

A Novel Hybrid Collaborative Filtering Approach to Recommendation Using Reviews: The Product Attributes Perspective

Min Cao¹, Sijing Zhou¹, Honghao Gao^{1,2,3,*}, Youhuizi Li⁴

¹, School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

², Computing Center, Shanghai University, Shanghai 200444, China

³, Shanghai Key Laboratory of Computer Software Testing & Evaluating, Shanghai 201112, China

⁴, College of Computer, Hangzhou Dianzi University, Hangzhou 310018, China

mcao@staff.shu.edu.cn, zhousijing@shu.edu.cn, gaohonghao@shu.edu.cn, huizi@hdu.edu.cn

Abstract—The product recommendation research has been focusing on modelling users' reviews to construct the relation of users and products. Thus, the recommended performance can be improved by obtaining virtual ratings from corresponding reviews. However, these perspectives on reviews do not take into account the product field characteristic, which may impact the recommendation performance. To this point, this paper proposes a hybrid collaborative filtering approach to compute the correlation value considering product attributes. First, *Product Attribute Weight* and *Product Attribute Score* are introduced to formalize the product attributes for user and product respectively in a quantitative way. After that, the recommended ranking formula for the new model is presented. Finally, we carry out experimental analysis to show our method can effectively improve the performance of recommendation under a sparseness dataset.

Keywords — *Product Recommendation, Reviews, Hybrid Collaborative Filtering, Product Attributes*

I. INTRODUCTION

The recommender system originated from information retrieval has been served to provide users with personalized online product recommendations to improve user experience [1][2]. With the increasing information on line, the performance of the recommendation using reviews becomes a crucial problem to modern service industry. However, there are some deficiencies in the existing product recommendation approaches using product reviews.

First, the mainstream of recommendation methods using reviews is usually using aspect preference, such as *aspect need* and *aspect importance* [2-4]. The implicit condition in aspect preference methods is reviews' characteristic of centralized features. But the features of product reviews are scattered and not uniform because the reviews have multiple categories and are numerous. Therefore, aspect preference methods bring disunity problem of features and are not suitable for product field.

Second, product attributes is proved that affects consumers' desire for consumption [5-8]. The expression of product

attributes was once focused on the calculation of weight values [2]. Until the introduction of matrix factorization theory, the modeling methods based on the multi-irrelevant-models form began to emerge [3][9-10]. However, the models were generally only based on user. Lack of product perspective, user preference simulation is not comprehensive, which affects recommended performance negatively.

In response to above issues, this paper presents a hybrid collaborative filtering approach based on product attributes – PACF (Product Attributes Collaborative Filtering). To model from two angles of user and product, PAM (Product Attributes Model) based on the matrix factorization's vectors multiplication idea is discussed. After that, important elements of *Product Attribute Weight* and *Product Attribute Score* for the PAM are defined for users and products respectively. It needs applicable formula to construct new model to integrate these factors. Then, a new hybrid collaborative filtering formula Γ_{PAM} is proposed to generate the recommended results for the PAM.

The rest of this paper is organized as follows. Section II reviews related work. Section III shows the formal definition. Section IV introduces the model PAM and the formula Γ_{PAM} for PACF. Section V discusses the experimental analysis, and Section VI presents conclusions and provides future research directions.

II. FORMAL DEFINITION

Integrating the valuable information embedded in reviews not only promotes user experience in the recommender system but also improves the recommended performance [2]. A virtual rating can be generated through users' implicit preference information from reviews.

Aimed at characteristics of product reviews, product attributes are fixed to facilitate feature consistency first. Then, sentiment polarity is introduced to achieve accurate user preferences. Product attributes can be defined including quality, performance, appearance and other aspects. Positive polarity and negative polarity are embodied in sentiment polarity. The

specific symbols are clearly defined in Table I. The product attributes parameter is fixed through building the dictionary, shown in Table II.

