Expert Systems With Applications 135 (2019) 249–258

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Sentiment based matrix factorization with reliability for recommendation

Rong-Ping Shen[a], Heng-Ru Zhang[a], Hong Yu[b], Fan Min[a,c,*]

[a] School of Computer Science, Southwest Petroleum University, Chengdu 610500, China
[b] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[c] Institute for Artificial Intelligence, Southwest Petroleum University, Chengdu 610500, China

ARTICLE INFO

Article history:
Received 19 December 2018
Revised 26 April 2019
Accepted 1 June 2019
Available online 3 June 2019

Keywords:
Collaborative filtering
Matrix factorization
Recommender system
Sentiment analysis

ABSTRACT

Recommender systems aim at predicting users' preferences based on abundant information, such as user ratings, demographics, and reviews. Although reviews are sparser than ratings, they provide more detailed and reliable information about users' true preferences. Currently, reviews are often used to improve the explainability of recommender systems. In this paper, we propose the sentiment based matrix factorization with reliability (SBMF+R) algorithm to leverage reviews for prediction. First, we develop a sentiment analysis approach using a new star-based dictionary construction technique to obtain the sentiment score. Second, we design a user reliability measure that combines user consistency and the feedback on reviews. Third, we incorporate the ratings, reviews, and feedback into a probabilistic matrix factorization framework for prediction. Experiments on eight Amazon datasets demonstrated that SBMF+R is more accurate than state-of-the-art algorithms.

1. Introduction

Recommender systems are successful in predicting users' preferences for items using various techniques, such as content-based (Pazzani & Billsus, 2007) and collaborative filtering (CF)-based (Sarwar, Karypis, Konstan, & Riedl, 2001) approaches. Content-based approaches extract the features of items and build a model or profile of user interests for recommendation. CF-based approaches predict the interests of users by collecting information from similar users or relevant items. One class of popular CF-based methods is the latent factor model, including matrix factorization (MF) (Koren, Bell, & Volinsky, 2009), non-negative MF (Lee & Seung, 2001), and probabilistic MF (PMF) (Mnih & Salakhutdinov, 2008).

CF assumes that the ratings can reflect users' true preferences and items' real qualities. However, this assumption does not always accord with real-world scenarios for several reasons. First, critical users tend to give a poor rating, whereas tolerant users might give a good rating, which does not necessarily reflect the quality of the item (Raghavan, Gunasekar, & Ghosh, 2012). Second, users who give the same rating may have different degrees of satisfaction (Cheng, Ding, Zhu, & Kankanhalli, 2018; Lloret, Saggion, & Palomar, 2010). Third, rating noise is introduced by fake likes, that is, superior or bad ratings on products presented by a large group of low-paid workers (Mobasher, Burke, Bhaumik, & Williams, 2007). Fourth, a wrong click also has an implicit influence on the unreliability of ratings.

Fig. 1 shows some examples of both reviews and ratings selected from Amazon datasets. Here the reviews given by users are inconsistent with their ratings. Fortunately, the increasing information, such as pictures, tags, reviews, and feedback on e-commerce sites, has shed some light on this problem (McAuley, Targett, Shi, & Hengel, 2015; Yu, Zhou, Deng, & Hu, 2018). Most e-commerce sites allow users to provide reviews of products and give feedback (up or down) to evaluate the usefulness of the reviews (Chen, Qi, & Wang, 2012; Connors, Mudambi, & Schuff, 2011; Hart-Davidson, McLeod, Klerkx, & Wojcik, 2010). Numerous models have been proposed to use review information to facilitate the recommendation task (Chen, Chen, & Wang, 2015; Diao et al., 2014; Kun-Peng, Ramanathan, & N, 2010; Wang, Zhu, & Li, 2013). Popular review extraction techniques include latent Dirichlet allocation and feature discovery (Bao, Fang, & Zhang, 2014; 2014; McAuley & Leskovec, 2013; Tang, Mao, & Huang, 2016).

Recently, sentiment analysis (Cambria, Schuller, Xia, & Havasi, 2013; Wilson, Wiebe, & Hoffmann, 2005; Zhang & Liu, 2011) has

* Corresponding author at: Institute for Artificial Intelligence, Southwest Petroleum University, Chengdu 610500, China.

E-mail addresses: shen_rongping@126.com (R.-P. Shen), minfan@swpu.edu.cn (H.-R. Zhang), yuhong@cqupt.edu.cn (H. Yu), minfan@swpu.edu.cn (F. Min).

https://doi.org/10.1016/j.eswa.2019.06.001
0957-4174/

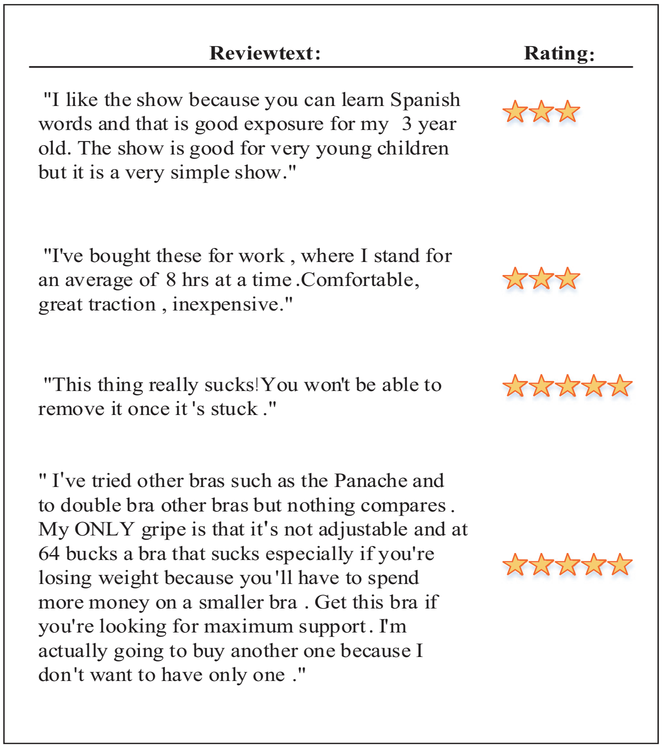

Fig. 1. Examples of inconsistent reviews and ratings on Amazon datasets.

been proved to be effective in uncovering people's emotions and opinions in different areas (He et al., 2017; Hu, Tang, Tang, & Liu, 2013; Li, Xie, Chen, Wang, & Deng, 2014b). It has also led to many sentiment embedding approaches in recommender systems (Chen and Wang, 2014; Dong, P. O!'Māhony, Schaal, McCarthy, and Smyth, 2016; H. Alahmadi & Xiaojun, 2015; Jakob, Weber, Müller, & Gurevych, 2009; Yang, Zhang, Yu, & Wang, 2013a). For example, Zhang et al. (2014) extracted explicit product features and user opinions using phrase-level sentiment analysis on user reviews to generate explainable recommendations. Pappas and Popescu-Belis (2013) labeled each review as either positive or negative, and adjusted the prediction using these labels. Faridani (2011) conducted sentiment analysis on textual reviews to infer ratings and expanded them into multiple dimensions. Ganu, Elhadad, and Marian (2009) considered textual information to obtain review score predictions.

