Dynamical Rating Prediction with Topic Words of Reviews: A Hierarchical Analysis Approach

Huibing Zhang[1], Hao Zhong[1], Qing Yang[2(✉)], Fei Jia[1], Ya Zhou[1], and Fang Pan[3]

[1] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
zhanghuibing@guet.edu.cn

[2] Guangxi Key Laboratory of Automatic Measurement Technology and Instrument, Guilin University of Electronic Technology, Guilin 541004, China
gtyqing@hotmail.com

[3] Teaching Affairs Office, Guangxi Normal University, Guilin 541004, China

Abstract. Social commerce is an important part of the social network which contains a large number of user behaviors and user relationships. Users generate reviews, social relations, user-product or product-product mapping information that can reflect an evolution of product characteristics and user preferences in using social commerce. It is a popular topic by using these information to conduct rating prediction in the field of intelligent recommendation. In this paper, optimizing the rating prediction based on topic analysis in two aspects. On the one hand, in the process of data preprocessing, constructing a dynamic hierarchical tree of topic words (DHTTW), which can not only capture the change of users' preferences for product property, but also reflect the impact of different product property on users' preferences at the same time. Based on DHTTW, designing the mapping rules from user reviews to DHTTW to generate user preference vectors. On the other hand, in the process of prediction, proposing a prediction method named combination of gradient boosting decision tree and multi-class linear regression (GBDT-MCLR), which further improves the accuracy of rating prediction.

Keywords: Social commerce · Reviews · Rating prediction · Dynamic Hierarchical Tree of Topic Words · Multi-Class Linear Regression

1 Introduction

Social commerce is an integration of e-commerce and social networks, which is a developmenttendencyinthefuture. As an important part of the Internet of Things, social commerce achieves the deep integration of users and products by using user behavior (for example reviews) and user relationship. Reviews

X. Wang et al. (Eds.): CollaborateCom 2019, LNICST 292, pp. 780–798, 2019.
https://doi.org/10.1007/978-3-030-30146-0_52

in social commerce websites are users' evaluations of quality, performance, and price about products in a specific space–time environment, which reflect the basis of user rating and preference slightly. Therefore, reviews can be used to predict user rating and recommend certain products to them that satisfy their preferences. Here, the key is to extract topic features from reviews and achieve a mapping between topic features and rating value. The mapping mechanism must be established between reviews and the dynamic hierarchical relationship of topic words to describe an accurate meaning of reviews well under a specific space-time condition.

Traditional methods represent user preferences and product property by extracting topic features from reviews, which generate a topic distribution of reviews by performing topic analysis on reviews. And then obtain the relationship between each topic and real rating by using prediction model. However, in the process of data preprocessing, these methods only consider the topic distribution of reviews [1,2] but ignore the dynamic changes in the probability of topic words in different time windows and lack the description of hierarchical relationship between topic words. Thus, they can neither adapt well to the change in user preferences nor describe the effect of different properties of products on user rating. In the process of user rating prediction, using LR (linear regression), GBDT (gradient boosting decision tree), RF (random forest) and other prediction algorithms to predict. Based on recent research [3,4], a rating prediction method is proposed in this work by using DHTTW and GBDT-MCLR. The main contributions of the paper as follows:

(1) A review–preferences dynamic mapping method based on time windows was designed. On the basis of dynamic topic model (DTM) [5], excavating the potential change rule of topic words in different time windows, and exhibiting the evolution of user preference for product property by the change in probability of topic words for a timely rating prediction.
(2) A DHTTW constructing method of topic words of reviews was proposed. On the basis of dynamic changes in topic words, fusing the similarity and intensity of mutual information between topic words in a specified time window to establish hierarchical relationship. That is, a topic generated different hierarchical trees of topic words in different time windows, such that the hierarchy of topic words could dynamically represent the effect of topic words on user rating.
(3) A method for generating user preference vector based on DHTTW of topic words was proposed. Reviews were mapped to DHTTW in a specified time window to generate a topic vector of reviews. Using the vector to represent user preferences such that all reviews in different time windows were mapped to a vector space with the same dimension.
(4) A prediction method named GBDT-MCLR is proposed. In view of the discreteness of rating data [6], There is still much improvement when using GBDT-LR for predicting. Before rating prediction based on user preference vectors, clustering all preference vectors. Based on the idea of regression, generating a fitting function in each class. So that the GBDT-LR can adapt to the discrete data to a certain extent.

The remainder of this paper is organized as follows. Section 2 introduces relevant research. Section 3 describes the meaning of a dynamic analysis of topic words, constructs a hierarchical tree, and summarizes the process of the model and algorithm. Section 4 presents the construction details of the dynamic hierarchical tree of topic words, the mapping of user reviews to the tree structure, the generation rules of user preference vectors, and the improvement method of GBDT+LR prediction model. Section 5 conducts an experimental analysis of real datasets. Finally, Sect. 6 presents the conclusion of this work.

2 Related Works

There are two ways of recommendation for users: based on user location information [7–9] and user rating. Traditional methods for rating prediction in an intelligent recommendation system analyze a user's historical rating behavior and predict user rating on unrated products through a collaborative filtering method [10] without analyzing a user review text. With the development of topic discovery [11], sentiment analysis [12], user opinion mining [13], and other technologies of word prediction [14,15]. For example, Tang et al. [16] generates user preference vectors by analyzing the sentimental intensity of review texts, and predicts user ratings by combining the neural network prediction model. Seo et al. [17] uses convolutional neural network to analyze the features of user reviews, correlates users and product according to the features of user reviews by using matrix decomposition method, and finally makes rating prediction. In the research of this paper, the focus is topic analysis technology on user reviews, a rating prediction based on text topic discovery has become the focus of research in recent years.

