# Automatically Ranking Reviews Based on the Ordinal Regression Model

Bing Xu, Tie-Jun Zhao, Jian-Wei Wu, and Cong-Hui Zhu

School of Computer Science and Technology,
Harbin Institute of Technology
Harbin, China
{xb,tjzhao,jwwu,chzhu}@mtlab.hit.edu.cn

**Abstract.** With the rapid development of Internet and E-commerce, the quantity of product reviews on the web grows very fast, but the review quality is inconsistent. This paper addresses the problem of automatically ranking reviews. A specification for judging the reviews quality is first defined and thus ranking review is formalized as ordinal regression problem. In this paper, we employ Ranking SVM as the ordinal regression model. To improve system performance, we capture many important features, including structural features, syntactic features and semantic features. Experimental results indicate that Ranking SVM can obviously outperform baseline methods. For the identification of low-quality reviews, the Ranking SVM model is more effective than SVM regression model. Experimental results also show that the unigrams, adjectives and product features are more effective features for modeling.

**Keywords:** *S*entiment Analysis, Review Ranking, Ordinal Regression model, SVM Ranking.

## 1 Introduction

With the rapid development of Internet and E-commerce, the quantity of product reviews on the web grows very fast, but the review quality is inconsistent. Due to the fact that the number of reviews is too large, sometimes it is impossible for the users to read each comment by themselves. Especially, when some useful reviews locate at the end of all the reviews, the user may not be patient enough to find and read them. So it is very important to evaluate the quality of product reviews and rank reviews.

Ranking reviews is different from ranking search results. Because reviews are directly relevant to the product in the ranking reviews, assessing relevance is no longer important. Users want to browse the useful information such as product quality, service, etc. from reviews. Therefore, providing the detailed and multi-point subjective reviews is very valuable for users. A key problem of ranking reviews is how to determine which review is helpful and valuable, and then rank it to the top of review list.

Most websites can provide several ways of ranking reviews, including publication time, the number of voting according to the review helpfulness, the number of

H. Deng et al. (Eds.): AICI 2011, Part III, LNAI 7004, pp. 126–134, 2011.

responses reviews, product rating, etc. In contrast with other ways, manual voting is a more effective measure of providing valuable reviews for users. For example, on dangdang.com website, an interface allows customers to vote whether a particular book review is helpful or not. However, the issue of manual voting is that newly published reviews are always ranked at the bottom of the reviews and can not easily be found. The accumulation of votes takes time for a review to be listed at the front. So an automatically method for ranking reviews is very useful for ranking the new reviews and rarely voted reviews. Existing studies [1] [2] used these users' votes for training ranking models to assess the quality of reviews, which therefore are suffered from bias, including the imbalance vote bias, the winner circle bias, and the early bird bias.

In this paper, we propose a standard to assess the quality of book reviews and apply SVM Ranking to rank the book reviews. Firstly, the specification of reviews quality is defined, and then different classes of features are selected for modeling, including structural, lexical, syntactic, semantic features. Experiment results show that the proposed approach can rank high quality reviews to the front effectively.

The rest of the paper is organized as follows: Section 2 introduces the related work. In Section 3, we describe the specification of assessing the reviews quality and ordinal regression model. Section 4 reports experimental results and analysis. Section 5 summarizes our work in the paper and points out the future work.

## 2 Related Work

The task of ranking product reviews is related to a variety of research areas, including evaluation of reviews quality, opinion mining and learning to rank etc.

In the recent years, the evaluation of reviews quality attracts more and more researchers' attentions. Most researchers considered the problem as a ranking and solved it with classification, regression models etc. [1-7]. Liu et al. adopt SVM classification model to discriminate low-quality from high-quality reviews and define a standard specification to measure the quality of product reviews. Kim et al. and Zhang et al. presented SVM regression model to rank reviews and used the assessing information derived from users' votes of helpfulness provided by websites in the training and testing processing. Liu et al. proposed a nonlinear regression model for the helpfulness prediction and depended on three groups of factors to build model. The experimental results showed that their model is better than the SVM regression model. Zhang et al. proposed a linear model to predict the helpfulness of online product reviews. Lim et al. used scoring methods to measure the degree of spam for detecting users generating spam reviews or review spammers.

In domain of opinion mining, the focus is on distinguishing between subjective and objective texts, and mining polarity and opinions in product reviews [8-11]. Pang et al. examined several supervised machine learning methods for sentiment classification of movie reviews. Turney applied an unsupervised machine learning technique based on mutual information between document phrases and the words "excellent" and "poor" to find indicative words expressing opinions for classification. Liu and Popescu extracted and summarized users' opinions from product reviews. A sentence or a text segment in the reviews is considered as the basic unit that includes extracted opinion feature and its polarity of users' sentiments.