TABLE I. SYMBOL AND ITS MEANING

Symbol	Meaning
$R=\{r_1, r_2, \dots, r_{ R }\}$	Reviews set
$U=\{u_1, u_2, \dots, u_{ U }\}$	Users set
$P=\{p_1, p_2, \dots, p_{ P }\}$	Products set
$PA=\{pa_1, pa_2, \dots, p_{ pa }\}$	Product attributes set. The specific definition is shown in Table II.
$F_k = \{f_1, f_2, \dots, f_{ F_k }\}$	F_k is a set of feature words f_i of the product attribute pa_k . The specific definition shown in Table II.
$\delta_i \in \{1, -1\}$	δ_i is the sentiment polarity corresponding to the product attribute feature word f_i . Among the set, -1 is negative sentiment, and 1 is non-negative sentiment (including positive and neutral).

TABLE II. PRODUCT ATTRIBUTES AND FEATURE WORDS

Symbol	Meaning
PA	$PA=\{Quality, Service, Performance, Package\}$
$F_{Quality}$	$F_{Quality}=\{nature, product, greener, brand .etc\}$
$F_{Service}$	$F_{Service}=\{communication, efficient, responsive .etc\}$
$F_{Performance}$	$F_{Performance}=\{fresh, flavor, awful, tasted .etc\}$
$F_{Package}$	$F_{Package}=\{delivery, ship, on-time, speed .etc\}$

Accordingly, the research problem is to tackle the following challenges. First, how to reliably model inference user preferences from two-tuples (pa_k, δ_i) ? Second, how to effectively incorporate product attributes information to generate recommendation results? Another problem is to establish the relevance of users' model and products' model.

III. PRODUCT ATTRIBUTES COLLABORATIVE FILTERING

This section presents the recommended framework of PACF shown in Fig 1. After data preprocessing, a collector of two-tuples (pa_k, δ) in reviews is obtained.

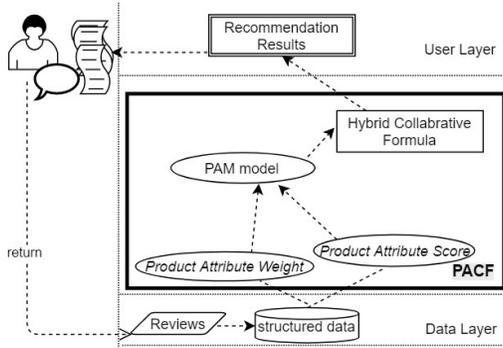


Fig. 1. The recommendation framework of PACF

Then, a novel recommendation method is used to generate recommendations. We formalize PACF with the PAM model based on reviews and a hybrid collaborative filtering formula Γ_{PAM} . *Product Attribute Weight* proposed in PAM characterizes the user's weight while *Product Attribute Score* describes the product's score. The specific modeling is detailed in Section A. Learning the idea of a hybrid collaborative filtering approach, Section B proposes a new formula Γ_{PAM} for the PAM. The premise of the formula is to make *Product Attribute Weight* relate to *Product Attribute Score*.

Finally, the recommendation results are generated through the formula.

A. Product Attribute Model based on Reviews

Users and products are modeled separately through the matrix factorization idea [9]. From the perspective of the user's weight and the product's score, PAM is subdivided into *Product Attribute Weight* and *Product Attribute Score*.

1) Product Attribute Weight Analysis

The first measure, *Product Attribute Weight* recorded as W , is the degree of attention given to the attributes of the product. Given a user u_i and a product attribute pa_k , the formula *Product Attribute Weight* is defined as follows.

$$W(u_i, pa_k) = \frac{|F_{ik}|}{|F_i|} = \frac{|\delta_{ik}|}{|\delta_i|} = \frac{\sum_{p_j \in R_i} |\delta_{ijk}|}{\sum_{p_j \in R_i} |\delta_{ij}|} \quad (1)$$

In formula (1), R_i is the set of u_i 's reviews; $p_j \in R_i$ represents the products which the user reviewed; $|F_{ik}|$ represents the frequency that the feature word is mentioned by the user with respect to product attribute pa_k ; $|F_i|$ represents the number of times that all product attributes' feature word is mentioned by the user; δ indicates the sentiment polarity value. $|\delta_{ijk}|$ is the user's sentiment polarity set of product attribute pa_k for product $p_j \in R_i$. The number of user comments on product attribute pa_k is $|\delta_{ik}|$. $|\delta_{ij}|$ represents the user's sentiment polarity set for $p_j \in R_i$ on all product attributes PA . $|\delta_i|$ is the frequency of all product attributes PA in the comments. When the value of W is zero or unknown, *Product Attribute Weight* is 0.1.