Ma et al. [18] used LDA (Latent Dirichlet Allocation) model to conduct topic analysis on reviews, generate topic words, calculate a distribution probability of each topic word, manually annotate the sentiment intensity of topic words, generate corresponding word vector in accordance with topic words in reviews, and predict user rating. Ji et al. [19] considered a structural information among users, reviews, and products to propose a topic propagation model on the basis of user–review–product structure for describing user characteristics, products properties, and finally predicting user rating on the basis of random walk. Fang et al. [20] proposed a topic gradient descent model to conduct a topic analysis by using LDA model. The characteristic of topic was expressed by the probability distribution of topic words, and a latent factor was assigned to each topic. The latent factor was dynamically assigned in accordance with the proportion of topic in user review set. Finally, rating was predicted in accordance with the performance value of reviews on each topic. Zhang et al. [21] argued that if topic of review is limited to review text it cannot fully reveal the complex relationship between reviews and ratings. Thus, they proposed a method for integrating a topic and latent factor model, which enables them to complement each other linearly during user rating prediction to improve the accuracy of prediction. McAuley et al. [22] proposed an HTF (Hypersonic Tunnel Facility) model to explore the hidden relationship between user rating and reviews, which used LDA model to analyze all reviews published by each user and all reviews for each product. So it obtained characteristic matrixes of each user and product, and finally input the two characteristic matrixes into SVD (Singular Value Decomposition) model to obtain predicted value of user rating on the product. Zhang et al. [23] conducted a topic analysis based on HTF model. It represented reviews as a set of topic vectors in accordance with the topic words, and normalized these topic vectors to obtain the characteristics of each user and product. Then, it predicted user ratings of products on the basis of each vector and its corresponding rating by using three models: RF, LR, and GBDT. During the experiment, the RMSE value of score prediction was the smallest when LR was used, and the MAE value of score prediction was the smallest when GBDT was used. Therefore, the combination of GBDT and LR has become another focus of research.

Blei et al. [24] first proposed the concept of hierarchical topic, but did not consider the hierarchical relationship between the topic words. None of the above mentioned studies has considered the dynamic changes in the probability of topic words and the hierarchical relationship between topic words. Consequently, these studies have failed to dynamically adapt to the change in user preference and distinguish the effect of different topic words on user rating, thereby leading to a certain amount oferror in predicting user rating. Paranjpe et al. [25] first proposed the method of combining GBDT and regression model. Gupta et al. [26] applies the algorithm to CTR (Click Through Rate). When combining GBDT and regression model, Wang et al. [27] spliced the extracted features with the original features to increase the dimension of features, thereby reducing the prediction error. However, the prediction effect of LR on discrete data is still unsatisfactory.

3 Model and Algorithm

The definitions of relevant symbols are displayed in Table 1.

Table 1. Relevant symbol definitions in the present work.

Reviews	$R = \{R_1, R_2, ..., R_m\}$
Topics	$T = \{T_1, T_2, ..., T_K\}$
Topic words	$W_i = \{W_{i1}, W_{i2}, ..., W_{iN}\}$
Time windows	$t = \{t_1, t_2, ..., t_n\}$
Hiberarchy of words	$H_{t_n,i} = \{H_{t_n,i1}, H_{t_n,i2}, ..., H_{t_n,iN}\}$
Preference vector	$U^m = \{U_1^m, U_2^m, ..., U_K^m\}$
Divide reviews	$R^1, R^2, ..., R^n$
Ratings	$G = \{G_1, G_2, ..., G_m\}$
New features	$\{U^m, u^m\}$
Classes	$C = \{C_1, C_2, ..., C_l\}$
Final feature	$\{U^m, u^m, C_l\}$

Figure 1 illustrates the process of rating prediction model based on dynamic and hierarchical analysis of topic words.

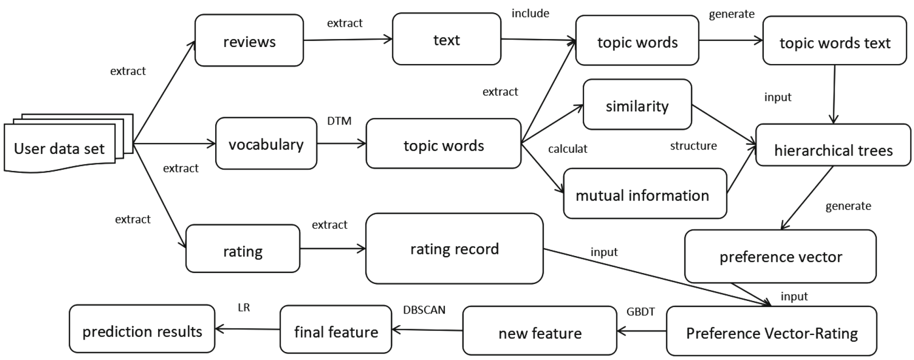


Fig. 1. Process of rating prediction

Firstly, considering the dynamics of user preferences, dividing user reviews set $R = \{R_1, R_2, ..., R_m\}$ into different subsets R^1, R^2, ..., R^n, that correspond to time windows. Using the DTM to generate a uniform set of topic $T = \{T_1, T_2, ..., T_K\}$ and a distribution of topic words $W_i = \{W_{i1}, W_{i2}, ..., W_{iN}\}$ in each time window. The probability value of each topic word in each time window was also calculated. So a change of user preferences for product property could be expressed by probability value changes. Secondly, considering a difference in the effect of each topic word on user rating, a hierarchical tree of topic words was constructed by combining similarity and intensity of mutual information between topic words. Among these words, a hierarchy that corresponds to the set of topic word $W_i = \{W_{i1}, W_{i2}, ..., W_{iN}\}$ in the time window t_n was $H_{t_n,i} = \{H_{t_n,i1}, H_{t_n,i2}, ..., H_{t_n,iN}\}$. Different weights were given to topic words in accordance with their hierarchies. To characterize the effect of topic words on user rating, a deeper hierarchy (a fine granularity) indicated a significant influence on user rating. Finally, user review R_m was mapped to the hierarchical tree of topic words to obtain the number of topic words and average depth, then it calculated the performance value U_K^m of reviews on each topic. A user preference vector $U^m = \{U_1^m, U_2^m, ..., U_K^m\}$ that corresponds to the user review R_m was formed by traversing K topics. Finally, U^m and G were inputted into GBDT for feature analysis, it will generate new feature vectors $\{U^m, u^m\}$, using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to cluster the new feature vectors and get l classes. Finishing linear fitting for each class of reviews $\{U^m, u^m, C_l\}$. The corresponding value of $\{U^m, u^m, C_l\}$ in the fitting function is used as the predicted value of user rating. The work will evaluate the prediction results on the basis of two kinds of errors, namely, mean absolute error (MAE) and root mean square error (RMSE). The process is presented in Algorithm 1.

Algorithm 1. Rating Prediction based on Dynamic and Hierarchical Analysis of the Topic Words of Reviews.

