Augmenting Chinese Online Video Recommendations by Using Virtual Ratings Predicted by Review Sentiment Classification

Weishi Zhang[1], Guiguang Ding[1], Li Chen[2], Chunping Li[1]

[1]*School of Software, Tsinghua University, China*

[2]*Department of Computer Science, Hong Kong Baptist University, Hong Kong*

zhang-ws08@mails.tsinghua.edu.cn; {dinggg, cli}@tsinghua.edu.cn; lichen@comp.hkbu.edu.hk

***Abstract*—In this paper we aim to resolve the recommendation problem by using the virtual ratings in online environments when user rating information is not available. As a matter of fact, in most of current websites especially the Chinese video-sharing ones, the traditional pure rating based collaborative filtering recommender methods are not fully qualified due to the sparsity of rating data. Motivated by our prior work on the investigation of user reviews that broadly appear in such sites, we hence propose a new recommender algorithm by fusing a self-supervised emoticon-integrated sentiment classification approach, by which the missing User-Item Rating Matrix can be substituted by the virtual ratings which are predicted by decomposing user reviews as given to the items. To test the algorithm's practical value, we have first identified the self-supervised sentiment classification's higher performance by comparing it with a supervised approach. Moreover, we conducted a statistic evaluation method to show the effectiveness of our recommender system on improving Chinese online video recommendations' accuracy.**

***Keywords*-Information retrieval; sentiment analysis; opinion mining; online video recommendation.**

I. INTRODUCTION

Recommender systems that suggest unknown interesting items to users have been developed rapidly in recent years, among which collaborative filtering (CF) is one of typical approaches that principally derives recommendations for a user based on the preferences of other users who have similar tastes [5]. In most CF systems, the item ratings in the *User-Item Rating Matrix* are assumed to be obtainable from real-users. However, in reality, many websites especially the existing video-sharing ones, do not provide rating supports, or few users have actually inputted rates (e.g., the sparsity problem [16]). Considering the three biggest Chinese video-sharing websites such as Youku [26], Ku6 [28] and Tudou [29], none of them provide rating interface supports for the users. It hence unfortunately limits the applicability of CF algorithms and even other pure rating based systems to generate accurate recommendations.

Therefore, in this paper, we have mainly investigated the role of user reviews in complementing the rating sparsity problem. User reviews, as another form of user inputs to indicate their interests, have actually broadly appeared in the resource-sharing websites especially video-sharing ones such as YouTube [27] and YouKu [26]. Thus, it will be meaningful to study their impacts on enhancing the accuracy of video recommendations. More specifically, we have aimed at generating the missing rating data from user reviews through the method of sentiment classification, so that given a piece of text, the latent opinion can be discovered to show different possible sentiment polarities (e.g., positive, neutral, or negative) and hence reflect users' preferences on the corresponding item. To make a difference from the traditional user-input ratings, the ratings generated from user reviews are called as virtual ratings.

With the aim, we have first in depth studied the Chinese online reviews in the resource-sharing sites (like the video-sharing ones), and found that 1). the review is usually very short and contents include much noise information such as advertisements, hyperlink text etc.; 2) besides textual comments, there are 41% reviews (the statistic of our experiment data) containing various expression faces, i.e., emoticon (e.g., smiley); 3) the ratio of positive and negative reviews is not 1:1, that is different from the common assumption in related sentiment classification methods [4, 23, 25]. In fact, most of related works on sentiment classification are supervised (i.e., a large number of labeled training data is needed) and developed based on standard review datasets [1, 4, 7, 13], so it is not straightforward to adopt them into the rating generation process based on online reviews.

Thus, we have improved the SELC model [17] on self-supervised sentiment classification in Chinese reviews to address the realistic characteristics of the Chinese online reviews. It concretely uses two models, i.e., unsupervised model and supervised model, to determine the overall sentiment polarity of a review document by analyzing both the sentiment words and emoticons. The method does not need the accumulation of training data, and can automatically build the virtual *User-Item Rating Matrix* to be fused into standard CF algorithms.

The main contributions of our work can be summarized as follows: 1) we propose a self-supervised sentiment classification approach that particularly considers the special features of real Chinese online video reviews, and identify its higher accuracy being compared with both the SELC model and a supervised classification approach; 2) we propose a strategy to predict the virtual *User-Item Rating Matrix* by employing the sentiment classification results; (3) we perform a statistical simulation recommendation experiment which significantly show the effectiveness of our approach in providing video recommendations based on real Chinese online data.

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.27

II. RELATED WORK

A. Sentiment Classification

Standard machine learning techniques such as Support Vector Machine (SVM) and Naive Bayes have been usually used in supervised sentiment classification methods [1]. Different factors affecting the machine learning process are investigated. For example, linguistic, statistical and n-gram features were researched in [7]. Selected words and negation phrases were investigated in [13]. However, the performance of supervised approaches normally decreases when training data is insufficient [2, 18].

On the contrary, unsupervised approaches make the assumption that there are certain words people tend to use to express strong sentiment. In [23], an unsupervised sentiment classification approach was proposed by calculating the mutual information between each phrase in a document and the selected two seed words, excellent and poor. Fewer seed words imply less domain-dependency. The authors in [25] only assign one word *good* as a seed positive word, and use negation words such as *not* to find initial negative expressions.