In this paper, we propose a new algorithm called sentiment based MF with reliability (SBMF+R) for this issue. The algorithm leverages review information to adjust the ratings and obtain more reliable and fine-grained scores. By contrast, it uses a reliability measure to weight the ratings and thereby improves the prediction accuracy. The SBMF+R algorithm consists of three stages. In the first stage, we construct a sentiment dictionary and conduct sentiment analysis on reviews. Instead of using an existing sentiment dictionary (Denecke, 2008; Yang et al., 2013a), we build a dictionary that contains rating-related scores for each dataset. Then we use the dictionary-based sentiment analysis method to convert reviews into sentiment scores. In the second stage, we use a reliability measure designed in this work to evaluate each rating. This process includes calculating the Euclidean distance between ratings and sentiment scores of the same user, and counting the feedback on each review. In the third stage, we incorporate ratings and sentiment scores into an objective function, where their proportions are adjusted according to the reliability. To summarize, both ratings and reviews are used for reliable recommendations.

Our work differs from the aforementioned works in several aspects. First, many approaches have been designed to use review information in a recommender system. Some researchers use the aspect (topic) information of reviews to assist the prediction (McAuley & Leskovec, 2013; Musat, Liang, & Faltings, 2013; Zhang et al., 2014). Different from this work, we use sentiment analysis on reviews to obtain a numerical sentiment score. Second, most methods that conduct sentiment analysis on reviews consider how to add the sentiment effect to ratings. Pappas and Popescu-Belis (2013) extracted positive and negative labels from reviews and then added a positive and negative effect to ratings. Faridani (2011) used canonical correlation analysis to map text and numerical ratings. By contrast, our algorithm fuses sentiment information into the MF procedure. Therefore, the user-feature and item-feature vectors fit the ratings and reviews simultaneously. Third, only a few existing works pay attention to the reliability of ratings and reviews. For example, Ganu et al. (2009) assumed that the text of a review was a better indicator of sentiment than the coarse star rating. In our work, we assign different weights to ratings and reviews based on their reliability.

The contribution of our algorithm lies in the following two aspects. On one hand, among those review-incorporated methods, few works focus on the inconsistency between ratings and reviews. Our algorithm compares reviews and ratings and detects the inconsistency between them. On the other hand, our algorithm alleviates this issue by introducing a reliability measure. To the best of our knowledge, our method is the first to attempt to measure the reliability between reviews and ratings. Through the reliability measure, ratings with lower reliability are given lower weights in the rating-review combination process. Therefore, more reliable and fine-grained scores are obtained to improve the performance of recommender systems.

We conducted a set of experiments on eight Amazon datasets to validate the effectiveness of our proposed model. The results demonstrated that SBMF+R significantly outperformed the popular rating-based methods. Moreover, SBMF+R obtained better results on five large datasets compared with state-of-the-art review-incorporated methods.

The remainder of the paper is organized as follows: We review preliminaries, including sentiment analysis and PMF, in Section 2. We describe the fundamental sentiment based MF (SBMF) model and the extended model with a reliability measure (SBMF+R) in Section 3. We describe the experimental methodology of our approach and the results in Section 4. Finally, we draw conclusions in Section 5. The implementation of the SBMF+R algorithm is available at https://github.com/FanSmale, where all source code is accessible.

2. Preliminaries

In this section, we introduce background information for our research, including sentiment analysis and PMF. Table 1 lists the notation used throughout this paper.

2.1. Sentiment analysis

Sentiment analysis is effective in detecting users' opinions and attitudes using materials such as reviews and survey responses. We use a dictionary-based sentiment analysis method to obtain users' sentiment scores (Li, Shi, Huang, Su, & Wang, 2014a; Neviarouskaya, Prendinger, & Ishizuka, 2011; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). Fig. 2 shows the results of sentiment analysis on reviews from Amazon datasets. The highlighted words in different colors correspond to different sentiment values.

Review Text:
"**Interesting** basic concept with a *poor* use of flashbacks."
"**Good** acting. **Interesting** plot."
"**Great** storyline, but *slow* plot progress."
"**Clunky**."
"**Five stars**."
"*Four stars*."
"**Excellent** king series hope it *inspires* more..."
"**Disappointing** and i **don't like** the **weird** cover for the image of the second season, either, the warholish picture of the 2 character."
"**Horror junkie**'s review: Instantly *forgettable*; even Jeff fahey *couldn't help*."
"I watched the first episode because I **like** the lead actress and I **enjoy** science fiction."
"It is an *addictive* show - can not figure out where it is headed at all but I **love** watching it - so **entertaining**."

Fig. 2. Reviews from an Amazon dataset with highlighted sentiment words: bold = +2, italic = +1, underline-italic = −1, and underline-bold = −2.

Table 1
Notation.

Notation	Definition
n	Number of users
m	Number of items
$R_{m\times n}$	User-item rating matrix
$\widehat{R}_{m\times n}$	Predicted rating matrix
$S_{m\times n}$	User-item sentiment score matrix
$W_{m\times n}$	Reliability factor matrix
$U_{l\times m}$	User latent feature matrix
$V_{l\times n}$	Item latent feature matrix
$I_{m\times n}$	Indicator function matrix of user ratings
λ	Regulation parameter, e.g., λ_U, λ_V
σ^2	Variance of Gaussian distribution, e.g., σ_U^2, σ_V^2
C_i	Consistency of user i
F_{ij}	Total number of votes of review that user i leave on item j
F_{ij}^P	Positive votes of review that user i leave on item j
H_{ij}	Helpfulness of review that user i leave on item j
T	Test dataset

According to (Taboada et al., 2011), the total sentiment score of a review is the arithmetic sum of each word in the dictionary.

$$S(F) = \sum_{i=1}^{m} dictionary.value(D_i), \tag{1}$$

where $S(F)$ represents the total sentiment score of sentence F and *dictionary.value* is the sentiment score of word D_i defined in the dictionary.

Based on the sentiment dictionary, the sentence “Great storyline, but slow plot progress,” “great,” and “slow” has the sentiment scores of +2 and −1, respectively. Therefore, using Eq. (1), the total sentiment score of this sentence is $S(F) = +2 + (-1) = +1$. Note that the value of the score is not limited. In reality, it often reaches −10 or +30.