Input: user dataset
Output: MAE and RMSE of prediction results
1: Divide(R,n)//Division of review set
2: DTM(R^1,R^2,...,R^n,K,N)//Dynamic topic analysis of reviews
3: GetPro(W_{iN},n)//Dynamic analysis of topic words
4: GetTree($W_{i1}, W_{i2}, ..., W_{iN}$)// Hierarchical analysis of topic words
5: GetUser(R_m,$H_{t_n,i}$)//Generate preference vector
6: GBDT(U^m,G)//Feature analysis and processing
7: DBSCAN(U^m,u^m)//Clustering new features generated by GBDT
8: LR(C_1,C_2,...,C_l)//Linear fitting for each class
9: Predic(U^m,u^m,C_l)//Prediction based on new feature and classes of feature

4 Algorithm Design and Implementation

4.1 Dynamic Analysis of Topic Words

The beginning of the research work, preprocessing the set of reviews $R = \{R_1, R_2, ..., R_m\}$, to get the review sets $R^1, R^2, ..., R^n$ in each time window. Then inputing $R^1, R^2, ..., R^n$ into the DTM, obtaining the topic set $T = \{T_1, T_2, ..., T_K\}$ of reviews and the set of topic words $W_i = \{W_{i1}, W_{i2}, ..., W_{iN}\}$ under the ith topic. The change of user preferences in different time windows was described by the dynamic nature of topic words: the probability of a topic word is different in each time window, thereby indicating that the users' concern about the product was dynamic.

For topic i in the time window t_n, $P_{t_n,W_{iN}}$ represents the probability of occurrence of topic word W_{iN}. The calculation method is expressed in Formula (1).

$$P_{t_n,W_{iN}} = C_{t_n,W_{iN}} / \sum_{j=1}^{N} (C_{t_n,W_{ij}}), \tag{1}$$

where $C_{t_n,W_{iN}}$ represents the number of occurrences of the topic word W_{iN} in the time window t_n, the definition of $C_{t_n,W_{iN}}$ is as Formula (2).

$$C_{t_n,W_{iN}} = \begin{cases} C_{t_n,W_{iN}} + 1, W_{iN} \in R^n; \\ C_{t_n,W_{iN}}, W_{iN} \notin R^n. \end{cases} \tag{2}$$

Calculating the probability values of all topic words in each time window by Formula (1) and Formula (2). Thus, the probability distribution of the topic word W_{iN} is presented as follows:

$$P_{n,W_{iN}} = \{P_{t_1,W_{iN}}, P_{t_2,W_{iN}}, ..., P_{t_n,W_{iN}}\} \tag{3}$$

The dynamics of user preference was described by the probability values of topic words in different time windows, such that user rating prediction could reflect the dynamics of user preference in different time windows, thereby enhancing the rating prediction timeliness and authenticity.

4.2 Dynamic and Hierarchical Analysis of Topic Words

The hierarchical relationship between the topic words is determined by similarity and mutual information of the topic words in the specified window. And the DHTTW is constructed accordingly: the larger the mutual information of the topic word, the more likely it becomes the upper layer concept. Therefore, it is necessary to compare the mutual information strength of each topic word to determine the upper and lower position of topic words. At the same time, using the similarity between the topic words as the constraint condition for constructing the hierarchical relationship. So that the topic words with high similarity are distributed in the same branch of the hierarchical structure, while the topic words with low similarity are distributed in different branches of the hierarchical structure. The influence of the topic words on the user's rating is characterized by the hierarchy in DHTTW.

Calculating the intensity of mutual information of each topic word in time window t_n by Formula (6), and the results were ranked in descending order. Obtaining an ordered set of topic words $W_i^{'} = \{W_{i1}^{'}:MI(t_n, W_{i1}^{'})$, $W_{i2}^{'}:MI(t_n, W_{i2}^{'})$, ..., $W_{iN}^{'}:MI(t_n, W_{iN}^{'})\}$ under topic i, and $MI(t_n, W_{i1}^{'}) > MI(t_n, W_{i2}^{'}) > ... > MI(t_n, W_{iN}^{'})$. Selecting the topic word $W_{i1}^{'}$ with the highest intensity of mutual information as the upper concept of hierarchical structure and deleting $W_{i1}^{'}$ from the set $W_i^{'}$. Selecting $W_{i2}^{'}$ as the candidate word of the hierarchical structure. If the relation between $W_{i2}^{'}$ and $W_{i1}^{'}$ satisfied the requirement of Definition 1, then $W_{i2}^{'}$ was added to the hierarchical structure as lower concept of $W_{i1}^{'}$ and was deleted from the set $W_i^{'}$. If the relation between $W_{i2}^{'}$ and $W_{i1}^{'}$ failed to satisfy the requirement of Definition 1, then $W_{i2}^{'}$ remains in $W_i^{'}$, selecting $W_{i3}^{'}$ as the candidate word of the hierarchical structure.

Definition 1. *Discriminating hierarchical relations of topic words W_{ia}, W_{ib} in time window t_n*

(1) In Formula (4), satisfy $SIM(R^n, W_{ia}, W_{ib}) < \alpha$, where α is the tuning parameter.
(2) In Formula (6), satisfy $MI(t_n, W_{ia}) < MI(t_n, W_{ib})$.

According to Definition 1, the hierarchical relationship among the topic words is judged in turn, until the set $W_i^{'}$ is empty. The same method was adopted to construct hierarchical tree for all topics in different time windows. Generating K hierarchical trees in each time window. The hierarchical tree of topic i in the time window t_n was $H_{t_n,i} = \{H_{t_n,i1}, H_{t_n,i2}, ..., H_{t_n,iN}\}$, where $H_{t_n,i1} \neq H_{t_n,i2} \neq ... \neq H_{t_n,iN}$, and $H_{t_1,iN} \neq H_{t_2,iN} \neq ... \neq H_{t_n,iN}$. Therefore, the hierarchies of topic words in hierarchical tree were different, and the hierarchy of the same topic word changed with time. The similarity between two topic words W_{ia} and W_{ib} in topic i in time window t_n is calculated as follows:

$$SIM(R^n, W_{ia}, W_{ib}) = \frac{(E_{W_{ia},R^n} E_{W_{ib},R^n})}{\sqrt{(E_{W_{ia},R^n})^2}\sqrt{(E_{W_{ib},R^n})^2}}, \tag{4}$$

where E_{W_{ia},R^n} represents the space vector formed by TF-IDF (Term Frequency–Inverse Document Frequency) value of topic word W_{ia} in each user review within a set of review R^n; therefore, $E_{W_{ia},R^n} = \{E_{W_{ia},R^n,1}, E_{W_{ia},R^n,2}, ..., E_{W_{ia},R^n,m'}\}$. The element $\mathrm{E}_{W_{ia},R^n,m'}$ of the vector represents TF-IDF value of topic word W_{ia} in the $m^{'}$th review within the set of review R^n. The calculation formula is expressed in Formula (5).

$$E_{W_{ia},R^n,m'} = \frac{F_{W_{ia},R^n,m'}}{\sum_{k=1}^{N} F_{W_{ik},R^n,m'}} log \frac{|R^n|}{|\{j : W_{ia} \in R_j^n\}|}, \tag{5}$$

where $F_{W_{ia},R^n,m'}$ represents the number of occurrence of topic word W_{ia} in the set of user review R^n, $|R^n|$ represents the total number of review texts, and $|\{j : W_{ia} \in R_j^n\}|$ represents the total number of texts containing the word W_{ia}.