2) Product Attribute Score Analysis

The second measure, *Product Attribute Score*, is similar to the user rating the project. The user's views of product attributes are scored and recorded as S . Given a product p_j and a product attribute pa_k , the measure *Product Attribute Score* can be defined as follows.

$$S(p_j, pa_k) = \frac{\sum_{u_i \in R_j, \delta_{ijk}=1} |\delta_{ijk}|}{\sum_{u_i \in R_j} |\delta_{ijk}|} = \frac{\sum_{u_i \in R_j, \delta_{ijk}=1} |\delta_{ijk}|}{|\delta_{jk}|} \quad (2)$$

Among the formula (2), R_j is the reviews set for the product p_j ; u_i is the user who has commented on it; and δ indicates the sentiment polarity value. δ_{ijk} is the user's sentiment polarity set of product attribute pa_k . $\sum_{u_i \in R_j, \delta_{ijk}=1} |\delta_{ijk}|$ expresses the size of

the sentiment polarity set when $\delta_{jk}=1$. When the value of S is zero or unknown, *Product Attribute Score* is 0.1.

3) Product Attributes Model

Draw on the experience of the multiplication form of matrix factorization, the PAM model is divided into different models corresponding to users and products. With the two formulas proposed above, PAM can be defined:

$$PAM(u_i, p_j, PA) = \begin{cases} (W_1, W_2, \dots, W_{|PA|}) & W_k = W(u_i, pa_k) \text{ for users} \\ (S_1, S_2, \dots, S_{|PA|}) & S_k = S(p_j, pa_k) \text{ for products} \end{cases} \quad (3)$$

B. Hybrid Collaborative Filtering Formula for PAM: Γ_{PAM}

The vectors corresponding to users and products separately in the PAM are unrelated. In order to solve the problem, the average *Product Attribute Score* of users recorded as \bar{S} is introduced based on the shopping history in reviews. The average *Product Attribute Weight* of the product is calculated similarly and denoted as \bar{W} . The formula is shown as follows.

$$PAM'(u_i, p_j, PA) = \begin{cases} (W_1, W_2, \dots, W_{|PA|}) \xrightarrow{\text{reviews}} (\bar{S}_1, \bar{S}_2, \dots, \bar{S}_{|PA|}) & \text{for users} \\ (S_1, S_2, \dots, S_{|PA|}) \xrightarrow{\text{reviews}} (\bar{W}_1, \bar{W}_2, \dots, \bar{W}_{|PA|}) & \text{for products} \end{cases} \quad (4)$$

Next, the generalization formula (5) we proposed is extracted for calculating the hybrid collaborative filtering [1][2][11-13]. A new hybrid collaborative filtering formula (6) for the PAM is obtained by derivation the formula (4) through combining the formula (5) and the cosine formula.

$$\Gamma_{HCF} = UBCF \times IBCF \quad (5)$$

$$\Gamma_{PAM}(u_i, p_j) = \frac{\sum_1^{|PA|} (W_i \times \bar{W}_j)}{\sqrt{\sum_1^{|PA|} W_i} \times \sqrt{\sum_1^{|PA|} \bar{W}_j}} \times \frac{\sum_1^{|PA|} (\bar{S}_i \times S_j)}{\sqrt{\sum_1^{|PA|} \bar{S}_i} \times \sqrt{\sum_1^{|PA|} S_j}} \quad (6)$$

For the current user, the cosine \cos_{UBCF} is obtained by the vector $(W_1, W_2, \dots, W_{|PA|})$ and the average user weight value $\bar{W} = (\bar{W}_1, \bar{W}_2, \dots, \bar{W}_{|PA|})$. It is the same way to figure out the value for \cos_{IBCF} .

IV. EXPERIMENTS

To evaluate the performance of PACF, experiments are carried out and compared by analyzing our method against other algorithms through offline dataset.