Although it has been stated in [14] that unsupervised approaches would perform worse than supervised approaches, given that the latter can be built on large training sets, the process of building training sets is unnecessarily time-consuming so as to make large-scale applications impracticable to certain degree [17].

SELC Model (**SE**lf-Supervised, **L**exicon-based and **C**orpus-based Model)[17] is proposed for self-supervised sentiment classification on Chinese IT product reviews. The model includes two phases. In the first phase, some reviews are initially classified based on a sentiment dictionary. Then more reviews are classified through an iterative process with a negative/positive ratio control. In the second phase, a supervised classifier is learned by taking some reviews classified in the first phase as training data. Then the supervised classifier applies on other unclassified reviews to revise the results produced in the first phase. In this paper, we improve this work by considering the special features of real Chinese online video reviews and use the sentimental polarities of reviews as resources for recommendations.

B. Recommender Systems

Since 1990s, recommender systems have been explored in many product domains, i.e., movies [6], TVs [20], web pages [3] with the objective of recommending items matched to users' profiles [24]. In recent years, much more techniques have been developed in recommender systems in order to derive better performance [8, 10, 22]. However, most of works are limited when user preference data (i.e., ratings) are hardly obtainable from real sites. To address this limitation, tags (in form of user-defined keywords) have been utilized as supplementary source to predict user interests [22]. In [22], the authors proposed a generic method that allows tags to be incorporated into standard CF algorithms, by reducing the three-dimensional correlation to three two-dimensional correlations and then applying a fusion method to re-associate these correlations. The authors in [10] have developed a strategy to infer user interests by applying machine learning techniques to learn from both the "official" item descriptions provided by a publisher, and tags that users used to annotate relevant items.

However, to the best of our knowledge, only a few papers have considered user reviews and integrated their sentiment analysis results into the generation of recommendations. In [12] the authors have attempted to identify features from reviews to infer ratings, but no detailed description of how the method was implemented. The sentiment analysis approaches in [9] are supervised and hence need manually annotated training data. The author in [11] proposed three approaches to extract movie aspects as opinion targets and use them as features for the collaborative filtering on IMDB data set. However, none of the prior papers have explored the combination of virtual ratings and review sentiment analysis on Chinese data set. Our work exerts to address this limitation by proposing a self-supervised sentiment classification approach and applying the results to predict the virtual ratings on items, so as to be effectively fused into standard CF algorithm in Chinese data set collected from Chinese video-sharing websites.

As for related works on online video recommendations, [24] proposed a technique to calculate multimodal relevance between videos and users' click-through data. In [15], the proposed video recommender can construct a per-user profile as an aggregation of tag clouds of videos viewed by the user, and then suggest videos based on the viewing patterns of similar users who were identified according to a similarity function over the user profiles. However, few have recognized the potential usefulness of user reviews to further enhance their video recommender systems' performance.

III. OVERVIEW OF OUR APPROACH

Our recommender algorithm is proposed to study the roles of online reviews when being fused into standard recommender algorithm based on the self-supervised sentiment classification method, with the goal of compensating the limitation of rating sparsity problem and augmenting the applicability of recommenders in realistic online environments like the video recommending. The algorithm consists of two phases, i.e., Self-supervised Review Sentiment Classification and Item Recommendation. Figure 1 shows the proposed online item recommender algorithm. Phase 1 and Phase 2 are separated by a dash line.

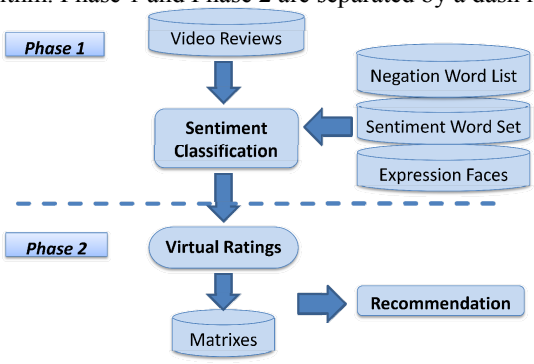

Figure 1. Proposed Online Video Recommender Algorithm.

IV. PHASE 1SELF-SUPERVISED REVIEW SENTIMENT CLASSIFICATION

Based on a *sentiment word set*, a *negation word list* and an *emoticon set*, Phase 1 uses a self-supervised approach to identify the sentiment polarity of reviews. Figure 2 shows the flow chart of the whole self-supervised sentiment classification process of Phase 1. It concretely consists of two models, i.e., unsupervised model and supervised model. In the unsupervised model, an unsupervised approach applies on the original data to automatically label some data. In the supervised model, a supervised approach applies on the labeled data to acquire a training model. Finally, the model applies on the original data to do classification. In Figure 2 the solid lines refer to the unsupervised model while dash ones refer to the supervised model.

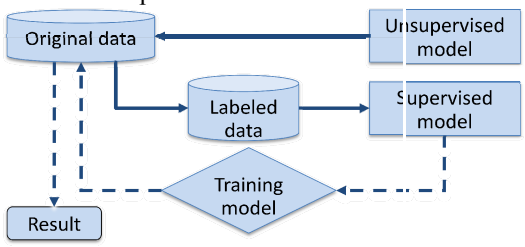

Figure 2. Flow Chart Of The Self-Supervised Sentiment Classification.