This score is then converted into a real value in the range [1, 5], where 1 and 5 represent the most negative and positive sentiments, respectively. Thus, the sentiment scores are in the same range as the numerical ratings.

2.2. Probabilistic matrix factorization

We adopt the PMF technique (Mnih & Salakhutdinov, 2008) to learn the latent characteristics, and factorize the user-item rating matrix. As shown in Fig. 3, the user factors and item factors are modeled as latent matrices using the Gaussian hypothesis. Let $U \in R^{l\times m}$ and $V \in R^{l\times n}$ be the user and item latent feature matrices,

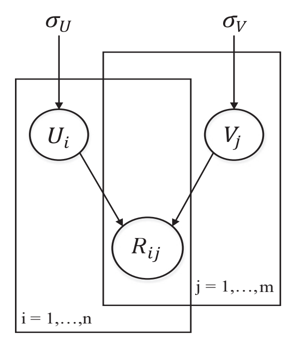

Fig. 3. Graphical model of PMF.

respectively. The conditional distribution of the user rating is given by

$$p(R \mid U, V, \sigma_R^2) = \prod_{i=1}^{m}\prod_{j=1}^{n}[\mathcal{N}(R_{ij} \mid g(U_i^T V_j), \sigma_R^2)]^{I_{ij}^R}, \tag{2}$$

where $(x \mid \mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2, and I_{ij}^R is the indicator function that equals 1 if user i rated item j and equals 0 otherwise. We also model the user and item latent feature vectors based on the Gaussian priors hypothesis:

$$p(U \mid \sigma_U^2) = \prod_{i=1}^{m}[\mathcal{N}(U_i \mid 0, \sigma_U^2 \mathbf{I})], \tag{3}$$

$$p(V \mid \sigma_V^2) = \prod_{j=1}^{n}[\mathcal{N}(V_j \mid 0, \sigma_V^2 \mathbf{I})]. \tag{4}$$

Hence, through simple Bayesian inference, we have

$$\begin{aligned} & p(U, V \mid R, \sigma_R^2, \sigma_U^2, \sigma_V^2) \\ & \propto p(R \mid U, V, \sigma_R^2) p(U \mid \sigma_U^2) p(V \mid \sigma_V^2) \\ & = \prod_{i=1}^{m}\prod_{j=1}^{n}[\mathcal{N}(R_{ij} \mid g(U_i^T V_j), \sigma_R^2)]^{I_{ij}^R} \\ & \quad \times \prod_{i=1}^{m}[\mathcal{N}(U_i \mid 0, \sigma_U^2 \mathbf{I})] \times \prod_{j=1}^{n}[\mathcal{N}(V_j \mid 0, \sigma_V^2 \mathbf{I})]. \end{aligned} \tag{5}$$

Table 2
Words with sentiment scores.

Office products	Instant video	Beauty	Cell phones
great +2	enjoyable +2	awesome +2	perfect +2
handy +1	realistic +1	affordable +1	handy +1
weak −1	predictable −1	greasy −1	limited −1
useless −2	boring −2	disappointed −2	flimsy −2

3. Proposed algorithm

In this section, we describe our SBMF+R algorithm. First, we present a star-based method for dictionary construction. Second, we introduce the review sentiment analysis process. Third, we introduce the sentiment based matrix factorization (SBMF) model. Fourth, we use the reliability measure designed in this section is employed to evaluate each rating. Finally, we extend the SBMF model to the SBMF+R algorithm.

3.1. Dictionary construction

In the first stage, we construct the star-based sentiment dictionary from the review text. Table 2 shows some words with sentiment score selected from the star-based dictionary. Star-based means that we consider rating stars in the classification of sentiment words. Our intuition is that reviews with 5 stars mostly convey positive emotions, and vice versa (Lu, Malu, Umeshwar, & Zhai, 2011). Therefore, the keywords in reviews should be in line with the reviews' overall emotions. We design two techniques for dictionary construction. Most existing sentiment dictionaries only label words with a binary value (positive or negative). We assign different scores that correspond to different sentiment intensities to obtain more accurate results. The same word could express different emotions for different categories of products. Therefore, we build a different dictionary for each category to alleviate this problem.

Instead of using existing sentiment dictionaries, such as SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010), we extract keywords and build a new dictionary. First, we classify the reviews into five groups according to their rating levels (1 star to 5 stars). Second, for each group of reviews, we extract keywords with a term frequency of more than 0.002. In the extraction process, we filter out stop words, neutral nouns, and verbs using the NTLK toolkit. Furthermore, we consider the effects of negation, modality and the localization role of adverbs. Third, the keywords in different groups (1, 2, 3, 4, and 5 stars) are assigned corresponding sentiment scores (−2, −1, 0, 1, and 2). Fourth, we classify words that appear in multiple groups, we classify them into the group with the highest term frequency.

We conduct the construction process on each dataset, therefore each dataset has a dictionary that fits its features. The reliability of the star-based dictionary is reported in Section 4.

3.2. Review sentiment analysis

After constructing the sentiment dictionary, we conduct sentiment analysis on each review. The process is similar to Section 2.1 but we consider the effects of negation and modality.

Negation Negation can be used to deny or reject statements (Jia, Yu, & Meng, 2010). We consider two types of negation reversal to check whether the sentiment score should be adjusted (Benamara, Chardon, Mathieu, Popescu, & Asher, 2012). One type of negation completely reverses the polarity of the sentiment. For example, if we assume that the score of "comfortable" is +1, then the score of "not comfortable" is reversed to −1. Another type of negation is that of a very positive (negative) adjective, which slightly reverses the polarity. For example, if we assume that the score of "excellent" is +2, then the score of "not excellent" is multiplied by −0.5 to 1.

Modality Modality is a grammatical expression that embeds possibility, necessity or ability in a sentence (Yang, Yu, Chen, & Bing, 2013b). It is grammatically expressed via modal verbs such as "maybe", "certainly", and "may". We adopt the approaches in (Özdemir, 2015) and consider the modal verbs that weaken the sentimental intensity. Therefore, a sentiment word that occurs in the scope of modality is multiplied by 0.5 to dampen its intensity.

Table 3 presents the final sentiment score after negation and modality operations. To evaluate the precision of the sentiment analysis process, we conducted the experiments on eight datasets. The quality of sentiment analysis is reported in Section 4.1.