Under topic i in time window t_n, the intensity of mutual information of topic word W_{ia} referred to the accumulation of point mutual information between topic word W_{ia} and other topic words. As shown in Formula (6):

$$MI(t_n, W_{ia}) = \sum_{k=1}^{N} PMI(t_n, W_{ia}, W_{ik}) \tag{6}$$

The calculation formula of point mutual information of two topic words is as follows:

$$PMI(t_n, W_{ia}, W_{ib}) = log \frac{P_{t_n,(W_{ia},W_{ib})}}{P_{t_n,W_{ia}} P_{t_n,W_{ib}}} \tag{7}$$

According to Formula (1), $P_{t_n,W_{ia}}$ represented the probability of occurrence of topic word W_{ia} in time window t_n. The probability that the topic words W_{ia} and W_{ib} occurred at the same time window t_n is expressed by $P_{t_n,(W_{ia},W_{ib})}$.

The construction pseudo code of topic words hierarchical tree of under time window t_n is defined in Algorithm 2.

4.3 Construction of User Preference Vector

In time window t_n, the corresponding hierarchy of each topic word in $W_i = \{W_{i1}, W_{i2}, ..., W_{iN}\}$ under topic i was $H_{t_n,i} = \{H_{t_n,i1}, H_{t_n,i2}, ..., H_{t_n,iN}\}$. Topic word W_{iN} was given a weight by using hierarchy $H_{t_n,iN}$. The number of topic words S_{i,t_n} under topic i in the record r of user review set R^n is calculated using Formula (8).

$$S_{i,t_n} = \begin{cases} S_{i,t_n} + 1, \exists W_{iN} \in r; \\ S_{i,t_n}, \exists W_{iN} \notin r. \end{cases} \tag{8}$$

The number of topic words $S_{t_n} = \{S_{1,t_n}, S_{2,t_n}, ..., S_{K,t_n}\}$ under each topic contained in reviews could be obtained by traversing K topics.

The average depth of each user review on hierarchical tree of topic was calculated in accordance with topic words in user review r, which contains the topic

Algorithm 2. Construction algorithm of the hierarchical tree of topic words.

Input: a set of topic words $W_i=\{W_{i1},W_{i2},...,W_{iN}\}$
Output: The hierarchy of topic words $H_{t_n,i}=\{H_{t_n,i1},H_{t_n,i2},...,H_{t_n,iN}\}$ that corresponds to $W_i=\{W_{i1},W_{i2},...,W_{iN}\}$

GET $W_i^{'}$ By Formula (6)
$H_{t_n,i}$=[],*Node*=[]
$H_{t_n,i}$[1]=1,*Node*[1]=$W_{i1}^{'}$
FOR(j=2;j≤N;j++)
  $SIM(R^n, W_{i1}^{'}, W_{ij}^{'})$//By Formula (4)
  IF $SIM(R^n, W_{i1}^{'}, W_{ij}^{'})<\alpha$
    THEN $H_{t_n,i}$[j]=1,*Node*[j]=$W_{ij}^{'}$
END FOR//The first hierarchy is end
FOR(temp=1;temp≤N;temp++)
  IF $H_{t_n,i}$[temp]=1
    THEN M_1=M_1+1
END FOR
$M_1 \rightarrow num$
$M_1 \rightarrow sum$
FOR(j=1;j≤*num*-1;j++)
  index1=*findindex*($W_i^{'}$,*Node*[j])
  index2=*findindex*($W_i^{'}$,*Node*[++j])
  $H_{t_n,i}$[*index1*]=2,*Node*[*index1*]=$W_{iindex1}^{'}$
  FOR(k=*index1*+1;k≤*index2*-1;k++)
    $SIM(R^n, W_{iindex1}^{'}, W_{ik}^{'})$//By Formula (4)
    IF $SIM(R^n, W_{iindex1}^{'}, W_{ik}^{'})<\alpha$
      THEN $H_{t_n,i}$[k]=2,*Node*[k]=$W_{ik}^{'}$
  END FOR
END FOR//The second hierarchy is end
FOR(temp=1;temp≤N;temp++)
  IF $H_{t_n,i}$[temp]=2
    THEN M_2=M_2+1
END FOR
$M_1+M_2 \rightarrow sum$
IF $sum<N$
  $M_2 \rightarrow num$
  Repeat

word set W_i and the corresponding hierarchies of topic words in hierarchical tree $H_{t_n,i}$, as expressed in Formula (9).

$$L_{i,t_n} = \sum_{j=1}^{N} (H_{t_n,ij}(\exists W_{ij} \in r))/S_{i,t_n} \tag{9}$$

The average depth of reviews under the topic hierarchical tree was obtained by traversing K topics, where $H_{t_n,ij}$ represents the hierarchy of topic word W_{ij}

under topic i in time window t_n, and L_{i,t_n} represents the average depth of user review r under the hierarchical tree of topic i.

Based on the number of topic words $S_{t_n} = \{S_{1,t_n}, S_{2,t_n}, ..., S_{K,t_n}\}$ of each topic contained in reviews and the average depth $L_{t_n} = \{L_{1,t_n}, L_{2,t_n}, ..., L_{K,t_n}\}$ of reviews under hierarchical tree, the user review R_m was assumed to be in time window t_n, and the user preference U_K^m for topic K is calculated as follows:

$$U_K^m = e^{L_{K,t_n}} \times ln(1 + S_{K,t_n}), \tag{10}$$

where U_K^m is calculated individually to obtain user preference vector $U^m = \{U_1^m, U_2^m, ..., U_K^m\}$, which corresponds to user review R_m. The method fully considers the different effects of S_{t_n} and L_{t_n} on user preference.

4.4 Rating Prediction Model

The input of the prediction model were user preference–rating set $\{U^m, G_m\}$, where U^m represents preference vector generated by the ith user review, and G_m represents rating value that corresponds to review R_m.