In this paper, we made four improvements over the method of SECL model [17]. First, the emoticon scoring analysis is integrated into the supervised model and the emoticon items' sentiment scores can also be updated through the iteration process. Second, a new iteration control strategy is introduced in the unsupervised model with the goal of providing accurately labeled training data for the supervised model. Then for simplicity, in the supervised model we apply the training model on the whole original data to do the classification instead of the integration process (phase 2 of SECL model) of corpus-based model's results and Lexicon-Based model's results. Finally, in supervised model the emoticon items are also taken as features for SVM classifier as well as the sentiment word items.

A. Unsupervised Model

The unsupervised model adopts the basic method of the first phase (i.e., lexicon-based iterative process) of SECL model. In the first phase of SECL, a sentiment vocabulary is initialized by a general sentiment dictionary. The vocabulary is used to label reviews. Then more sentiment words are found from the labeled reviews and update the vocabulary. The new vocabulary then helps classify more reviews. By this iterative process, the vocabulary and labeled reviews are updated (and generally enlarged) step by step. In the iterative process, a technology, i.e., positive/negative ratio control is introduced. That control ranks the reviews and keeps the same number of top-ranked positive and negative reviews in each iteration.

Additionally, in this paper the emoticon scoring analysis is integrated into the iterative process of the unsupervised model to get more accurate results. The unsupervised model consists of the steps as follows.

1) Step 1 Initializing Sentiment Element Sets: To augment the SECL mode, in this paper the *sentiment element sets* consist of two sets, which are *sentiment word set* and *emoticon set*.

The *sentiment word set*, denoted by W_{sen}, includes a list of word items, each of which is assigned with a sentiment score. W_{sen} is initialized by a general sentiment dictionary, which usually includes a lot of positive and negative words. A positive word is initially assigned with score +1.0, while a negative word is assigned with score -1.0. Monosyllabic words are filtered from W_{sen}, because most of them are too ambiguous to provide reliable sentiment. In addition, since the general sentiment dictionary is applicable to many domains, this method has the potential to be domain independent.

The *emoticon set*, denoted by E, is the set of the emoticons (e.g., smiley or sad faces) used by the users to express their preferences. Because the emoticons are widely used by the users in many resource-sharing websites to express their opinions, they play an important role in the task of our item review sentiment classification. First, we manually remove all of the none-sentiment-bearing emoticons, e.g., [Oh…] and [Well...], from the whole set of crawled expression emoticons. Then we add the remaining part into E and according to the sentiment they express, the emoticons are divided into two kinds: positive and negative. Each positive face in E is initially assigned with score +1.0, and a negative emoticon is assigned with score -1.0. We selected 10 positive emoticons and 5 negative emoticons as the initial *emoticon set* (see details in the experiment Section VI).

For the generation of the *negation word list*, we manually selected ten most frequently used negation words, such as "不" ('not'), "不会" ('would not'), "没有" ('don't have'), "没" ('don't have'), etc. (see the dataset used in the experiment Section VI).

2) Step 2 Identifying Review Sentiment Scores: Through analyzing online reviews, two kinds of reviews have been found in most of view-sharing websites: users expressed their opinions on the items, and/or expressed their opinions on other users' reviews. We call the first kind as *item-oriented reviews* and the second kind as *user-oriented reviews.* It's easy to differentiate between these two kinds of reviews, since the *user-oriented reviews* always start with a "[reply to] + [other user]" writing styles. Since the sentiment of *user-oriented reviews* is usually not very related to the items directly, we only apply the sentiment classification algorithm on *item-oriented reviews*. There are also some noise reviews including advertisements, hyperlink text, which are not related and removed for the consideration.

Therefore, at first, a pre-processing was conducted to filter out all the *user-oriented* and noise reviews according to their writing styles. Given an item i, all related reviews of it are denoted by $Rev(i)$. Each review r ($r \in Rev(i)$) is then divided into clauses by punctuation marks.

Secondly, for each clause, if it contains sentiment word items as appearing in W_{sen} (the sentiment word set), each

sentiment word item w of the clause is scored by Equation (1), where L_w is the length of the word item, L_{clause} is the length of the clause, S_w^W is the word item's current sentiment score in W_{sen}, and N_w is a negation check coefficient with a default value of 1.0. If the word item is preceded by a negation within the specified zone, N_w is set to -1.0.

$$S_w = \frac{L_w^2}{L_{clause}} S_w^W N_w \quad (1)$$

Then the sentiment score of a clause c, denoted by $CS(c)$, is calculated by $CS(c)=\sum S_w$ for all $w \in c$. For each review r, the *ReviewWordScore* (the sentiment score of a review taking into account of its contained sentiment words), denoted by $RS^W(r)$, is subsequently calculated according to Equation (2).

$$RS^W(r) = \sum_{c \in r} CS(c) \quad (2)$$

For review r, the *ReviewEmoticonScore* (the sentiment score of a review taking into account of its contained emoticon items), denoted by $RS^E(r)$, is also calculated according to Equation (3), where S_e^E is the current sentiment score of emoticon item e as appearing in E.

$$RS^E(r) = \sum_{e \in F_{sen} \cap e \in r} S_e^E \quad (3)$$

Finally, the sentiment score of the review r, denoted by $RS(r)$ can be computed using:

$$RS(r) = \alpha RS^W(r) + (1-\alpha) RS^E(r) \quad (4)$$

where parameter $\alpha \in [0,1]$ determines the weight put on each factor, i.e., the balance between the review's word sentiment score $RS^W(r)$ and its emoticon sentiment score $RS^E(r)$.