Learning to rank or machine-learned ranking is a type of supervised problem that is to automatically construct a ranking model from training data. Ranking models have been applied in many areas, such as information retrieval, sentiment analysis and machine translation [12] [13]. Herbrich et al. [14] proposed a new learning algorithm for the task of ordinal regression based on large margin rank boundaries. In his paper, this method is applied to the information retrieval task that is learning the order of documents according to an initial query. Joachims [15] proposed learning a ranking function for search as ordinal regression using click-through data. The Ranking SVM model is employed for ordinal regression.

In this paper, we formalize the ranking of review quality as ordinal regression and employed Ranking SVM model for implementation.

# 3 Modeling Review Quality

Firstly, let us formalize the problem of ranking of review quality. Given a training data set $D = \{x_i, y_i\}_{i=1}^{n}$, we construct a model that can minimize error in prediction of $y$ given $x$. Here $x_i \in X$ and $y_i \in \{excellent, good, fair, bad\}$ represent a review and a label, respectively. When applied a new instance $x$, the model predicts the corresponding $y$ and outputs the score of the prediction.

## 3.1 Ranking SVM Model

Classifying instances into the categories, “excellent”, “good”, “fair”, “bad”, is problem of a typical ordinal regression, for there is an order between the four categories.

Ranking SVM is employed as the model of ordinal regression. Given an instance $x$, Ranking SVM assigns a score to it based on

$$U(x) = \omega^T x \tag{1}$$

Where $\omega$ represents a vector of weights. The higher the value of $U(x)$ is, the better the quality of the instance $x$ is. In ordinal regression, the value of $U(x)$ are mapped into intervals on the real line and the intervals correspond to the ordered categories. An instance that falls into one interval is classified into the corresponding ordered category. In our method of opinion ranking, we adopt scores output by a Ranking SVM.

## 3.2 Specification of Quality

In this section, four categories of review quality are defined, including “excellent review”, “good review”, “fair review” and “bad review”. The different definitions of review quality represent different values of the reviews to users’ purchase decision.

An excellent review is a comprehensive and detailed review or a relatively complete and detailed review on a product. It describes most aspects of a product and gives credible opinions with sufficient evidence. Usually an excellent review will be served as the main reference by users before making their purchase decision.

A good review presents some aspects of the product, but it is not a complete and detailed evaluation. It only supplies a brief description which can not offer users sufficient reference on a product.

A fair review contains a brief description on an aspect of the product. It often occurs with a short sentence and only provides less information.

A bad review is an unhelpful description that can be ignored on a product. It talks about some topic that is not related to the product. Sometimes users spend a lot of time to read, but do not get any valuable information.

### 3.3 Feature Selection

We experimented with seven features, including structural, lexical, syntactic and semantic features. Table 1 shows the list of the features.

Lexical feature captures the unigram feature of word. We calculate number of each word occurring in the review.

The number of sentences treated as the structural feature represents review length. This feature is considered to be related with information content of review.

Syntactic feature aims to capture the linguistic features of the review. In grammar, part-of-speech of a word is a linguistic category defined by its syntactic or morphological behavior. Common POS categories are: noun, verb, adjective, adverb .etc. In the syntactic features, the numbers of adjective and adverb tokens are calculated.

**Table 1.** Description of Feature Set

| ID | Feature Set |
|---|---|
| 1 | The occurring frequency of each word in a review |
| 2 | The number of sentences in a review |
| 3 | The number of adverb in a review |
| 4 | The number of adjective in a review |
| 5 | The number of sentiment word in a review |
| 6 | The number of product features in a review |
| 7 | The occurring frequency of product features in a review |

We hypothesized the high quality review will include the product features and sentiment words information, frequency and the number of product features are calculated and treated as the semantic features in the review. Product features in the review are better indicator of review quality. However, product feature extraction is not an easy task. In this paper, we adopt an unsupervised clustering approach to extract product features. Firstly, we adopt the shallow parser to identify noun phrases from reviews, and then k-means clustering approach is used to group noun phrase. To avoid the side effect of arbitrarily selecting the initial cluster centers, some representative product features are selected as the initial cluster centers. Experiment shows that supervised selecting of the cluster centers can increase precision of product attribute clustering. But there are many noise data in product features extracted by clustering. We remove these noun phrases that are far from cluster centers by setting the threshold. Finally, we check identification results manually for high performance.

Sentiment words are positive and negative words that correlate with product names and product features. We capture these features by extracting sentiment words using the publicly available list from the Hownet[1]. On one hand, the features of sentiment words help to distinguish subjective reviews from all reviews. On the other hand, the number of sentiment words is larger in a high quality review.

# 4 Experiment

In this section, we will describe our experimental setup, show our results and analyze performance.

