculation method, the similarity calculation method between users can be more accurately quantified.

In addition, in order to solve the problem of data sparseness and accurately portray user portraits, Pirasteh[6] et al. proposed the AC-PCC method, which considered other features when calculating user similarity. On the one hand, it alleviates the sparsity problem of the score matrix, and on the other hand, it more accurately characterizes user interest. Liu [8] et al. proposed a new similarity calculation method (NHSM). This calculation method uses contextual feature information. For the recommended target user, the method considers not only the rating matrix information, but also the context of user ratings, such as time, location, weather, etc. The traditional collaborative filtering method usually calculates the similarity between two users, and then selects the K most similar users. The KNN algorithm [9-12] is based on the K nearest neighbor users to sort the item prediction scores.

The problem that the recommendation algorithm solves is nothing more than matching the characteristics of the user's interest with the characteristics of the items, and the items with a high degree of matching are first recommended. In order to reduce the computational complexity of data, Jaimes [13] et al. proposed to match the user tag to the item tag (the tag includes the extraction of category, entity and other information), which greatly reduces the computational complexity.

In addition, many scholars have considered the influence of time on recommendation accuracy. Wang [14] proposed the concept of time decay. As the behavior time goes on, the influence of the behavior on the user portrait becomes smaller, so the time decay factor is used to control the influence of behavior time on user interest.

Finally, American scientist Edwin Bidwell Wilson proposed Wilson Confidence Interval in 1927 [15-16]. This formula effectively solves the credibility of small samples, thereby increasing the robustness of the model.

III. Collaborative Filtering Recommendation Algorithm Based on User Interest Probability

In this chapter, we will elaborate on the collaborative filtering recommendation algorithm based on user interest probability. This chapter mainly describes in detail from three aspects: the mathematical symbols, basic principles and implementation process of the collaborative filtering recommendation algorithm based on user interest probability.

A. Symbols

First of all, in order to facilitate readers to better understand the collaborative filtering recommendation algorithm based on user interest probability, we briefly introduce the meaning of the mathematical symbols used in the model. The specific mathematical symbols are shown in Table I.

TABLE I. Symbols of PUCF

Mathematical symbols	Description
C	Collection of categories
B	Collection of behaviors
I	Collection of items
U	Collection of users
wil_p	Wilson's Confidence Value of Probability p
t_{uij}	The time decay value corresponding to the user's(u) behavior(j) on the item in category i
$p_u(C_i^u/B_j^u)$	Probability of user's(u) behavior on category i

B. Principle of Collaborative Filtering Model Based on User Interest Probability

Above we elaborated on the meaning of the mathematical symbols used in this article, and then we will elaborate on the principle of the collaborative filtering model based on the probability of user interest. First of all, in order to prevent insufficient confidence caused by insufficient data samples, we introduced Wilson confidence interval in the model. The specific formula is as follows:

$$wil_p = \frac{p+\frac{1}{2n}z^2_{1-\frac{\alpha}{2}} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}+\frac{z^2_{1-\frac{\alpha}{2}}}{4n^2}}}{1+\frac{1}{n}z^2_{1-\frac{\alpha}{2}}} \quad (1)$$

Among them, wil_p represents the Wilson confidence value of probability p, n represents the sample size, α represents the confidence level, and $z_{1-\frac{\alpha}{2}}$ represents the z statistic under the confident level, and this value is generally a constant.

As times goes by, the behaviors of users have less reference significance for establishing user interest models. Therefore, we have introduced time decay factors into the model. The specific calculation formula is as follows:

$$t_{uij} = e^{\alpha * \frac{T_c - T_{uij}}{T_c - \max(T_B)}} \quad (2)$$

Among them, t_{uij} represents the time attenuation factor of the behavior j of the user u on the item of the category i, α represents the time attenuation factor, and T_c represents the current time or the time when the behavior last occurred. This

Authorized licensed use limited to: MIT-World Peace University. Downloaded on March 30,2023 at 04:38:24 UTC from IEEE Xplore. Restrictions apply.

value needs to be determined according to the data set. T_{uij} represents the time when the user u has the behavior j on the item of the category i, and $max(T_B)$ represents the earliest time in the behavior.

Next, according to formula 2, we can calculate the interest of the user u for the category i after the behavior j occurs, that is, the conditional probability. The specific calculation formula is as follows:

$$p_u\left(C_i^u / B_j^u\right) = \frac{\sum_{x \in C_i^u} t_{uxy}}{\sum_{y \in B_j^u} t_{uy}} \quad (3)$$

Among them, $p_u\left(C_i^u / B_j^u\right)$ represents the probability that the user u has the behavior j on the item of the category i, and t_{uxy} represents that the attenuation factor of the user u has the behavior y on the item of the category x, t_{uy} represents the attenuation factor of the user u with the behavior y.