3) Step 3: Review Sentiment Classification with Ratio Control: Basically, after this step, a review r is labeled as positive (if $RS(r)>0$) or negative ($RS(r)<0$)). This policy looks good but would cause sentiment bias for items. Since there are usually different amount of classified positive and negative reviews, when *sentiment element sets* are updated in step 4, their items' scores (S_w^W and S_e^E) may be biased. For example, if there are 20 positive reviews and 10 negative reviews classified, then the number of words only occurring in the positive reviews is more likely to be bigger than the number of words only occurring in negative ones. If the word "screen" only occurs in one of the positive reviews, then "screen" will assigned with a sentiment score of 1.0, and therefore be judged as a positive word item. But in fact, such a word may not have any sentiment polarity. Such bias is caused by unequal number of positive and negative documents. To overcome the bias, a ratio control is designed, which requires the number of positive and negative reviews in the classified sentiment review list to be the same.

Denote the number of positive and negative reviews in one round of iteration as $RN_{positive}$ and $RN_{negative}$ respectively. To realize the ratio control, first, rank all reviews according to their sentiment score $RS(r)$. Second, take the smaller one of $RN_{positive}$ and $RN_{negative}$, i.e., $Min(RN_{positive}, RN_{negative})$, as a threshold, and remain the positive and negative documents above the threshold in the sentiment review list, and remove others. Figure 3 shows the whole process to classify the reviews with ration control. Those reviews form the sentiment review list.

1. Let $RN_{min}=Min(RN_{positive}, RN_{negative})$.
2. Rank all reviews in descending order by their RS.
3. Document labeling:
   - 3.1 Label the top RN_{min} reviews in the list as positive.
   - 3.2 Label the tail RN_{min} reviews in the list as negative.
   - 3.3 Others are left unlabeled.

Figure 3. Review Sentiment Classification with Ratio Control.

4) Step 4 Updating the Sentiment Element Sets: In this section the sentiment word set W_{sen} and *emoticon set E* are to be updated (and usually enlarged).

For sentiment word set W_{sen}, each lexical item[1] that occurs at least twice in those classified reviews is taken as a candidate word item. For an candidate word item w, denote the number of positive reviews containing w as N_w^p, and the number of negative reviews containing w as N_w^n (preceding by a negation make the account reduce by one). The idea of updating *sentiment word set* W_{sen} is: if N_w^p is much bigger than N_w^n, then w is very likely to be a positive word item, and vice versa. The following formula is designed as a measure.

$$difference(w) = \frac{\left|N_w^p - N_w^n\right|}{\left(N_w^p + N_w^n\right)} \quad (5)$$

If $difference(w) \geq 1$, w is included in W_{sen} (current items in W_{sen} will be removed if they no longer satisfy this condition). The sentiment score of w in W_{sen} is updated as

$$S_w^W = N_w^p - N_w^n \quad (6)$$

In this paper, for updating the *emoticon set*, we used a method analogous to the one updating the *sentiment word set.* For an emoticon item e, denote the number of positive reviews containing e as N_e^p, and the number of negative reviews containing e as N_e^n. The following formula is designed as a measure for updating *emoticon set E*.

$$difference(e) = \frac{\left|N_e^p - N_e^n\right|}{\left(N_e^p + N_e^n\right)} \quad (7)$$

If $difference(e) \geq 1$, w is included in E (current items in E will be removed if they no longer satisfy this condition). The sentiment score of w in E is updated as

$$S_e^F = N_e^p - N_e^n \quad (8)$$

5) Step 5 Iteration Control: The unsupervised approach iterates between step 1 and 4. In the SECL model, the iteration completes when both W_{sen} and the sentiment review classification results (the sentiment review list) do not change. Generally, when the iteration completes, most of the reviews are classified (more than 80%). For this paper, since the goal of the unsupervised model is to provide accurately

[1] Let N be the length of a zone, a lexical item is a sequence of Chinese characters excluding punctuation marks, from unigram to N-gram, in an enclosing zone.

labeled data, it is not necessary to label that many reviews. In addition, generally, more reviews are classified; lower accuracy of the classification is acquired, for the errors generated in the former rounds of iteration will propagate to the following ones. Therefore, the iteration should complete at some early point of iteration. However, the iteration cannot complete too early, because the supervised approach still needs adequate data to train the model.

To make the control, a parameter β is set, where 0<β<1. When β*100 percent of documents have been labeled, the iteration completes. In the experiments, β is set as 0.618 (i.e., golden mean). That is, if 61.8% of documents have been labeled, the iteration procedure completes. And the labeled 61.8% of documents are provided as the training data of the supervised model of phase 1.

B. Supervised Model

In supervised model, the Support Vector Machine (SVM) classifier with a linear kernel is selected as the realization of supervised approach. As a widely used method, Support Vector Machine achieves good performance in many areas. But in fact, the performance of supervised model depends much more on the quality of labeled data provided by the unsupervised model, while less on the particular machine learning method.

In this paper, the items of *sentiment element sets* updated by the last iteration are used as the feature set. TFIDF measure (see Equation (9)) is used to compute weights for the items in both *sentiment word set* and *emoticon set*.

$$w_i = tf_i \times \log \frac{N_i}{df_i} \tag{9}$$

Finally, the model applies on the original data to do classification and get the finally review sentiment classification results.