3.3. SBMF Model

Based on the PMF approach and sentiment analysis, we design our SBMF model. Fig. 4 shows the graphical model of SBMF. The user-rating matrix is modeled as given by Eq. (5). Through simple Bayesian inference, the conditional distribution of user rating is computed as

$$\begin{aligned} &p(U, V \mid R, \sigma_R^2, \sigma_U^2, \sigma_V^2) \\ &\propto p(R \mid U, V, \sigma_R^2) p(U \mid \sigma_U^2) p(V \mid \sigma_V^2) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n} [\mathcal{N}(R_{ij} \mid g(U_i^T V_j), \sigma_R^2)]^{I_{ij}^R} \\ &\times \prod_{i=1}^{m} [\mathcal{N}(U_i \mid 0, \sigma_U^2 \mathbf{I})] \times \prod_{j=1}^{n} [\mathcal{N}(V_j \mid 0, \sigma_V^2 \mathbf{I})]. \end{aligned} \tag{6}$$

Now we consider MF for the sentiment score. Let S_{ij} represent the sentiment score that we obtain from the early stage. We also model the user and item latent feature matrices as $U \in R^{l\times m}$ and $V \in R^{l\times n}$, respectively. To model the user and item latent feature matrices to fit the sentiment score, we define the conditional distribution over the observed sentiment score as

$$p(S \mid U, V, \sigma_S^2) = \prod_{i=1}^{m} \prod_{j=1}^{n} [\mathcal{N}(s_{ij} \mid g(U_i^T V_j), \sigma_S^2)]^{I_{ij}^S}, \tag{7}$$

where I_{ij}^R is the indicator function that equals 1 if user i rated movie j and equals 0 otherwise. We also place zero-mean spherical Gaussian priors on the user and item feature vectors:

$$p(U \mid \sigma_U^2) = \prod_{i=1}^{m} [\mathcal{N}(U_i \mid 0, \sigma_U^2 \mathbf{I})], \tag{8}$$

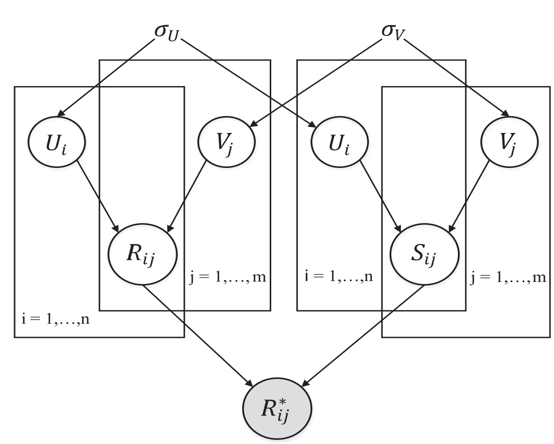

Fig. 4. Graphical model of SBMF.

Table 3
Some sentiment scores after negation and modality operations.

word	category	original score	in context of	trigger word	final score
impressed	positive	+2	negation	not	2*(−0.5) = −1
problem	negative	−1	negation	didn't	(−1)*(−1)=1
well	positive	+1	modality	hope	1*0.5=0.5
leak	negative	−2	modality	may	(−2)*0.5=−1

$$p(V \mid \sigma_V^2) = \prod_{j=1}^{n}[\mathcal{N}(V_j \mid 0, \sigma_V^2\mathbf{I})]. \tag{9}$$

Hence, similar to the inference in Eq. (6), we have

$$\begin{aligned} &p(U, V \mid S, \sigma_S^2, \sigma_U^2, \sigma_V^2) \\ &\propto p(S \mid U, V, \sigma_S^2)p(U \mid \sigma_U^2)p(V \mid \sigma_V^2) \\ &= \prod_{i=1}^{m}\prod_{j=1}^{n}[\mathcal{N}(S_{ij} \mid g(U_i^T V_j), \sigma_S^2)]^{I_{ij}^S} \\ &\times \prod_{i=1}^{m}[\mathcal{N}(U_i \mid 0, \sigma_U^2\mathbf{I})] \times \prod_{j=1}^{n}[\mathcal{N}(V_j \mid 0, \sigma_V^2\mathbf{I})]. \end{aligned} \tag{10}$$

Maximizing the log-posterior over the user and item features with hyperparameters (${\sigma_R}^2$, ${\sigma_U}^2$, ${\sigma_V}^2$) kept fixed is equivalent to minimizing the following sum-of-squared-errors objective function:

$$\begin{aligned} L(\theta) = &\sum_{i,j} I_{ij}[(R_{ij} - U_i^T V_j)^2] + \sum_{i,j} I_{ij}[(S_{ij} - U_i^T V_j)^2)] \\ &+ \lambda_U \|U\|_F^2 + \lambda_V \|V\|_F^2, \end{aligned} \tag{11}$$

where $\sum_{i,j} I_{ij}[(R_{ij} - U_i^T V_j)^2]$ and $\sum_{i,j} I_{ij}[(S_{ij} - U_i^T V_j)^2)]$ are the cost functions of rating and sentiment score, $\lambda_U = {\sigma_R}^2/{\sigma_U}^2$,$\lambda_V = {\sigma_R}^2/{\sigma_V}^2$ are the regularization parameters and $\|\cdot\|$ is the Frobenius norm of a matrix. This equation is used to build the final model.

3.4. Reliability measure

Only a few works have discussed whether ratings or reviews are more reliable in the prediction task (Ganu et al., 2009; Hu, Pavlou, & Zhang, 2006; Tang, Gao, Hu, & Liu, 2013). We take both user consistency and review feedback into account as the reliability measure. Then we calculate the reliability measure of each rating to assign a personalized weight to them.

We denote the rating vector of user i as $R_i = [R_{i1}, \ldots, R_{in}]$. Let $S_i = [S_{i1}, \ldots, S_{in}]$ be the user's sentiment score vector that we calculated using sentiment analysis. The consistency of user i is defined as the Euclidean distance C_i between the user's rating vector and the user's sentiment score vector. The larger the distance between ratings and reviews, the lower the consistency of users. User consistency is defined as

$$C_i = \sqrt{\sum_{i=1}^{n}(R_i - S_i)^2}. \tag{12}$$

In addition to user consistency, we also consider the helpfulness of each review. Most e-commerce websites allow users to provide feedback on reviews as thumbs up or thumbs down. This feedback is considered as the helpfulness of the reviews, which reflects the authenticity of the reviews. Therefore, we use positive feedback as the helpfulness of the reviews. Some researchers have proposed automatically assessing the feedback of each review (Kim, Pantel, Chklovski, & Pennacchiotti, 2006; Martin & Pu, 2014; Raghavan et al., 2012; Tang et al., 2013). Let F_{ij} and F_{ij}^P denote the total number of votes and positive votes, for the review that user U_i left on item V_j. Then the helpfulness of review H_{ij} is given by

$$H_{ij} = F_{ij}^P / F_{ij}. \tag{13}$$

Let W_{ij} represent the reliability of user is review on item j. Then the reliability factor of rating R_{ij} is

$$W_{ij} = \frac{H_{ij}}{1 - C_i}. \tag{14}$$

Similarly, the reliability factor of sentiment score S_{ij} is $1 - W_{ij}$. We normalize the reliability interval to [0, 1]. Using reliability estimation, we finally obtain a personalized weight for each rating.