A. Dataset

We used the Amazon fine-food reviews dataset from SNAP. The dataset collected 568,454 reviews posted by 256,059 users for 74,258 items of food [14]. The format contains the *UserId*, the *ProductId*, the *Text*, and the *Score* attributes, which are

required for the experiment. The *Score* is an integer from 1 to 5. In large shopping sites, the number of users and products increases daily. Meanwhile, the dataset of actual purchases is sparse, usually below 0.1%. The datasets are described in Table III.

TABLE III. DESCRIPTIVE STATISTICS OF DATASETS

Dataset	Descriptive Statistics			
	Num. of users	Num. of products	Num. of ratings and reviews	Sparsity
Data1	1232	754	1250	0.1346%
Data2	2420	1193	2500	0.0866%
Data3	4719	1791	5000	0.0592%
Data4	9051	1765	10000	0.0626%
Data5	17139	3148	20000	0.0371%

B. Experimental Result and Evaluation

The rating data have a sparseness of less than 0.1% on shopping sites. Furthermore, the product category is in billions of units. Thus, using TOP-N sorting to evaluate accuracy is not an appropriate method. Our experiment is to simulate the data from actual shopping sites. The extremely low accuracy of this site does not have a value.

In this case, *coverage* is more appropriate. The *coverage* is used to measure the ability of the methods to discover products. The PACF method contains the idea of collaborative filtering and matrix factorization. Therefore, the following comparison is considered: UBCF, IBCF and SVD [15]. UBCF and IBCF are classical algorithms for collaborative filtering; SVD is a representative algorithm for matrix factorization. Firstly, we take the datasets in Table III as the experimental dataset. Then, the evaluation metric *coverage* is calculated under $N = 1, 5, 10$ and 20. The experimental results are shown in Table IV. The *coverage* unit is %.

TABLE IV. EXPERIMENTAL RESULTS OF COVERAGE

Dataset	Methods	Coverage			
		N=1	N=5	N=10	N=20
Data1	UBCF	0.1326	0.0119	0.0146	0.0291
	IBCF	0.2652	0.0093	0.0172	0.0305
	SVD	0.6631	0.0186	0.0358	0.0650
	PACF	1.3263	0.0597	0.0941	0.1552
Data2	UBCF	0.0008	0.0042	0.0092	0.0176
	IBCF	0.0017	0.0050	0.0101	0.0192
	SVD	0.0117	0.0268	0.0360	0.0762
	PACF	0.0142	0.0386	0.0695	0.1215
Data3	UBCF	0.0006	0.0036	0.0430	0.0122
	IBCF	0.0017	0.0061	0.0089	0.0151
	SVD	0.0168	0.0329	0.0061	0.0642
	PACF	0.0101	0.0274	0.0519	0.0966
Data4	UBCF	0.0017	0.0045	0.0073	0.0130
	IBCF	0.0017	0.0062	0.0102	0.0164
	SVD	0.0221	0.0567	0.0771	0.1082

Dataset	Coverage				
	Methods	N=1	N=5	N=10	N=20
Data5	PACF	0.0096	0.0306	0.0515	0.0816
	UBCF	0.0004	0.0016	0.0030	0.0048
	IBCF	0.0006	0.0022	0.0042	0.0076
	SVD	0.0072	0.0198	0.0358	0.0550
	PACF	0.0056	0.0150	0.0260	0.0398

C. Discussion

The overall *coverage* of PACF is on the rise in Table IV. PACF depends on the number of reviews. The larger the reviews, the better the performance. PACF also applies to scenarios where SVD is suitable. In Data1, Data2 and Data3 of Table IV, PACF performs better than SVD. In Data4 and Data5 of Table IV, PACF's *coverage* is lower than SVD's. In further analysis, PACF's formula is similarly affected by the purchase record. With a sparsity of 0.05%, PACF sacrifices coverage. Moreover, Table IV illustrates that when sparseness is higher than 0.05% , the *coverage* performance of our proposed PACF is good.