This paper presents a prediction algorithm called GBDT-MCLR. Firstly, GBDT carries out feature analysis on user preference-rating set $\{U^m, G_m\}$ and generates new feature $\{U^m, u^m\}$. The process of element u^m generation of feature $\{U^m, u^m\}$ is as follows:

(1) According to the relationship between feature vector U^m and decision value G_m, GBDT model constructs a specified number of RT(Regression Decision Tree) based on residual learning. Expressing the decision value of each RT by $f(U^m)$.
(2) Suppose the number of RT is q, it will get $u^m = \{u_1^m, u_2^m, u_3^m ..., u_q^m\}$. Each element u_q^m of $\{u_1^m, u_2^m, u_3^m, ..., u_q^m\}$ is calculated by Formula (11).

$$u_q^m = f_q(U^m) \tag{11}$$

Secondly, using DBSCAN algorithm to cluster the set of feature vectors $\{\{U^1, u^1\}, \{U^2, u^2\}, ..., \{U^m, u^m\}\}$, and obtaining the set of classes $C = \{C_1, C_2, ..., C_l\}$. The training process of the MCLR model is as follows:

(1) In each class C_l,
(2) Setting the feature weight vector to W_l, and the error to θ_l
(3) Determining the loss function according to the parameters W_l and θ_l, and obtaining the minimum value of the loss function.
(4) Using the least square method to solve the loss function and getting the minimum value. Determining the parameters W_l and θ_l.

Finally, when predicting the rating based on $\{U^m, u^m\}$, judging the class C_l of $\{U^m, u^m\}$, and calculating the corresponding rating value of $\{U^m, u^m\}$ according to the parameters W_l and θ_l of class C_l. The calculation method is as follows:

$$G_m^{'} = W_l\{U^m, u^m\} + \theta_l \tag{12}$$

5 Experiment and Results

5.1 Data Set and Evaluation Standard

Test data were extracted from Amazon.com. Two product categories, namely, inch-tablet and remote streaming media player, were selected. The corresponding reviews are listed in Table 2.

Table 2. Amount of user review data for different products.

Product name	Product code	Total review number
Inch-Tablet	B00TSUGXKE	74615
Media-Player	B00ZV9RDKK	108930

Note: The data set contains all user reviews for each product from 2015 to 2017. Among them, the product numbered B00TSUGXKE belongs to the product with frequent updates, while the product numbered B00ZV9RDKK updates slowly. Each record in the user data contains user's review and ratings of product. Product codes are used to represent the products in the experiment.

In the present work, the result of rating prediction was evaluated by using MAE, RMSE, Recall and F-score. The formulas of MAE, RMSE, Recall and F-score are presented in Formula (13), Formulas (14), Formulas (15) and Formulas (16).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|(y_i^{'} - y_i)| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i^{'} - y_i)^2} \tag{14}$$

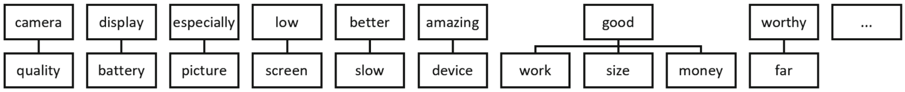
$$Recall = (\frac{m_1^{'}}{m_1} + \frac{m_2^{'}}{m_2} + \frac{m_3^{'}}{m_3} + \frac{m_4^{'}}{m_4} + \frac{m_5^{'}}{m_5})/5 \tag{15}$$

$$F - score = \frac{(\frac{m_1^{'}+m_2^{'}+m_3^{'}+m_4^{'}+m_5^{'}}{m}) * Recall * 2}{(\frac{m_1^{'}+m_2^{'}+m_3^{'}+m_4^{'}+m_5^{'}}{m}) + Recall} \tag{16}$$

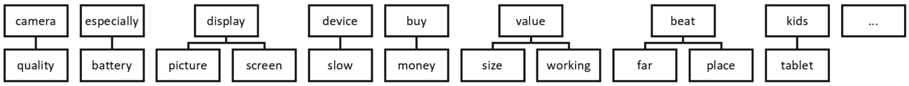
The m is the total number of reviews, among them, the rating mechanism sets user rating with positive integer 1–5. $m_1^{'}$, $m_2^{'}$, $m_3^{'}$, $m_4^{'}$, $m_5^{'}$ are the correct predictions of the number of 1, 2, 3, 4, 5, respectively. m_1, m_2, m_3, m_4, m_5 are the actual number of 1, 2, 3, 4, 5, respectively. $y_i^{'}$ is the predicted rating, y_i is the actual rating.

5.2 Test Analysis

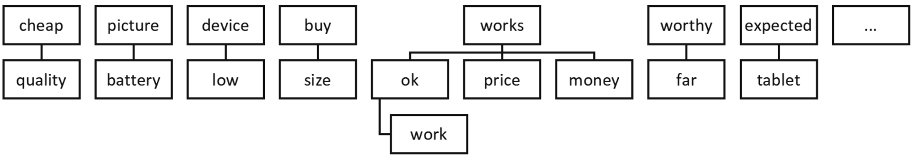
Test on the DHTTW Construction. In this paper, we set the number of topics $K = 5$, the number of topic words $N = 50$, the number of time windows $n = 3$, and the similarity threshold $\alpha = 0.1$. Selecting B00TSUGXKE as an example of the construction of the dynamic hierarchical tree of topic words, specifying the topic T_1. According to the method of constructing the hierarchical tree of topic words proposed in Chapter 4, the dynamic hierarchical analysis of topic words under the topic T_1 is carried out. Part of the hierarchical trees under three time windows are shown in Fig. 2, respectively.



(a) The hierarchical tree of topic word under Time Window T_1



(b) The hierarchical tree of topic word under Time Window T_2



(c) The hierarchical tree of topic word under Time Window T_3

Fig. 2. Example-dynamic hierarchical tree of topic world

As shown in Fig. 2, the topic words extracted from user reviews are divided into two categories, one is the user's daily language, such as "ok", "money", "good", etc. and the other is the user's descriptive vocabulary for goods, such as "camera", "quality", "battery", "screen", etc. It can be clearly seen from the Fig. 2 that the descriptive vocabulary of a product is at or above the second hierarchy of the hierarchical tree. The more such vocabulary users use in their reviews, the more they like the product. Meanwhile, descriptive vocabulary, such as "quality" and "battery", has the same influence on users'ratings in every time window. Vocabulary such as "device" and "screen", has a declining influence on users' ratings. Vocabulary such as "work" and "price", has an increasing influence on users' ratings. Generally speaking, the dynamic hierarchical tree of topic words can reflect the change of the impact of keywords on user ratings.

To prove that the hierarchical tree of topic words proposed in this work changed dynamically, the number of time windows n was set to 3. In each time window, the proportions of topic words in topic T_1 of two categories of products in different hierarchies are displayed in Fig. 3.