V. Phase 2 Item Recommendation

Because writing reviews is a direct and effective way to show users' preferences on the items in the resource-sharing websites especially the video-sharing ones, we can use the sentimental polarities of reviews as resources for recommendations.

In this phase, we explain how we use the item review sentiment classification results of Phase 1 build the virtual *Rating Matrixes* for the input of the user-based CF recommender algorithm to derive personalized recommendations.

A. Collaborative Filtering Algorithms

First, we introduce the following notations that we use throughout the rest of the paper.

- U : a set of users.
- I : the set of recommendation items.
- R_{UI}: the *User-Item Rating Matrix* where each value $R_{UI}(u,i)$ corresponds to the predicted rating of user u on item i, where $u \in U$ and $i \in I$.
- $V_{UI}(u)$: the item rating vector of user u in R_{UI} along a set.

R_{UI} theoretically can contain any categorical values. In this paper we consider only ternary ratings (+1: like, -1: dislike and 0: unknown).

Most CF recommender algorithms derive recommendations to a user by using opinions from people who have similar tastes, called neighborhood. Recommendations are generated by considering the ratings of users on items, by computing the pair-wise similarities between the current user and his/her neighbors. One typical method is using the vector cosine similarity. The correlation between user u and v is:

$$S_{UI}(u,v) = \frac{V_{UI}(u) \cdot V_{UI}(v)}{\|V_{UI}(u)\| V_{UI}(v)\|} \tag{10}$$

In the above equation, $u, v \in U$, and $V_{UI}(u)$and $V_{UI}(v)$ are their rating vectors in R_{UI}. In this paper, our task is to predict the *top N* interesting items that are unknown to the current user. In user-based CF, to derive the recommendations for a target user u, k most-similar users are selected, which constitute the neighborhood of u, denote by $N(u)$. When predicting the rating of a given user u for an unknown item i, the rating score of i can be computed by:

$$r_{UI}(u,i) = \overline{R_{UI}(u)} + t \sum_{v \in N(u)} w(u,v)(R_{UI}(v,i) - \overline{R_{UI}(v)}) \tag{11}$$

In the above equation, $\overline{R_{UI}(u)}$ is the mean rating for the user u and the weight $w(u,v)$ reflects the similarity between each user v and the given user u (i.e., the value of $S_{UI}(u,v)$). t is a normalized factor. Then, the *top N* items with the highest $r_{UI}(u,i)$ are selected in the recommendation list for the user u.

B. Predicting the Virtual Rating matrix

After the process of Phase 1, each review r is classified as *positive* or *negative*. In this step, we use the review sentiment classification results to predict the virtual rating matrix. Before applying formula (11) to compute recommendations, we first need to predict the virtual *User-Item Rating Matrix* (R_{UI}). In R_{UI}, each user has a virtual *User-Item Vector*, i.e., $V_{UI}(u)$. Each V_{UI} consists of three parts i.e., *Like*+, *Dislike* and *Unknown*- parts. The *Like*+ part of the V_{UI} consists of the items liked by the user u (positive and neutral ones), while the *Dislike*- and *Unknown* parts consist of the items disliked or unknown to user u (negative and unknown ones) respectively.

First, given an item i and a user u, the set of all the reviews that user u puts on item i is denoted as $Rev(u,i)$. The set of all the positive reviews in $Rev(u,i)$ is denoted as $Rev(u,i)^{pos}$, while the set of all the negative reviews in $Rev(u,i)$ is denoted as $Rev(u,i)^{neg}$. Then, for a user u, we calculate the sets of $Rev(u,i)^{pos}$ and $Rev(u,i)^{neg}$ for all the items. Then, we build the *User-Item Vector* ($V_{UI}(u)$) of u in the *Rating Matrix* R_{UI} according to the following rules.

- If the value of ($|Rev_{num}(u,i)^{pos}|$ - $|Rev_{num}(u,i)^{neg}|$) is greater or equal[2] than 0, then we add item i into the *Like*+ part of the $V_{UI}(u)$ with the value of +1.

[2] According to the habits of the majority online users, if they are interested in the item, they are likely to give reviews for it even if the polarities of reviews are not clear sometimes.

- If the value of ($|Rev_{num}(u,i)^{pos}|$ - $|Rev_{num}(u,i)^{neg}|$) is less than 0, then we add item *i* into the *Dislike-* parts of the V_{UI} *(u)* with the value of -1.
- If item *i* is unknown to user *u*, then we add item *i* into the *Unknown* parts of the V_{UI} *(u)* with the value of 0.

VI. EXPERIMENTS

A. Data and Tools

The experiments were conducted on the data sets crawled from a popular video-sharing site in China, called YouKu [26], which is a YouTube counterpart in China. We used the video search engine in Youku to crawl the video reviews. The following nine Chinese queries were used.

{体育 ti-yu '*sport*', 音乐 yin-yu '*music*', 新闻 xin-wen '*news*', 科技 ke-ji '*science*', 旅游 lv-you '*tourism*', 电影 dian-ying '*movie*' , 原创 yuan-chuang '*originality*' , 汽车 qi-che '*automobile*' , 时尚 shi-shang '*fashion*' }.

Finally, we got the data[3] including more than 10,320 videos, each of which had more than 20 reviews. All the reviews were written in Chinese.