3.5. SBMF+R

Fig. 5 shows the graphical model of SBMF+R.

SBMF+R is an extension of the SBMF model with a reliability measure. After reliability estimation, as shown in Fig. 5, we extend the SBMF model to the SBMF+R algorithm. We assign different weights to ratings and reviews based on their reliability. With reliability factor w_{ij} calculated, the sum-of-squared-errors objective function is given by

$$\begin{aligned} L(\theta) = &\sum_{i,j} I_{ij}[W_{ij}(R_{ij} - U_i^T V_j)^2] \\ &+ \sum_{i,j} I_{ij}[(1 - W_{ij})(S_{ij} - U_i^T V_j)^2)] + \lambda_U \|U\|_F^2 + \lambda_V \|V\|_F^2, \end{aligned} \tag{15}$$

where $\lambda_U = \sigma^2/{\sigma_U}^2$ and $\lambda_V = \sigma^2/{\sigma_V}^2$ are the regularization parameters and $\|\cdot\|$ is the Frobenius norm of a matrix. In our experiments, $\lambda_U = \lambda_V = \lambda$.

The log posterior probability of Eq. (15) is

$$\begin{aligned} &\ln p(U, V \mid S, R, \sigma_s^2, \sigma_R^2, \sigma_U^2, \sigma_V^2) \\ &= -\frac{1}{2{\sigma_S}^2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}[W_{ij}(R_{ij} - g(U_i^T V_j))^2] \\ &-\frac{1}{2{\sigma_R}^2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}[(1 - W_{ij})(S_{ij} - g(U_i^T V_j))^2] \\ &-\frac{1}{2{\sigma_U}^2}\sum_{i=1}^{m} U_i^T U_i - \frac{1}{2{\sigma_V}^2}\sum_{j=1}^{n} V_j^T V_j \\ &-\frac{1}{2}((\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij})\ln\sigma_S^2 + (\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij})\ln\sigma_R^2) \\ &-\frac{1}{2}(ml\ln\sigma_U^2 + nl\ln\sigma_U^2) + C, \end{aligned} \tag{16}$$

where C is an independent constant. Maximizing Eq. (16) with hyperparameters $(\sigma^2, \sigma_U^2, \sigma_V^2)$ kept fixed is equivalent to minimizing the following objective function:

$$\mathcal{L}(U, V, R, S) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}^R[W_{ij}(R_{ij} - g(U_i^T V_j))^2]$$

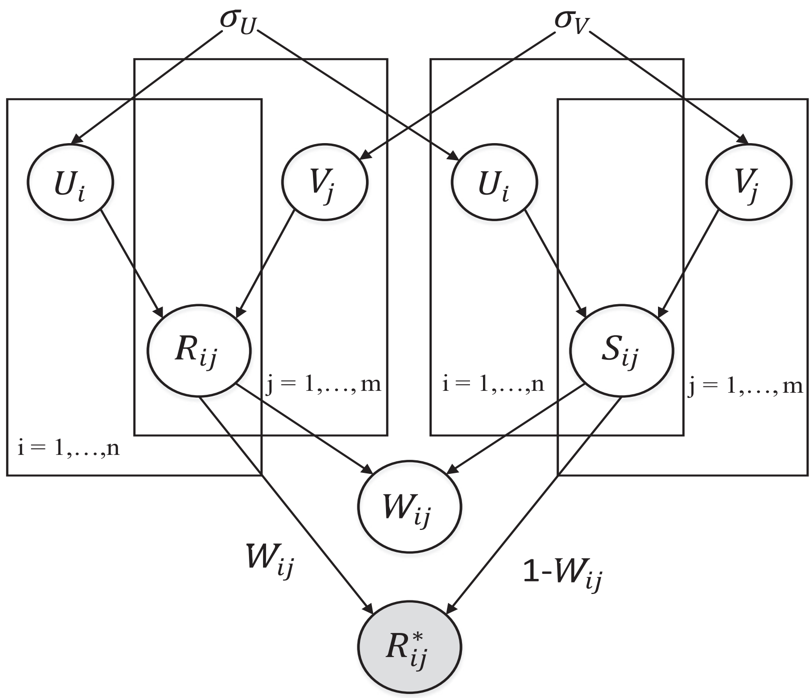

Fig. 5. Graphical model of SBMF+R.

Table 4
Statistics for the eight Amazon datasets.

Dataset	#users	#items	#review	#words	#words/item	#reviews/item
Patio, lawn and garden	1,686	963	13,272	1,360,320	102.49	13.80
Office products	4,905	2,420	53,258	5,103,630	95.83	22.01
Amazon instant video	5,130	1,685	37,126	2,415,801	65.07	22.03
Baby	19,445	7,050	160,792	5,449,909	33.89	22.81
Tools and home improvement	19,856	10,217	134,476	4,777,536	35.53	13.16
Beauty	22,365	12,101	198,502	12,571,484	63.33	16.40
Cell phones and accessories	27,879	10,429	194,439	12,207,401	62.78	18.64
Clothing and accessories	39,387	23,033	278,677	12,874,343	46.19	12.10

$$+\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}^{R}[(1-W_{ij})(S_{ij}-g(U_i^T V_j))^2] + \frac{\lambda_U}{2}\|U\|_F^2 + \frac{\lambda_V}{2}\|V\|_F^2. \quad (17)$$

The local minimum of objective function Eq. (18) and Eq. (19)can be obtained by performing gradient descent on U_i and V_j:

$$\frac{\mathcal{L}}{U_i} = \sum_{j=1}^{n} I_{ij}^{R}[W_{ij} g'(U_i^T V_j)(g(U_i^T V_j) - R_{ij})V_j] + \sum_{j=1}^{n} I_{ij}^{R}[(1-W_{ij}) g'(U_i^T V_j)(g(U_i^T V_j) - S_{ij})V_j] + \lambda_U U_i \quad (18)$$

$$\frac{\mathcal{L}}{V_j} = \sum_{i=1}^{m} I_{ij}^{R}[W_{ij} g'(U_i^T V_j)(g(U_i^T V_j) - R_{ij})U_i] + \sum_{i=1}^{m} I_{ij}^{R}[(1-W_{ij}) g'(U_i^T V_j)(g(U_i^T V_j) - S_{ij})U_i] + \lambda_V V_j. \quad (19)$$

4. Experiments

We conducted experiments on eight 5-core Amazon product review datasets. Table 4 shows the statistics for the datasets used in our experiments. The 5-core Amazon datasets included reviews, ratings, helpfulness votes, product metadata and links, etc. 5-core data means that each of the users and items in the datasets had at least five reviews each. Three types of information were used as input: rating, review, and helpfulness votes. Each dataset contained products of one category. The number of reviews ranged from 13,272 to 278,677.