Therefore, according to formulas 1 and 3, the conditional probability of the user u to the category i is calculated. The specific computed process is as follows:

$$p_u(C_i) = \sum_{B_j^u \in B^u} wil_{p_u\left(C_i^u / B_j^u\right)} * w_{B_j} \quad (4)$$

Among them, $p_u(C_i)$ represents the interest of the user u for the category i, $wil_{p_u\left(C_i^u / B_j^u\right)}$ represents the Wilson confidence value of $p_u\left(C_i^u / B_j^u\right)$, w_{B_j} represents the calculated weight of the behavior j.

Due to the high time complexity of traditional similarity calculation, we have improved the similarity calculation method in combination with formula 4. The specific calculation formula is as follows:

$$sim\left(u_i, u_j\right) = \frac{\sum_{x \in C} p_{u_i}(c_x) . p_{u_j}(c_x)}{\|C\|} \quad (5)$$

Among them, $sim\left(u_i, u_j\right)$ represents the similarity between the user u_i and the user u_j, and $\|C\|$ represents the number of categories. Through the above formula, we can see that the computational complexity of the user similarity drops from by $\|U\|$ or $\|I\|$ to$\|C\|$.

Finally, according to formulas 4 and 5, we can calculate the rating of the user u on the item i. The specific formula is as follows:

$$\widehat{r_{ui}} = \sum_{x \in C_i} p_u(C_x) + \overline{r_u} + \frac{\sum_{v \in U_1} sim(u,v)(\overline{r_{vi}} - \overline{r_v})}{\sum_{v \in U_1} |sim(u,v)|} \quad (6)$$

Among them, ($\widehat{r_{ui}}$ represents the predicted score of the user u for the item i, and U_1 represents the set of the most similar users to the user u.

C. The process of collaborative filtering based on the user interest probability

Above, we elaborated on the relevant theoretical knowledge of collaborative filtering based on the user interest probability (PUCF). Next, we will further explain the process of PUCF , as is shown in the table II below.

TABLE II. THE PROCESS OF PUCF

Input: user u and the user behavior $Record$

Output: the recommendation $Result$ of the user u

1. $Result = \emptyset,\ N_u = \emptyset,\ P_u(C) = \emptyset,\ Orderlist = \emptyset$
2. $PreprocesInputDataSet(Record)$
   $< userid, itemid, action\_time, cateid, action\_type >$
3. $P_u(C) = GetCateProbalityUser(u)$
4. $N_u = GetKSimilarUser(u)$
5. for $\forall\ i \in I$ do
6. $\widehat{r_{ui}} = \sum_{x \in C_i} p_u(C_x) + \overline{r_u} + \frac{\sum_{v \in N_u} sim(u,v)(\overline{r_{vi}} - \overline{r_v})}{\sum_{v \in N_u} |sim(u,v)|}$
7. Insert i into $Orderlist$
8. end for
9. $ReverseSortbyRui(Orderlist)$
10. $Result = GetFirstn(Orderlist)$
11. return $Result$

First, we process the user's behavior record into the data format in Table 2 and calculate interest of the user u in each category according to the formula 1-4. For the Wilson's confidence probability under normal circumstances, the value of the z statistic is 1.96 at the 95% confidence level. According to the formula 5, we calculate the K users who are the most similar to the user u. Then, based on the formula 6, the scores of the candidate item set are calculated, and the scores are sorted from large to small, finally top n items are recommended to the user u.

IV. EXPERIMENT

A. DataSet

Different from the traditional collaborative filtering algorithm, we use our company's cleaned e-commerce data as the experimental data set, which includes 402980 users and 35204 items. The total number is about 5 million. We use the behavior type as the classification basis, and the specific data is shown in the following table:

TABLE III. DETAILED DESCRIPTION OF THE DATA SET

ActionType	Numbers of users	Numbers of items	Numbers of Datasets
rec_show	402980	52235	3636058
view	112908	35204	908097
pre_buy	41084	7007	118288
buy	3940	6039	63675

Consistent with other comparison methods, we use 80% of the experimental data as the training set, and the remaining 20% as the test set.

B. Evaluation Metric

We choose Mean Absolute Error (MAE) to evaluate the experimental effect of PUCF proposed in this article. *MAE* is one of the commonly used evaluation indicators in recommendation systems. Its specific formula is as follows:

Authorized licensed use limited to: MIT-World Peace University. Downloaded on March 30,2023 at 04:38:24 UTC from IEEE Xplore. Restrictions apply.

$$MAE = \frac{\sum_{ij}|r_{ij} - \widehat{r_{ij}}|}{n} \quad (7)$$

Among them, r_{ij} represents the real score of the user u_i on the item j, $\widehat{r_{ij}}$ represents the predicted score of the user u_i on item j and n represents the number of test sets.