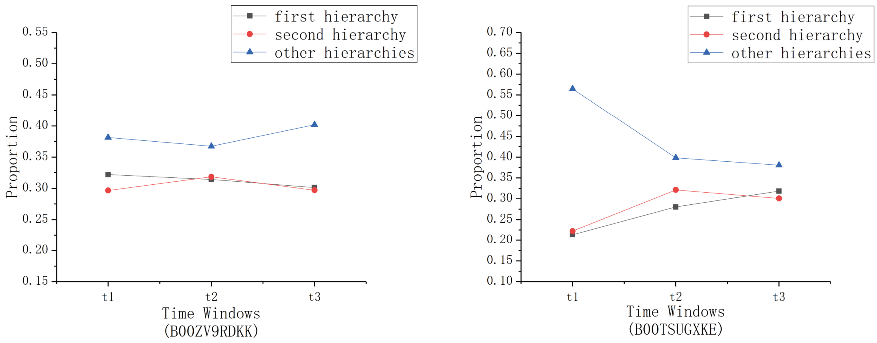


Fig. 3. The hierarchical distribution of topic words in topic T_1

Figure 3 presents that the number of topic words in each hierarchy of hierarchical tree differ in each time window, thereby indicating that the hierarchical tree of topic words changes with time. The change in the hierarchy of topic words described the change in user preference slightly. Thus, user rating prediction based on the dynamic and hierarchical analysis of topic words would adapt to the evolution of user preferences and enhance the timeliness of the rating prediction.

Comparison of Prediction Results Based on DHTTW. At the same time, in order to prove the effectiveness of DHTTW, three prediction models, LR, GBDT and RF, are used to predict rating. The product code chosen in the experiment is B00TSUGXKE. The similarity threshold ∂ of the DHTTW was set to 0.025, the number of time windows n to 3; in addition, the number of topics K was set to 5, and the number of topic words N to 50. The comparison with the method [23] is shown in Fig. 4.

Figure 4 displays that the method for rating prediction based on the dynamic and hierarchical analysis of topic words proposed in this work was superior to the method for analyzing reviews based on the LDA model in four evaluation indexes, namely, MAE, RMSE, Recall and F-score. The optimization degrees of rating prediction results of two categories of products were different because the hierarchical analysis on topic words of two products could describe the influence of different topic words on user rating, thereby enhancing the practicality of the rating prediction work. Thus, DHTTW can reduce the error of rating prediction of two products on the basis of the LDA prediction model. The dynamic analysis of topic words result in timely rating prediction to reflect the changing rule of user preferences well in products. So the effect of DHTTW is obviously better than that of LDA.

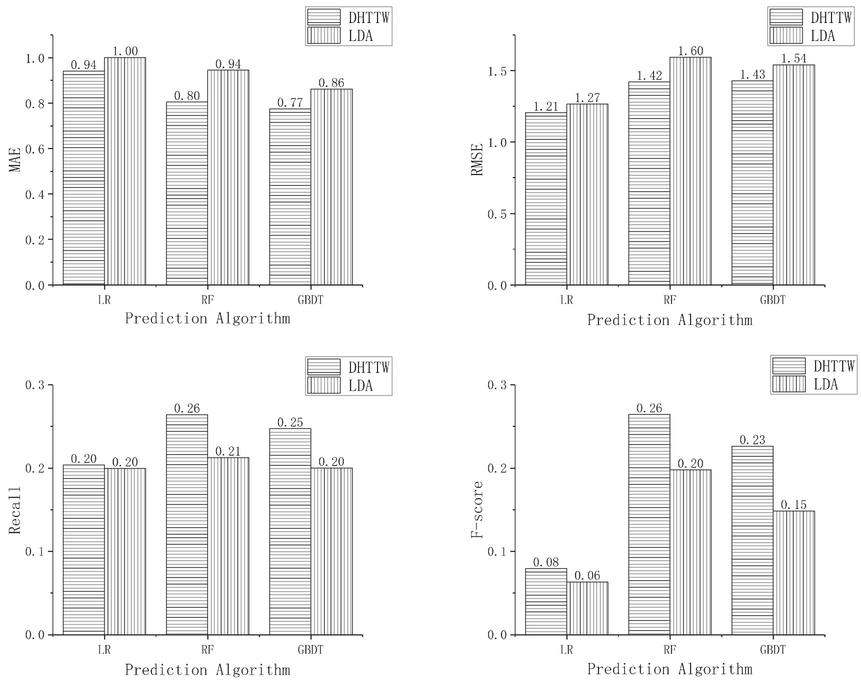


Fig. 4. Comparison of rating prediction results

Test on the GBDT-MCLR. As shown in Fig. 4, the prediction algorithm-LR performs best on RMSE and the prediction algorithm-GBDT performs best on MAE. On the basis of DHTTW, the GBDT-LR algorithm mentioned [27] and the GBDT-MCLR algorithm are tested. The experimental parameters are shown in Fig. 4, the experimental results are shown in Fig. 5.

As shown in Fig. 5, compared with GBDT and LR models alone, the prediction results with using GBDT-LR model are not much improved. The reason is that LR model can not achieve expected results in predicting discrete data. Therefore, the GBDT-MCLR algorithm proposed in this paper can make the GBDT-LR algorithm adapt to discrete rating data to a certain extent, thus making MAE and RMSE worthwhile to be effectively reduced. However, for Recall, GBDT-MCLR is not as effective as GBDT, but the difference is no more than 0.05. For F-score, this indicator has increased 0.1. This is because the GBDT-MCLR algorithm essentially uses the GBDT algorithm to optimize the LR algorithm, so it can be improved compared to the LR algorithm.

Test on the Different Number of Data. Figure 6 shows the variation of the rating prediction error of the same product under different number of user reviews. The product code chosen in the experiment is B00ZV9RDKK. User reviews in each time window are randomly selected for 20%, 40%, 60% and 80% to verify the prediction error of the algorithm under different number of user reviews.