Table 5 presents the distribution of the five levels of ratings. We observed that the 5 stars accounted for nearly 50% of the total ratings.

4.1. Sentiment analysis evaluation

Table 6 shows the evaluation of the star-based dictionary, where 'Prec' denotes 'Precison'. To evaluate the reliability of the dictionary, we used a similar human annotation scheme to that used in (Zhang et al., 2014). We randomly sampled 20% of words in the dictionary for each dataset to be labeled by three human annotators. Then the sentiment scores of these words were converted into a binary value (positive or negative). A sentiment word was considered appropriate if it was approved by at least two annotators. The average agreement among human annotators is 93.85%.

To evaluate the performance of sentiment analysis, we used a similar human annotation scheme to that used in (Lu et al., 2011). We randomly sampled some reviews from each dataset to be labeled by three human annotators. The annotators were asked to

Table 5
Rating distribution.

Dataset	1 star	2 stars	3 stars	4 stars	5 stars
	The proportion of level-k (%)				
Patio, lawn and garden	3.91	5.07	12.50	25.50	53.02
Office products	2.12	3.24	9.50	28.19	56.95
Amazon instant video	4.62	5.08	11.28	22.75	56.27
Baby	4.86	5.72	10.73	20.52	58.17
Tools and home improvement	3.82	3.69	8.01	21.07	63.41
Beauty	5.30	5.77	11.21	20.02	57.70
Cell phones and accessories	6.83	5.68	11.03	20.57	55.89
Clothing and accessories	4.01	5.55	10.92	20.94	58.58

Table 6
Evaluation results for the sentiment lexicon.

	Patio	Office	Video	Baby	Tools	Beauty	Phones	Clothing
words	76	75	85	72	71	79	75	69
Prec	0.9537	0.9533	0.9329	0.9522	0.9337	0.9547	0.9433	0.9410

Table 7
Evaluation results of sentiment analysis performance.

	Patio	Office	Video	Baby	Tools	Beauty	Phones	Clothing
entry	82	91	120	123	131	156	182	219
Prec(Postive)	0.8032.	0.8116	0.8625	0.8494	0.8922	0.8586	0.8783	0.8901
Prec(Neutral)	0.7059	0.7619	0.8261	0.8235	0.8125	0.8095	0.8077	0.8148
Prec(Negative)	0.6862	0.7442	0.6957	0.7429	0.7073	0.6341	0.7377	0.8194
Prec(Overall)	0.7683	0.7912	0.8167	0.8130	0.8473	0.8269	0.8516	0.8584

label each review with sentiment information (five-scale). The sentiment score interval was transformed into rating levels [1,5] and a review was considered proper if it was approved by at least two annotators. Table 7 shows the evaluation results, where positive, neutral and negative denote the scores within the rating levels 4–5, 3, and 1–2, respectively. The average agreement among human annotators was 82.17%.

4.2. Evaluation metric

We used normalized Root-Mean-Square-Error (NRMSE) in our experiment to measure the performance of prediction accuracy:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(R_{i,j} - \widehat{R}_{ij})^2}{|T|}}, \tag{20}$$

$$NRMSE = \frac{RMSE}{R_{max} - R_{min}}, \tag{21}$$

where T denotes the test data; $R_{i,j}$ and $\widehat{R}_{ij}$ denote the observed rating and the predicted rating of user i on item j, respectively; and $R_{\max}$ and $R_{\min}$ represent the maximum and minimum rating, respectively. The reason this metric was used is that we could compare the performance of difference datasets regardless of the range or variance it has. The smaller NRMSE, the better the performance.

4.3. Model evaluation

We compared the performance of SBMF+R with different methods for evaluation. The primitive MF algorithm based on only users ratings was considered as the baseline. To evaluate the effectiveness of review incorporation, we compared the null SBMF model with rating-based methods. The methods compared in the experiment are as follows:

1) Matrix factorization (MF) (Koren et al., 2009): the primitive matrix factorization algorithm, which was used as the baseline.
2) Probabilistic matrix factorization (PMF) (Mnih & Salakhutdinov, 2008): a classical matrix factorization method with the Gaussian hypothesis.
3) Hidden factors as topics (HFT) (McAuley & Leskovec, 2013): this algorithm considers reviews into recommendation tasks and fuses latent review topics into latent rating dimensions. We used the code provided by the authors.
4) Sentiment Probabilistic matrix factorization (SBMF): we simply incorporated sentiment information into the model without considering the reliability.

Both MF and PMF only take ratings as input, whereas HFT, SBMF, and SMBF+R take both ratings and reviews as input. Comparisons between MF, PMF, and SMBF demonstrate the effectiveness of review incorporation. Comparisons between HFT, SBMF, and SMBF+R demonstrate the advantages of sentiment analysis in processing reviews. Additionally, they also demonstrate the effectiveness of the reliability measure in improving the prediction accuracy. We randomly split the datasets into 80% training data and 20% test data.

Fig. 6 shows the performance of SBMF+R with latent feature dimension K varied from 5 to 25. In the experiment, latent space dimensions of the five methods were all set to 10 for fairness. We also fixed $\lambda_U = \lambda_V = \lambda = 0.01$ and learning rate = 0.2 for all methods.

4.4. Results

Table 8 shows the NRMSE for the models in the experiments. In particular, we show the improvements of our algorithm compared with the rating-based method PMF and review-incorporated method HFT. On comparison, we have the following observations:

1) SBMF+R obtained the lowest NRMSE on the Baby, Tools, Beauty, CellPhones, and Clothing datasets, with improvements of 0.32%, 0.75%, 2.72%, 6.88%, and 2.35% respectively, compared with the second best method.
2) SBMF+R obtained the second-lowest NRMSE on the Office, Patio, and Instant video datasets. Though SBMF+R ranked sec-