C. Experiment Result

It can be seen from formula 6 that the parameters of the collaborative filtering model based on user interest probability proposed in this paper are as follows: the time attenuation factor α in formula 2, the weight $W_{B_j}$of action B_j in formula 4, and the K value in the neighbor KNN algorithm in formula 6 . In formula 1, we normally choose the conventional 95% as Wilson's confidence level. According to Table III, we can see that there are four kinds of behaviors in our experimental datasets, namely exposure (rec_show), view, pre-purchase (pre_buy), and buy. Whether to buy is the final goal of this article. Therefore, $w_{rec_show}$and w_{buy} should take values 0 and 1, respectively. The weights of other behaviors should have the following relationship: $w_{rec\_show} < w_{view} \leq w_{pre\_buy} < w_{buy}$, we use 0.3 as the step length to approximate the optimal solution α、w_{view}、w_{pre_buy}.

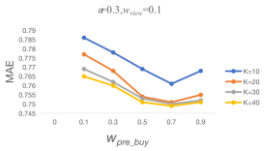

Fig. 1. K and w_{pre_buy} effect MAE.

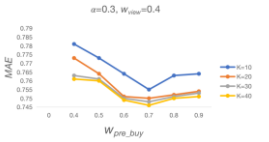

Fig. 2. K and w_{pre_buy} effect MAE.

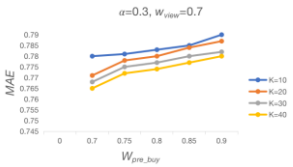

Fig. 3. K and w_{pre_buy} effect MAE.

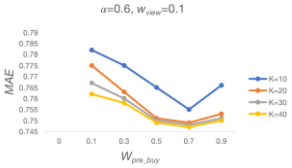

Fig. 4. K and w_{pre_buy} effect MAE.

From Equation 7, we can see that the lower the *MAE* value, the better the effect of recommender systems. From Figure 1 to Figure 3, we can see that when $\alpha = 0.3$, regardless of the value of w_{view}, the *MAE* is always the lowest when $K = 40$, $w_{pre\_buy} = 0$, and when$w_{view} = 0.1$, the *MAE* value is 0.749 . From Figure 4 to Figure 6, we can see that when $\alpha = 0.6$, regardless of the value of w_{view}, the *MAE* value is the lowest when $K = 40$ and $w_{pre\_buy} = 0.7$, and when $w_{view} = 0.4$, the *MAE* value is 0.743 .From Figure 7 to Figure 9, when $\alpha = 0.9$, no matter what value $w_{view} = 0.4$ takes, when $K = 40$ and $w_{pre\_buy} = 0.7$, the *MAE* is always the lowest, and when $w_{view} = 0.1$, the *MAE* value is 0.748. In summary, when $\alpha = 0.6$, $w_{view} = 0.4$, $w_{pre\_buy} = 0.7$,the value of *MAE* is the smallest, that is, the effect of recommender systems is the best.

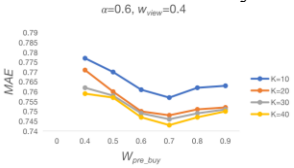

Fig. 5. K and w_{pre_buy} effect MAE

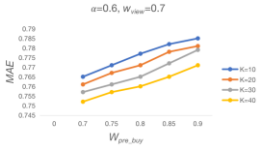

Fig. 6. K and w_{pre_buy} effect MAE

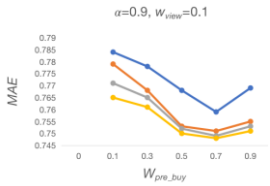

Fig. 7. K and w_{pre_buy} effect MAE.

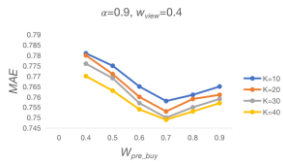

Fig. 8. K and w_{pre_buy} effect MAE.

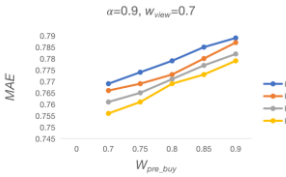

Fig. 9. K and w_{pre_buy} effect MAE.

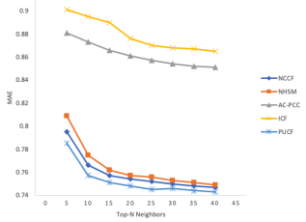

Fig. 10. K and w_{pre_buy} effect MAE.