In Phase 1, a negation word list that contains ten Chinese negations was used:

{不 bu '*not*', 不会 bu-hui '*would not*', 没有 mei-you '*don't have*', 没 mei '*don't have*', 虽然 sui-ran '*although*', 虽 sui '*although*', 尽管 jin-guan '*although*'', 缺 que '*don't have*', 缺乏 que-fa '*don't have*', 无 wu '*don't have*'}.

For all the experiments, the HowNet Sentiment Dictionary[4] was used as the sentiment dictionary, which is well-known in the area of Chinese sentiment classification containing 4,566 positive words and 4,370 negative words.

There are more than 30 emotions provided by Youku website for users to use while writing reviews. The following 10 positive emoticons and 5 negative emotions were used as the initial *emoticon set*.

{'*smile*', '*love*', '*joking*', '*sweat*', '*naughty*', '*Uh-oh*', '*cool*', '*flower*', '*kiss*', '*thumbs up*'}.

{'*sad*', '*sick up*', '*angry*', '*thumbs up'*, '*Tired'*}.

B. Results of Sentiment Classification

We first tested the accuracy of our sentiment classification method by using a set of data with 1,085 videos and 6,450 users. Each video has at least 100 *video-oriented reviews*, and the total number of reviews is 120,174 in this set. Among the 120,174 reviews, there are 49,271 reviews which contain more than one emoticon. It's a pretty high proportion of 41%. In the experiment we set the value of parameter α in Equation (4) as default 0.3. That is because we considered the emoticon sentiment score $RS^E(r)$ was more important than the word sentiment score $RS^W(r)$ for the sentiment classification.

[3] http://learn.tsinghua.edu.cn:8080/2006990066/OVRdataset.html

[4] http://www.keenage.com/download/sentiment.rar

After removing the noise reviews as mentioned in Section IV, we manually labeled the polarities of 1000 reviews. The numbers of positive, negative reviews in the labeled set are 653 and 347 respectively. We took the labeled results as the actual polarities of the reviews.

To make a comparison, we conducted three approaches, i.e., *Phase1*, *SECL* and *SVM*. In *Phase1*, the method of Phase 1(unsupervised and supervised models) proposed in this paper was used to get the sentiment classification result. In *SECL*, the *SECL* model in [17] was used to get the result. In *SVM*, the supervised Support Vector Machine classifier was used to conduct the sentiment classification with the HowNet Sentiment Dictionary and the initial *emoticon set* as the feature set using the well-known *tfidf* values, and the *SVM* classifier with a linear kernel was ran in 10-fold stratified cross-validation mode. As the results, Table 1 shows the sentiment classification's precision and recall values from the three approaches.

TABLE I. RESULTS OF THE SENTIMENT CLASSIFICATION.

Method	Review Sentiment	Precision	Recall	F_1
Phase1	***Positive***	92.8	**96.4**	94.6
	Negative	**92.9**	85.9	89.3
	Total	92.8	92.8	**92.8**
SECL	***Positive***	87.6	94.2	90.8
	Negative	87.2	74.9	80.6
	Total	87.5	87.5	87.5
SVM	***Positive***	86.5	95.7	90.9
	Negative	89.9	71.8	79.8
	Total	87.4	87.4	87.4

From Table 1, we can see that both the two self-supervised approaches (*Phase1* and *SECL*) achieve higher F_1 scores (92.8%, 87.5%) on *Total* reviews than the supervised *SVM* classifier (87.4%). It is worthwhile to note that both the *SECL* and *SVM* have suffered from the unbalance training data (pos:653, neg:347: the common ratio in real online environments) and get bad recall values on *Negative* reviews (74.9% and 71.8%). On the other side, *Phase1* can still achieve a comparatively good recall on *Negative* reviews (85.9%).

We can also see that the results of *Phase 1* outperform those of *SECL* on all the 3 kind of reviews (with F_1 values in *Total* reviews: 92.8% against 87.5%), which indicate that the proposed self-supervised approach of *Phase1* enables higher accuracy of sentiment classification when being compared to the *SECL* model and supervised *SVM* classifier, that is likely because that our method particularly considers the realistic features of the online item reviews.

As mentioned at the beginning of Section IV, in *Phase 1* there are several novel improvements over the method of SECL model, which affect the performance simultaneously. To check their individual effect, two variant models were implemented. They are referred to as V1 and V2 respectively. In V1, the new iteration control strategy is replaced by the iteration control method of the SECL model. In V2, the emoticon analysis is removed from both unsupervised model and supervised of *Phase 1*. Figure 4 shows that both the new iteration control strategy and the emoticon analysis have

taken effect on the performance improvement, i.e., improving 3.1% and 4.5% F_1-scores respectively.

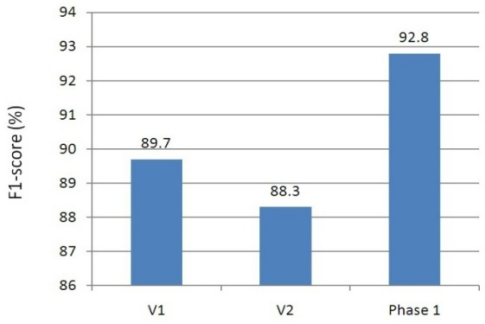

Figure 4. The Results of Two Variants of *Phase1*.