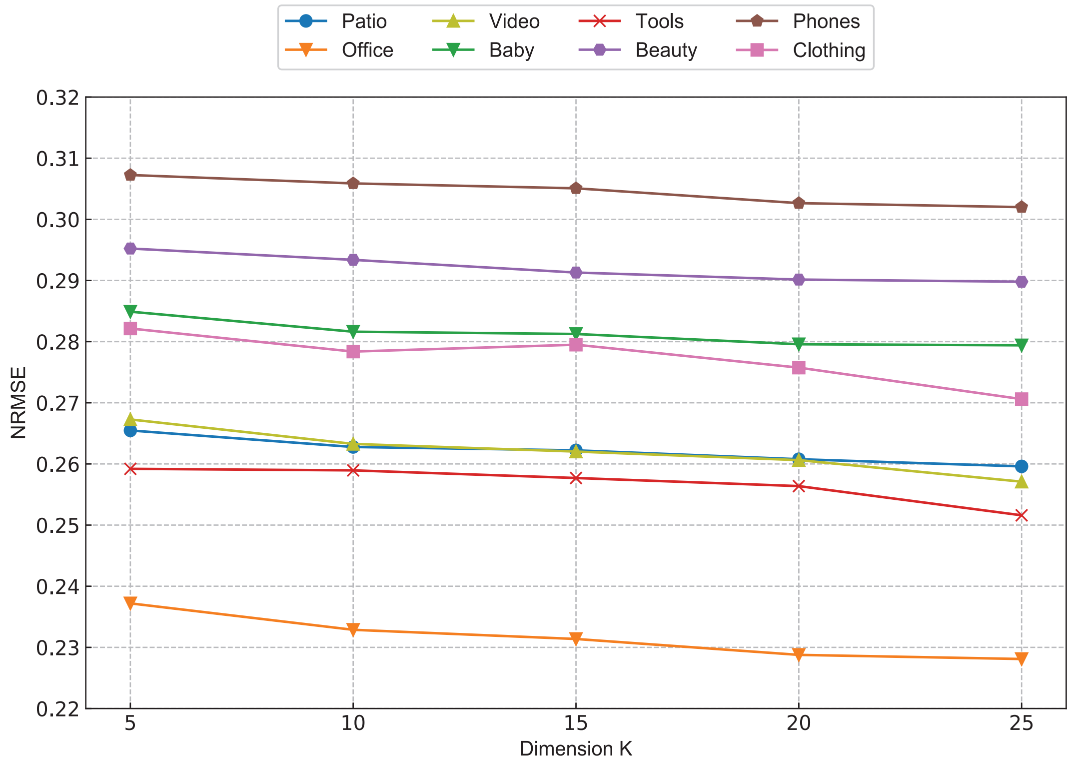

Fig. 6. Effect of dimension *K*.

Table 8
NRMSE comparison for the different methods.

Dataset	MF	PMF	HFT	SBMF	SBMF+R	Improvement	
	(a)	(b)	(c)	(d)	(e)	e vs. b	e vs. c
Patio	0.2992	0.2934	**0.2455**	0.2653	0.2596	11.52%	−5.74%
Office products	0.2565	0.2379	**0.2206**	0.2329	0.2281	4.12%	−3.40%
Instant Video	0.2993	0.2917	**0.2447**	0.2798	0.2571	11.86%	−5.07%
Baby	0.3088	0.2913	0.2812	0.2803	**0.2794**	4.09%	0.64%
Tools	0.2783	0.2649	0.2544	0.2535	**0.2516**	5.02%	1.10%
Beauty	0.3411	0.3338	0.3085	0.2979	**0.2898**	13.18%	6.06%
Cell Phones	0.3496	0.3243	0.3426	0.3054	**0.3020**	6.88%	11.85%
Clothing	0.2824	0.2810	0.2864	0.2771	**0.2706**	3.70%	5.52%

ond, it still demonstrated great improvements compared with rating-based MF methods, which demonstrates the effectiveness of review incorporation. The improvements on the Office, Patio, and Instant video datasets compared with the best results for MF, PMF were approximately 4.12%, 11.52%, and 11.86%, respectively.

3) To evaluate the validity of reliability, we conduct pairwise T-test a significance level of 5% on SPSS to analyze the significant difference. The pairwise T-test rejected the null hypothesis with a p-value equal to 0.02759. Therefore, we can claim that there is a significant difference between SBMF and SBMF+R.
4) From Table 8, we conclude that SBMF+R performed better on five datasets. These five datasets all contained more than 19,000 users, and other three datasets contained less than 6000 users.

5. Conclusions and future work

In this paper, we presented the SBMF+R algorithm to leverage reviews for prediction. SBMF+R adjusts the weights of rating and sentiment information using the reliability measure. Compared with both rating-based and review-based state-of-the-art algorithms, our algorithm obtained better results on some datasets. One area of future work is to use more advanced sentiment analysis techniques to identify the correlation between reviews and ratings. Another area of future work is to improve recommendation accuracy and diversity of recommendation by incorporating more types of information, such as brand and price.

Declaration of interests

- The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Credit authorship contribution statement

Rong-Ping Shen: Writing - original draft, Data curation, Methodology, Software, Visualization, Validation. **Heng-Ru Zhang:** Conceptualization, Writing - original draft, Formal analysis. **Hong**

Yu: Writing - review & editing, Supervision. **Fan Min:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Acknowledgements

The authors thank the anonymous reviewers for their valuable comments and suggestions. This work was supported in part by the National Natural Science Foundation of China (41604114); the Natural Science Foundation of Sichuan Province (2019YJ0314); and the Scientific Innovation Group for Youths of Sichuan Province (2019JDTD0017).

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on international language resources and evaluation* (pp. 2200–2204).

Bao, Y., Fang, H., & Zhang, J. (2014). TopicMF: Simultaneously exploiting ratings and reviews for recommendation. In *Proceedings of the 28th AAAI conference on artificial intelligence* (pp. 2–8).

Benamara, F., Chardon, B., Mathieu, Y., Popescu, V., & Asher, N. (2012). How do negation and modality impact opinions? In *Proceedings of the workshop on extra-propositional aspects of meaning in computational linguistics* (pp. 10–18).

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems, 28*, 15–21. doi:10.1109/xMIS.2013.30.

Chen, L., Chen, G., & Wang, F. (2015). Recommender systems based on user reviews: The state of the art. *User Modeling and User-Adapted Interaction, 25*, 99–154. doi:10.1007/s11257-015-9155-5.

Chen, L., Qi, L., & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert Systems with Applications, 39*, 9588–9601. doi:10.1016/j.eswa.2012.02.158.

Chen, L., & Wang, F. (2014). Sentiment-enhanced explanation of product recommendations. In *Proceedings of the 23rd international conference on world wide web* (pp. 239–240). doi:10.1145/2567948.2577276.

Cheng, Z., Ding, Y., Zhu, L., & Kankanhalli, M. (2018). Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the 2018 world wide web conference* (pp. 639–648). doi:10.1145/3178876.3186145.

Connors, L., Mudambi, S. M., & Schuff, D. (2011). Is it the review or the reviewer? a multi-method approach to determine the antecedents of online review helpfulness. In *Proceedings of the annual hawaii international conference on system sciences* (pp. 1–10). doi:10.1109/HICSS.2011.260.

Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th international conference on data engineering workshop* (pp. 507–512). doi:10.1109/ICDEW.2008.4498370.