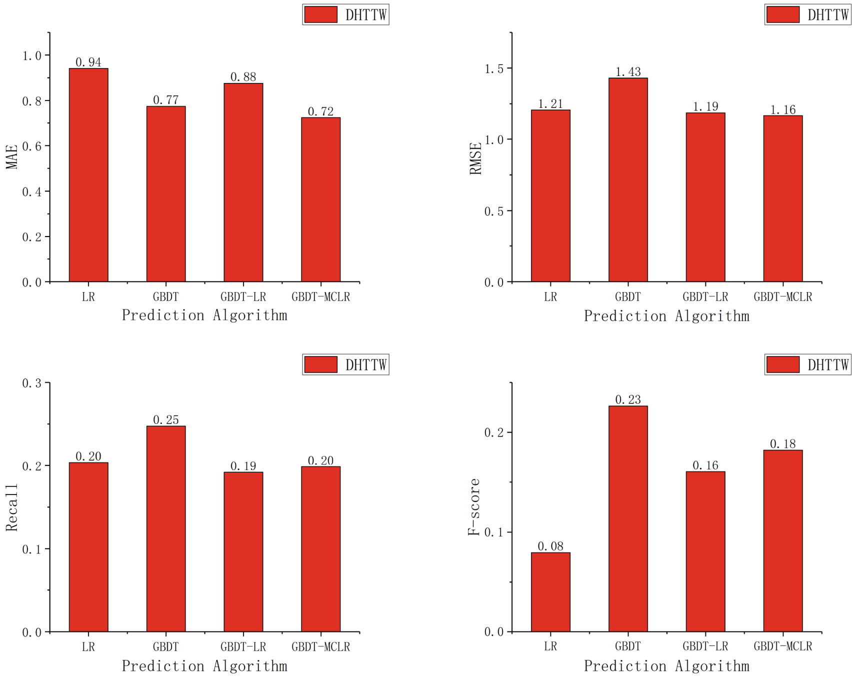


Fig. 5. Comparison of rating prediction results

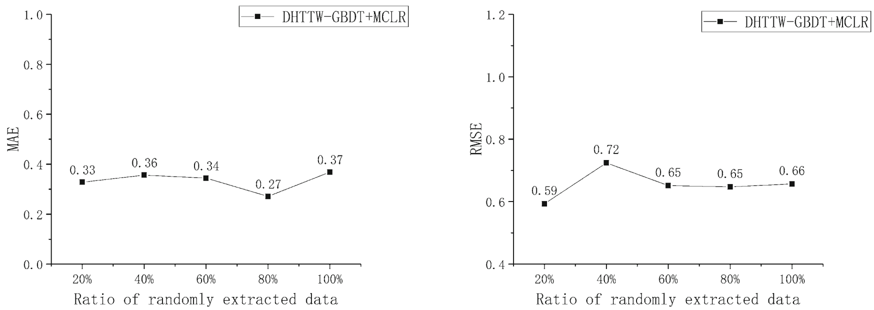


Fig. 6. Ratio of randomly extracted data

As shown in Fig. 6, for the same product, the error of rating prediction tends to be stable under the different number of user reviews, which indicates that the prediction model is stable and suitable for user rating prediction of each product. At the same time, we can see from Figs. 5 and 6 that the predicted errors of user ratings for different products are different, because the data sets come from real e-commerce websites, and the quality of user reviews for different products can not be guaranteed.

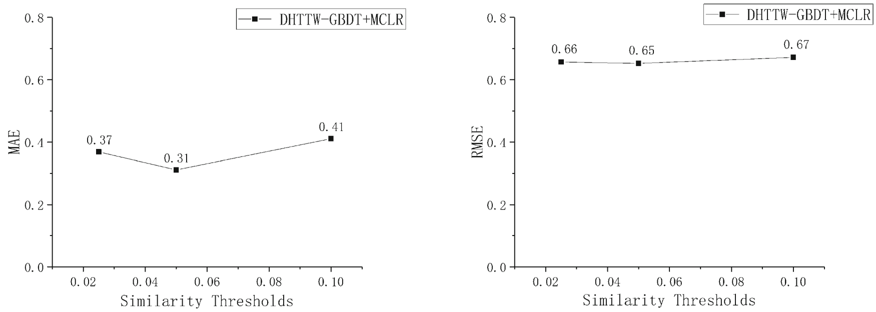


Fig. 7. Rating prediction results under different similarity thresholds

Test on the Similarity Threshold. With the increase of similarity threshold, the proportion of topic words in the first hierarchy of hierarchical tree increased, whereas those in other hierarchies decreased continuously. Thereby indicating that a small similarity threshold denotes additional topic words that were divided into the lower structure of the same word of upper concept, and an apparent hierarchical structure between topic words. When the hierarchical structure between topic words was apparent, the effect on the result of rating prediction is demonstrated in Fig. 7.

Figure 7 exhibits that an apparent hierarchical structure of topic words indicates an improved MAE and RMSE values of rating prediction results, but the improvement was insignificant. This result was due to a small similarity threshold resulted in additional topic words that were divided into the lower structure of the same upper concept. Therefore, a gradual increase in topic words was at high hierarchies. In real life, each word that users use to review a product has different effects on user rating, but the effect will be similar in several hierarchies. Thus, if additional topic words are found at high hierarchies, the method for analyzing reviews with the hierarchical tree of topic words cannot improve the rating prediction result well.

Test on the Number of Time Windows. To verify the effect of different numbers of time windows on rating prediction, the number of time windows was set to 3 (in year), 6 (in half a year), and 12 (in quarter) for the test. The number of topic K was set to 5, the number of topic words N to 50, and the similarity threshold ∂ was set to 0.025. The rating prediction results of two categories of products under different numbers of time windows are illustrated in Fig. 8.

Figure 8 depicts that the dynamic and hierarchical analysis of topic words could be conducted in a small time range with the increase in the number of time windows. The rating prediction result of B00TSUGXKE was clearly improved, but that of B00ZV9RDKK was only slightly improved because it belonged to the slowly updating product. User preference for such products changed slowly over time, thereby causing minimal change in the hierarchical tree of topic words. Therefore, the improvement in rating prediction result was minimal when

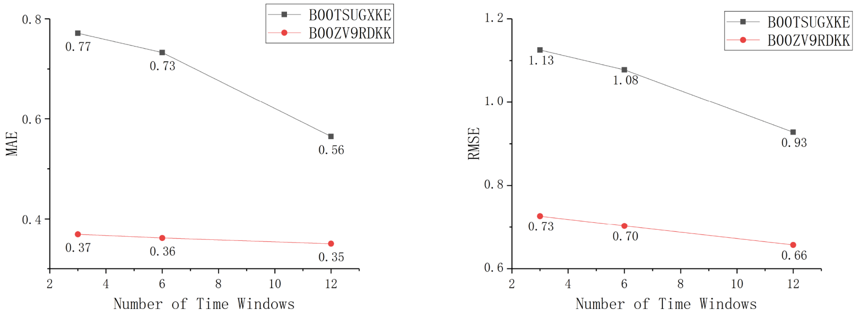


Fig. 8. Rating prediction results under different time window numbers

conducting the dynamic and hierarchical analysis of topic words in a fine time range. User preference for B00TSUGXKE change rapidly, thus resulting in the apparent change in the hierarchical tree of topic words over time. Given a small time interval, adapting to changes in user preferences and the significance rating prediction effect had been improved.