V. CONCLUSION AND FUTURE WORK

Considering the characteristics of product reviews, this paper uses constant product attributes and establish irrelevant and multi-perspective models. The objective is to solve the problem of complicated product reviews but also to refine information on user preferences. Based on the above conditions, a hybrid collaborative filtering method PACF is proposed. The PAM model and the Γ_{PAM} formula are constituted to make PACF. PAM consists of *Product Attribute Weight* and *Product Attribute Score*. The perspectives of users and products can effectively simulate user preferences. The experiments have showed that PACF achieved better recommended performance in dealing with large and sparse reviews and predicted user behavior well. The coverage is superior to other methods at a sparsity higher than 0.05%.

In the future, we will study the current issue, PACF sacrifices coverage when sparsity is less than 0.05%. Using user relation and social information from other platforms can further improve the performance of the recommendation on sparse data. Moreover, the implementation and effective response of large-scale electronic website platform is worthy of further work. To this point, cloud computing and cluster will be considered to accelerate the reaction speed.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable and constructive comments. This work is supported by The National Key Research and Development Plan of China under Grant No. 2017YFD0400101, and The National Natural Science Foundation of China under Grant No. 61502294, 61572306.

REFERENCES

- [1] Lu J., Wu D. S., Mao M. S., Wang W. and Zhang G. Q., "Recommender system application developments: A survey", *Decision Support Systems*, 2015, 74 , pp.12-32.
- [2] Chen L., Chen G. L. and Wang F., "Recommender systems based on user reviews: the state of the art", *User Modeling And User-adapted Interaction*, 2015, 25(2), pp. 99-154.
- [3] Ma Y., Chen G. Q. and Wei Q., "Finding users preferences from large-scale online reviews for personalized recommendation", *Electronic Commerce Research*, 2017, 17(1), pp. 3-29.
- [4] T. Sangeetha, N. Balaganesh and K. Muneeswaran, "Aspects based opinion mining from online reviews for product recommendation", *ICCIDS*, 2017, DOI: 10.1109/ICCIDS.2017.8272657.
- [5] M. Vamsee Krishna Kiran, RE Vinodhini, R. Archanaa and K. Vimalkumar, "User specific product recommendation and rating system by performing sentiment analysis on product reviews", *ICACCS*, 2017, DOI: 10.1109/ICACCS.2017.8014640.
- [6] ZKA Baizal, A. Iskandar and E. Nasution, "Ontology-based recommendation involving consumer product reviews", *ICoICT*, 2016, DOI: 10.1109/ICoICT.2016.7571890.
- [7] Alton Y.K. and C.S. Banerjee, "Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality", *Computers in Human Behavior*, 2016, 54, pp.547-554.
- [8] Su J.K., E. Maslowska and E. C. Malthouse, "Understanding the effects of different review features on purchase probability", *International Journal of Advertising*, 2018, 37, Issue 1: Electronic Word-of-Mouth
- [9] Zhao W. X., Wang J. P., He Y. L. and et al, "Mining Product Adopter Information from Online Reviews for Improving Product Recommendation", *ACM Transactions on Knowledge Discovery from Data*, 2016, 10(3), article num.29.
- [10] Z.W. Yu, H. Xu, Z. Yang and B. Guo, "Personalized Travel Package with Multi-Point-of-Interest Recommendation based on Crowdsourced User Footprints", *IEEE Transactions on Human-Machine Systems*, 2016, 46(1), pp.151-158.
- [11] Koren Y., Bell R. and Volinsky C., "Matrix Factorization Techniques for Recommender systems", *Computer*, 2009, 42(8), pp. 30-37.
- [12] Hammou B. A. and Lahcen A. A., "FRAIPA: A fast recommendation approach with improved prediction accuracy", *Expert Systems with Applications*, 2017, 87, pp. 90-97.
- [13] Kassak O., Kompan M. and Bielikova M., "Personalized hybrid recommendation for group of users: Top-N multimedia recommender", *Information Processing & Management*, 2016, 52(3), pp. 459-477.
- [14] [Online].Avliable:<http://online.cambridgecoding.com/notebooks/eWReNYcAfB/implementing-your-own-recommender-systems-in-python-2>
- [15] [Online].Avliable:<http://snap.stanford.edu/data/web-Amazon.htm>