Diao, Q., Qiu, M., Wu, C. Y., Smola, A. J., Jiang, J., & Wang, C. (2014). Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 193–202). doi:10.1145/2623330.2623758.

Dong, R., P. O. Mahony, M., Schaal, M., McCarthy, K., & Smyth, B. (2016). Combining similarity and sentiment in opinion mining for product recommendation. *Journal of Intelligent Information Systems, 46*, 285–312. doi:10.1007/s10844-015-0379-y.

Faridani, S. (2011). Using canonical correlation analysis for generalized sentiment analysis, product recommendation and search. In *Proceedings of the 5th ACM conference on recommender systems* (pp. 355–358). doi:10.1145/2043932.2044005.

Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the 12th international workshop on the web and databases: vol. 9* (pp. 1–6).

H. Alahmadi, D., & Xiaojun, Z. (2015). ISTS: Implicit social trust and sentiment based approach to recommender systems. *Expert Systems with Applications, 42*, 8840–8849. doi:10.1016/j.eswa.2015.07.036.

Hart-Davidson, W., McLeod, M., Klerkx, C., & Wojcik, M. (2010). A method for measuring helpfulness in online peer review. In *Proceedings of the 28th ACM international conference on design of communication* (pp. 115–121). doi:10.1145/1878450.1878470.

He, R., Liu, Y., Yu, G., Tang, J., Hu, Q., & Dang, J. (2017). Twitter summarization with social-temporal context. *World Wide Web, 20*, 267–290. doi:10.1007/s11280-016-0386-0.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on electronic commerce* (pp. 324–330). doi:10.1145/1134707.1134743.

Hu, X., Tang, L., Tang, J., & Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. In *ACM international conference on web search and data mining* (pp. 537–546). doi:10.1145/2433396.2433465.

Jakob, N., Weber, S. H., Müller, M. C., & Gurevych, I. (2009). Beyond the stars: Exploiting free-text user reviews to improve the accuracy of movie recommendations. In *Proceedings of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion* (pp. 57–64). doi:10.1145/1651461.1651473.

Jia, L., Yu, C. T., & Meng, W. (2010). The effect of negation on sentiment analysis and retrieval effectiveness. In *Acm conference on information & knowledge management* (pp. 1827–1830). doi:10.1145/1645953.1646241.

Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 423–430). doi:10.3115/1610075.1610135.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*, 30–37. doi:10.1109/MC.2009.263.

Kun-Peng, Z., Ramanathan, N., & N, C. A. (2010). Voice of the customers: Mining online customer reviews for product feature-based ranking. *Proceedings of the 3rd Wonference on Online Social Networks, 10, 11–11*. doi:10.1109/JSTQE.2006.884785.

Lee, D. D., & Seung, H. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556–562). doi:10.1109/TGRS.2013.2249522.

Li, R., Shi, S., Huang, H., Su, C., & Wang, T. (2014a). A method of polarity computation of chinese sentiment words based on gaussian distribution. In *Proceedings of the 15th international conference on computational linguistics and intelligent text processing* (pp. 53–61).

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014b). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*, 14–23. doi:10.1016/j.knosys.2014.04.022.

Lloret, E., Saggion, H., & Palomar, M. (2010). Experiments on summary-based opinion classification. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 107–115).

Lu, Y., Malu, C., Umeshwar, D., & Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: An optimization approach. In *Proceedings of the 20th international conference on world wide web* (pp. 347–356). doi:10.1145/1963405.1963456.

Martin, L., & Pu, P. (2014). Prediction of helpful reviews using emotions extraction. In *Proceedings of the 28th AAAI conference on artificial intelligence* (pp. 1551–1557).

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on recommender systems* (pp. 165–172). doi:10.1145/2507157.2507163.

McAuley, J., Targett, C., Shi, Q., & Hengel, A. V. D. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 43–52). doi:10.1145/2766462.2767755.

Mnih, A., & Salakhutdinov, R. (2008). Probabilistic matrix factorization. In *Advances in neural information processing systems* (pp. 1257–1264).

Mobasher, B., Burke, R., Bhaumik, R., & Williams, C. (2007). Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology (TOIT), 7*, 23. doi:10.1145/1278366.1278372.

Musat, C. C., Liang, Y., & Faltings, B. (2013). Recommendation using textual opinions. In *Proceedings of the 23rd international joint conference on artificial intelligence* (pp. 2684–2690). doi:10.1145/1651461.1651473.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). Sentiful: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing, 2*, 22–36. doi:10.1109/T-AFFC.2011.1.

Özdemir, C. B. (2015). *Enhancing Sentiment Lexica with Negation and Modality for Sentiment Analysis of Tweets* Ph.D. thesis.. Ph.D. thesis Concordia University.

Pappas, N., & Popescu-Belis, A. (2013). Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval* (pp. 773–776). doi:10.1145/2484028.2484116.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325–341). doi:10.1007/978-3-540-72079-9_10.

Raghavan, S., Gunasekar, S., & Ghosh, J. (2012). Review quality aware collaborative filtering. In *Proceedings of the 6th ACM conference on recommender systems* (pp. 123–130). doi:10.1145/2365952.2365978.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on world wide web* (pp. 285–295). doi:10.1145/371920.372071.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*, 267–307. doi:10.1162/COLI_a_00049.

Tang, J., Gao, H., Hu, X., & Liu, H. (2013). Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM conference on recommender system* (pp. 1–8). doi:10.1145/2507157.2507183.

Tang, Y. K., Mao, X. L., & Huang, H. (2016). Labeled phrase latent dirichlet allocation. In *Web information systems engineering – WISE 2016* (pp. 525–536). doi:10.1007/978-3-319-48740-3_39.

Wang, D., Zhu, S., & Li, T. (2013). Sumview: A web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications, 40*, 27–33. doi:10.1016/j.eswa.2012.05.070.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language* (pp. 347–354). doi:10.3115/1220575.1220619.

Yang, D., Zhang, D., Yu, Z., & Wang, Z. (2013a). A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM conference on hypertext and social media* (pp. 119–128). doi:10.1145/2481492.2481505.

Yang, L., Yu, X., Chen, Z., & Bing, L. (2013b). Sentiment analysis of sentences with modalities. In *Proceedings of the 2013 international workshop on mining unstructured big data using natural language processing* (pp. 39–44). doi:10.1145/2513549.2513556.

Yu, H., Zhou, B., Deng, M., & Hu, F. (2018). Tag recommendation method in folksonomy based on user tagging status. *Journal of Intelligent Information Systems, 50*, 479C–500. doi:10.1007/s10844-017-0468-1.

Zhang, L., & Liu, B. (2011). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, 30*, 167. doi:10.1007/978-1-4899-7502-7_907-1.

Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., & Ma, S. (2014). Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval* (pp. 83–92). doi:10.1145/2600428.2609579.