6 Conclusions

As a typical application of social network, the study of social commerce focuses on the dynamic characteristics of time and space. The method for rating prediction based on dynamic and hierarchical analysis of topic words proposed in this work started from the topic discovery of reviews to conduct a dynamic analysis of topic words, which could adapt to the dynamics of user preference for products. The hierarchical trees of topic words was constructed on the basis of dynamics of topic words. Different hierarchies of topic words could describe the influence of different topic words contained in reviews on user rating. The mapping rule from reviews to the hierarchical tree of topic words and the generation method of the user preference vector were designed. The dynamic and hierarchical analysis of the topic words were conducted for realistic and timely rating prediction, thereby reducing the error caused by the unified analysis of topic words.

In the future, we will focuson dynamically selecting the number of time windows. That is, selecting the appropriate number of time windows dynamically to describe the change rule of user preference well for products with different change cycles and achieve improved rating prediction results.

Acknowledgments. This work is supported by National Natural Science Foundation (61662013, U1501252, U1711263, U1811264,61662015, 61562014); Guangxi Innovation-Driven Development Project (Science and Technology Major Project)(AA17202024); The Guangxi Natural Science Foundation (2017GXNSFAA198372, 2016GXNSFAA380149); The Funds of Guangxi Key Lab of Trusted software Project (kx201511); The Teacher Growth Fund of the Education Development Foundation of Guangxi Normal University (EDF2015005); The Funds of Graduate student innovation program Guilin University of Electronic Technology (2017YJCX56, 2019YCXS045).

References

1. Brody, S., Elhadad, N.: An unsupervised aspect-sentiment model for online reviews. In: Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Los Angeles, California, USA, 2–4 June 2010, pp. 804–812 (2010)
2. Titov, I., McDonald, R.T.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, 21–25 April 2008, pp. 111–120 (2008)
3. Tang, X., Xiang, K.: Hotspot mining based on LDA model and microblog heat. Libr. Inf. Serv. **58**(5), 58–63 (2014)
4. Shin, S.-J., Moon, I.-C.: Guided HTM: hierarchical topic model with Dirichlet forest priors. IEEE Trans. Knowl. Data Eng. **29**(2), 330–343 (2017)
5. Bingyu, L., Cuirong, W., Cong, W.: Microblog community discovery algorithm based on dynamic topic model with multidimensional data fusion. J. Softw. **28**(2), 246–261 (2017)
6. Cena, F., Gena, C., Grillo, P., Kuflik, T., Vernero, F., Wecker, A.J.: How scales influence user rating behaviour in recommender systems. Behav. Inf. Technol. **36**(10), 985–1004 (2017)
7. Yin, Y., Chen, L., Xu, Y., Jian, W.: Location-aware service recommendation with enhanced probabilistic matrix factorization. IEEE Access **6**, 62815–62825 (2018)
8. Yin, Y., Xu, Y., Xu, W., Min, G., Yu, L., Pei, Y.: Collaborative service selection via ensemble learning in mixed mobile network environments. Entropy **19**(7), 358 (2017)
9. Gao, H., Zhang, K., Yang, J., et al.: Applying improved particle swarm optimization for dynamic service composition focusing on quality of service evaluations under hybrid networks. Int. J. Distrib. Sens. Netw. **14**(2), 1550147718761583 (2018)
10. Koren, Y., Bell, R.: Advances in collaborative filtering. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 77–118. Springer, Boston (2015). https://doi.org/10.1007/978-1-4899-7637-6_3
11. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, 9–12 February 2011, pp. 815–824 (2011)
12. Titov, I., McDonald, R.T.: A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, ACL 2008, Columbus, Ohio, USA, 15–20 June 2008, pp. 308–316 (2008)
13. Zhang, W., Xu, M., Jiang, Q.: Opinion mining and sentiment analysis in social media: challenges and applications. In: Proceedings of HCI in Business, Government, and Organizations - 5th International Conference Held as Part of HCI International 2018, HCIBGO 2018, Las Vegas, NV, USA, 15–20 July 2018, pp. 536–548 (2018)
14. Goulart, H.X., Tosi, M.D.L., Gonçalves, D.S., Maia, R.F., Wachs-Lopes, G.A.: Hybrid model for word prediction using Naive Bayes and latent information. CoRR, abs/1803.00985 (2018)
15. Keith, T., Debra, Y., Kathleen F.M., Christopher A.P.: Topic modeling in fringe word prediction for AAC. In: Proceedings of the 11th International Conference on Intelligent User Interfaces, IUI 2006, Sydney, Australia, January 29–February 1 2006, pp. 276–278 (2006)

16. Tang, D., Qin, B., Liu, T., Yang, Y.: User modeling with neural network for review rating prediction. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, 25–31 July 2015, pp. 1340–1346 (2015)
17. Seo, S., Huang, J., Yang, H., Liu, Y.: Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, 27–31 August 2017, pp. 297–305 (2017)
18. Ma, C., Chen, W.: A review topic analysis method for rating prediction. J. Chin. Inf. Process. **2**, 209–216 (2017)
19. Ji, Y., Li, Y., Shi, C.: Aspect rating prediction based on heterogeneous network and topic model. J. Comput. Appl. **37**(11), 3201–3206 (2017)
20. Fang, G.-S., Kamei, S., Fujita, S.: Rating prediction with topic gradient descent method for matrix factorization in recommendation. Int. J. Adv. Comput. Sci. Appl. **8**(12), 469–476 (2017)
21. Zhang, W., Wang, J.: Integrating topic and latent factors for scalable personalized review-based rating prediction. IEEE Trans. Knowl. Data Eng. **28**(11), 3013–3027 (2016)
22. McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM Conference on Recommender Systems, pp. 165–172. ACM (2013)
23. Zhang, R., et al.: Review comment analysis for predicting ratings. In: Dong, X.L., Yu, X., Li, J., Sun, Y. (eds.) WAIM 2015. LNCS, vol. 9098, pp. 247–259. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21042-1_20
24. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested Chinese restaurant process. In: Advances in Neural Information Processing Systems 16, NIPS 2003, Vancouver and Whistler, British Columbia, Canada, 8–13 December 2003, pp. 17–24 (2003)
25. Paranjpe, D.: Learning document aboutness from implicit user feedback and document structure. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, pp. 365–374, 2–6 November 2009
26. Gupta, M.S.: Predicting click through rate for job listings. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, pp. 1053–1054, 20–24 April 2009
27. Wang, Y., Feng, D., Li, D., Chen, X., Zhao, Y., Niu, X.: A mobile recommendation system based on logistic regression and gradient boosting decision trees. In: 2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, pp. 1896–1902, 24–29 July 2016