That suggests that the integration of emoticons can be very useful in further increasing the performance of the review sentiment classification, and the new iteration control strategy in the unsupervised model can also provide more accurate training data for the supervised model.

Thus, the above analysis results indicate that our classification approach is capable of overcoming the challenges of online video reviews' special features and providing reliable results for the building of virtual *Rating Matrixes* in the next phrase of producing video recommendations.

C. Results of Recommendations

1) The Recommendation Approaches: To compute recommendations, we classified 68,561 positive, 39,576 negative reviews on 1085 videos. The corresponding rating matrix was established for 6,450 users, with generated 61137 virtual user-item ratings (the number of +1 and -1).

In our experiments, we compared the results for different approaches. Following is the description of labels we used to denote each of these algorithms:

- *CF-Phase1:* The User-based Collaborative Filtering Approach where the results of *Phase1* are used to predict the virtual ratings.
- *CF-SECL:* The User-based Collaborative Filtering Approach where the results of *SECL* are used to predict the virtual ratings.
- *YOUKU:* The recommendation approach of Youku website where each video is along with 3 recommended videos based on video popularity.

2) Statistical Experiment Simulation: In the process of statistical experimental simulation, users are often split into training and test sets. The algorithm is trained over the users from the training set and evaluated over the users in the test set [21]. In this paper, we evaluated the accuracy of recommendations using a "cold-start" protocol on the data set. First, we randomly selected 860 (80%) of the items to be training item set, leaving 217 (20%) as testing item set. Then, we selected 500 users with the least item ratings to be test users.

Since each test user had rated two sets of items, i.e., training item set and testing item set, we can evaluated the performance of our approach by calculating the precision of a fixed length of recommendation list [8]. We first used the algorithm to derive a recommendation list based on the training items rated by a test user u. We then defined the per-user precision at the recommendation list containing *Top N* items as:

$$Precision(u) = \frac{HitNumber}{N} \tag{12}$$

where *HitNumber* is the number of items in the recommendation list that are hit in the test set of user u. Then, we averaged the resulting per-user precisions over all the 500 test users to get an average precision of the *Top N* recommendation. Figure 5 and Figure 6 show the average precisions respectively of *Top 3* and *Top 10* recommendations for different approaches with varying neighborhood sizes. Since *YOUKU* doesn't provide results for *Top 10* recommendations, Figure 6 only give out the results of the CF-based approaches.

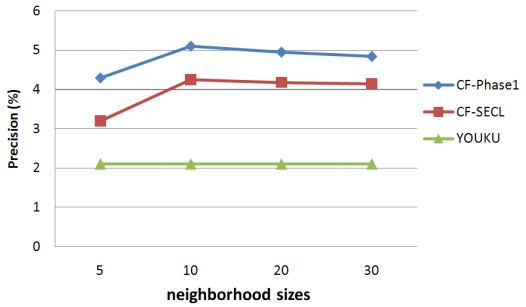

Figure 5. The Results of Average Precisions of *Top 3* Recommendations.

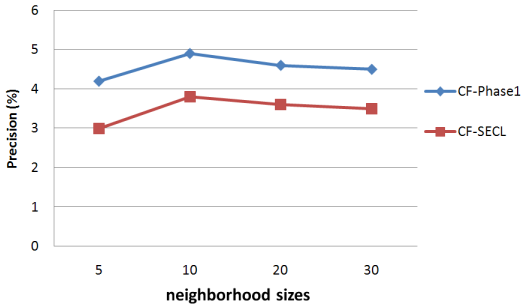

Figure 6. The Results of Average Precisions of *Top 10* Recommendations.

From Figure 5 and Figure 6 we can see that both the two CF-based fusion approaches obviously outperform the YOUKU approach (2.1%). These figures also show that the optimum neighborhood sizes k for the two CF-based are the same (i.e., 10) which lead to best precisions of 5.1% (*CF-Phase1*) and 4.2% (*CF-SECL*) at *Top 3* recommendations.

It is worthwhile to note that the *CF-Phase1* approach achieves better results than the *CF-SECL* approach in both *Top 3* and *Top 10* recommendations regarding all the three different neighborhood sizes. That suggests *Phase1* approach is more capable of accurately building the virtual *Rating Matrixes* for the CF recommendations than the *SECL* approach.

Because the top-ranked videos are usually more noticeable to the online users, the precisions at *Top 3*

recommendations is more worthwhile to be noted in producing better recommendations. From the above two Figures, we can see that the precisions at *Top 3* recommendations set are also slightly better than those at *Top 10* recommendations respectively by the *CF-Phase1* and *CF-SECL* approaches. In particular, *CF-Phase1* achieves the best result of 5.1% at *Top 3* recommendations.

VII. Conclusion And Future Work

In this paper, we improved the SECL model to meet the special characteristics of the real Chinese online video reviews and furthermore developed an online video recommender algorithm that particularly exploited the sentiment classification results to automatically build the virtual *Rating Matrix*. The experimental results through both the evaluations on sentiment classification results and recommendations show that our new approach achieves higher performance in augmenting the video recommendations in realistic online environments.

In fact, since our algorithm has no restriction on the type of collaborative-filtering algorithms used, the method can be easily scaled and incorporated into other types of CF recommenders, such as using Trust Inferences in [16] and Boltzman Machines in [8]. Our contribution is indeed primarily to the generation of the virtual *User-Item Rating Matrix* based on user reviews, so as to complement the rating sparsity limitation of current video-sharing sites when they attempt to apply the standard pure rating based CF techniques. Moreover, the virtual *Rating Matrix* in our system can in essence contain any categorical values, besides the ternary rating (+1/-1/0) that were assumed in this paper.