In order to further illustrate the collaborative filtering model which based on user interest probability (PUCF) proposed in this paper performs better, we choose NCCF, NHSM[8], AC-PCC[7], ICF and PUCF methods for comparison. It can be seen from the implementation that no matter what value K takes, PUCF proposed in the paper is the best, and when K=40, the *MAE* value is the lowest, that is to says, the recommendation effect is the best.

ACKNOWLEDGMENT

Traditional collaborative filtering is one of the common algorithms in current recommendation systems, but it does not consider the confidence of behavior, the influence of time factors on user portraits, and the conversion between behaviors. PUCF proposed in this paper uses the Wilson confidence interval to solve the problem of the probability confidence. It uses a time decay factor to consider the influence of time on user interest bias and finds the optional solution of the weight of the relationship between behaviors through experiments, which considers the relevance of transformations between behaviors. The experiments show that PUCF is better than other comparison methods on the *MAE* metric.

In addition, it is necessary to analyze the experimental results as to find the optimal solution for different data sets. In future, we will introduce machine learning-related algorithms to find the optimal solution of the experimental parameters.

REFERENCES

[1] H. Butcher, "Information overload in management and business," IEE Colloquium on Information Overload, London, UK, 1995, pp. 1/1-1/2, doi: 10.1049/ic:19951426.

[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[3] T. Hofmann, "Collaborative filtering via gaussian probabilistic latent semantic analysis," in *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR)*, pp. 259–266, Toronto, Canada, 2003.

[4] Y. Xiao, P. Ai, C.-H. Hsu, H. Wang, and X. Jiao, "Time-ordered collaborative filtering for news recommendation," *China Communications*, vol. 12, no. 12, pp. 53–62, 2015.

[5] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Beijing, China: China Machine Press, 2010, pp. 65-84

Authorized licensed use limited to: MIT-World Peace University. Downloaded on March 30,2023 at 04:38:24 UTC from IEEE Xplore. Restrictions apply.

[6] Z. L. Yang, Y. W. Song, and et al, “New Sigmoid-like function better than Fisher z transformation,” Communications in Statistics-Theory and Methods, vol. 45, no. 8, pp. 2332-2341, Apr. 2016.

[7] P. Pirasteh, D. Hwang, and J. E. Jung, “Weighted Similarity Schemes for High Scalability in User-Based Collaborative Filtering,” Mobile Networks and Applications, vol. 20, no. 4, pp. 1–11,2014.

[8] H. Liu, Z. Hu, A. Mian, H. Tian, and X. Zhu, “A new user model to improve the accuracy of collaborativefltering,” Knowledge-Based Systems, vol. 56, pp. 156–166, 2014.

[9] L. Y. Dong, Y. Wang, Y. Ren and Y. L. Li, "Collaborative Filter Algorithm Based on Matrix Decomposition and Clustering[J]", Journal of Jilin University (Science Edition), vol. 57, no. 01, pp. 105-110, 2019.

[10] L. Y. Dong, Y. Wang, Y. Ren and Y. L. Li, "Collaborative Filter Algorithm Based on Matrix Decomposition and Clustering[J]", Journal of Jilin University (Science Edition), vol. 57, no. 01, pp. 105-110, 2019.

[11] C. Yang, X. Yu and Y. Liu, "Continuous KNN Join Processing for Real-Time Recommendation," 2014 IEEE International Conference on Data Mining, Shenzhen, China, 2014, pp. 640-649, doi: 10.1109/ICDM.2014.20

[12] B. Wang, Q. Liao and C. Zhang, "Weight Based KNN Recommender System," 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, China, 2013, pp. 449-452, doi: 10.1109/IHMSC.2013.254.

[13] E. Frías-Martinez, M. Cebrián and A. Jaimes, "A Study on the Granularity of User Modeling for Tag Prediction," 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, NSW, Australia, 2008, pp. 828-831, doi: 10.1109/WIIAT.2008.67.

[14] H. Wang, Z. Wang and W. Zhang, "Quantitative analysis of Matthew effect and sparsity problem of recommender systems," 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2018, pp. 78-82, doi: 10.1109/ICCCBDA.2018.8386490.

[15] Agresti, Alan and Brent A. Coull (1998), "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions," The American Statistician, 52, 119-126.

[16] Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," Journal of the American Statistical Association, 22, 209-212.

Authorized licensed use limited to: MIT-World Peace University. Downloaded on March 30,2023 at 04:38:24 UTC from IEEE Xplore. Restrictions apply.