In the future, we will perform more studies to further optimize our algorithm, including the determination of the optimal value for the parameter α in Equation (4) so as to get the best balance between word sentiment score $RS^{W}(r)$ and emoticon sentiment score $RS^{E}(r)$ for the sentiment classification. We will be also engaged in further classifying the reviews into more delicate categories in addition to the "positive" and "negative" ones. On the other hand, the similar experimental procedures will be conducted on reviews in other languages (e.g. English) in order to validate our method's cross-language applicability.

References

[1] Ethem Alpaydin. 2004. Introduction to Machine Learning. The MIT Press.

[2] Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, Borovets, BG.

[3] M. Balabanovic. 1998. Exploring versus exploiting when learning user models for text recommendation. User Modeling and User-Adapted Interaction, 8(4):71–102.

[4] John Blitzer, Mark Drezde, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, CZ.

[5] J. S. Breese, D. Heckerman, and C. Kadie. 1998. Empirical analysis of predctive algorithms for collaborative filtering. In UAI '98, pages 43–52.

[6] C. Christakou and A. Stafylopatis. 2005. A hybrid movie recommender system based on neural networks. In Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications, Wroclaw, Poland.

[7] Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the Peanut gallery: opinion extraction and semantic classification of product documents. In Proceedings of WWW2003, Budapest, HU.

[8] Asela Gunawardana, Christopher Meek. 2009. A Unified Approach to Building Hybrid Recommender Systems. In RecSys '09, pages 117–124, New York, USA.

[9] Gayatree Ganu, Noémie Elhadad, Amélie Marian. Beyond the Stars: Improving Rating Predictions using Review Text Content. Twelfth International Workshop on the Web and Databases, Providence, Rhode Island, USA, 2009.

[10] Marco de Gemmis, Pasquale Lops, Giovanni Semeraro and Pierpaolo Basile. 2008. Integrating Tags in a Semantic Content-based Recommender. In RecSys '08, pages 163–170, Lausanne, Switzerland.

[11] Niklas Jakob, Stefan Hagen Weber, Mark-Christoph Müller, Iryna Gurevych. Beyond the Stars: Exploiting Free-Text User Reviews to Improve the Accuracy of Movie Recommendations. TSA'09, Hong Kong, China, 2009.

[12] C. W. ki Leung, S. C. fai Chan, and F. lai Chung. Integrating collaborative filtering and sentiment analysis: A rating inference approach. In ECAI-Workshop on Recommender Systems, 2006, pp. 62–66.

[13] J.C. Na, H. Sui, C. Khoo, S. Chan and Y. Zhou. 2004. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In I.C. McIlwaine (Ed.), Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference, pages 49-54, Wurzburg, Germany: Ergon Verlag.

[14] Bo Pang, Lilian Lee, and Shrivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP 2002.

[15] Jonghun Park, Sang-Jin Lee, Sung-Jun Lee, Kwanho Kim, Beom-Suk Chung, Yong-Ki Lee. 2010. An Online Video Recommendation Framework Using View Based Tag Cloud Aggregation. IEEE MultiMedia, IEEE computer Society Digital Library. IEEE Computer Society.

[16] M. Papagelis, D. Plexousakis, and T. Kutsuras. 2005. Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inferences. Proc. 3rd Int'l. Conf. Trust Management (iTrust 05), Springer, pp. 224–239.

[17] Likun Qiu, Weishi Zhang, Changjian Hu, Kai Zhao. 2009. SELC: A Self-Supervised Model for Sentiment Classification. In Proceedings of CIKM'09, Hong Kong, China.

[18] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL-2005 Student Research Workshop, Ann Arbor, MI.

[19] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. 2000. Analysis of recommendation algorithms for e-commerce. In Proceedings of the 2nd ACM Conference on Electronic Commerce (EC'00), pages 285–295.

[20] M. V. Setten and M. Veenstra. 2003. Prediction strategies in a TV recommender system – method and experiments. In Proceedings of International World Wide Web Conference, Budapest, Hungary.

[21] Guy Shani, Max Chickering, Christopher Meek. 2008. Mining Recommendations From The Web. In RecSys '08, pages 35–42, Lausanne, Switzerland.

[22] Karen H. L. TsoSutter, Leandro Balby Marinho and Lars Schmidt-Thieme. 2008. Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms. In Proceedings of SAC'08, pages 1995-1999, Fortaleza, Brazil.

[23] Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of documents. In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics, New Brunswick, N.J.

[24] Bo Yang, Tao Mei, Xiansheng Hua, Linjun Yang, Shiqiang Yang and Mingjing Li. 2007. Online Video Recommendation Based on Multimodal Fusion and Relevance Feedback. In Proceedings of CIVR'07, pages 73–80, Amsterdam, The Netherlands.

[25] Taras Zagibalov and John Carroll. 2008. Unsupervised Classification of Sentiment and Objectivity in Chinese Text. In Proceedings of the 3rd International Joint Conference on Natural Language Processing.

[26] http://www.youku.com

[27] http://www.youtube.com

[28] http://www.ku6.com/

[29] http://www.tudou.com/