## 4.1 Experiment Data

We use Ranking SVM model to verify the effectiveness of the proposed approach of review ranking. Because there is no standard data set for evaluation, we randomly collected 50 books as samples from a famous book-store website-dangdang.com.[2] After filtering the duplicated reviews, we got 2907 reviews and created a data set. Table 2 shows statistics on the data.

**Table 2.** Statistics of Review Quality

| Review | Number |
|---|---|
| Number of reviews | 2907 |
| Number of excellent reviews | 188 |
| Number of good reviews | 1592 |
| Number of fair reviews | 860 |
| Number of bad reviews | 330 |

Two annotators labeled the review independently with our definitions of review quality as their guideline. The value of kappa statistic is used to validate the effectiveness of labels by human. We found the two annotators achieved highly consistent results according to 0.8806 kappa value.

## 4.2 Measure of Evaluation

Evaluating the quality of ranking reviews is difficult. Five measures for evaluation of review ranking are adopted, including NDCG (Normalized Discounted Cumulative Gain), modified R-precision measures and Error_rate.

The following give the details.

$$NDCG @ n = Z_n \sum_{j=1}^{n} \begin{cases} 2^{r(j)} - 1, & \mathrm{j} = 1, 2 \\ (2^{r(j)} - 1) / \log(1 + j), & \mathrm{j} > 2 \end{cases} \tag{2}$$

[1] http://www.keenage.com/

Where j is the position in the review list, r(j) is the score of the j-th review in the review list, and Zn is a normalizing factor.

R-precision is precision at rank R. Where R is the total number of same score reviews from the book. In the experiment, we calculate the three modified R-precision measures for different value of reviews.

$$R\text{ - precision\_E (book}_{\text{i}}) = \frac{|\text{" excellent"  reviews at R top ranked reviews}|}{\text{R}} \tag{3}$$

Where R is the number of “excellent”reviews from book$_i$.

$$R\text{ - precision\_E \& G (book}_{\text{i}}) = \frac{|\text{" excellent" and " good" reviews at  R top ranked reviews}|}{\text{R}} \tag{4}$$

Where R is the number of “excellent”and “good” reviews from book$_i$.

$$R\text{ - precision\_B (book}_{\text{i}}) = \frac{|\text{" bad" reviews at R bottom ranked reviews}|}{\text{R}} \tag{5}$$

Where R is the number of “bad” reviews from book$_i$.

$$\text{R - precision } = \frac{\sum_{\text{i}=1}^{T} \text{R - precision(book}_{\text{i}})}{T} \tag{6}$$

Where T is the number of books in data set. Using the formula (3), (4) and (5), we got R-precision_E, R-precision_E&G and R-precision_B.

Error_rate represents the error rate of all ranked pairs.

$$Error\_rate = \frac{|\text{error pairs at ranked reviews}|}{|\text{ all pairs at ranked reviews }|} \tag{7}$$

### 4.3 System Performance and Analysis

In the experiment, we use 5-fold cross validation. The 50 books and their corresponding reviews are divided randomly into five subsets. One of the five subsets was treated as the test set and the other four subsets were combined to form a training set. The ranking model is trained on the training set and tested on the test set.

We use two baseline methods to compare performance of ordinal regression model for ranking review, including regression model and method of helpfulness voting from customers. For ordinal regression model and regression model, SVM ranking and SVM regression tool SVM$^{\text{light}}$ is deployed. The linear, polynomial and radial basis function (RBF) kernels are tested on development data respectively. Finally, we find that the best results are shown using the linear kernel. Table 3 reports the average results of the five trials for all features combination with 95% confidence bounds.

In table 3, we find that both Ranking SVM and SVM regression outperform Voting method significantly. The results show that ordinal regression and regression models are significant and effective for review ranking. In particular, the results of Ranking SVM indicate that ordinal regression model is more effective than regression model with our all features in the five evaluation measures, but R-precision_E is different

**Table 3.** Results of Book Reviews Ranking

| Evaluation Measures | Voting Method | SVM Regression | Ranking SVM |
|---|---|---|---|
| NDCG-10 | 0.6443 | 0.8708 | **0.8840** |
| NDCG-20 | 0.6713 | 0.8847 | **0.8959** |
| R-precision_E | 0.1829 | **0.4358** | 0.4230 |
| R-Precision_E&G | 0.6281 | 0.8418 | **0.8643** |
| R-precision_B | 0.1443 | 0.3575 | **0.5311** |
| Error_rate | 0.4484 | 0.1492 | **0.1266** |

from other measures. Therefore we conclude that SVM regression is better to identify the excellent reviews, and Ranking SVM is more efficient to rank the top-10 and top-20 reviews. For bad reviews, Ranking SVM has more advantages over SVM regression.

To compare the effectiveness of individual features, we analyze the R-precision_E&G and the R-precision_B drop when each feature is removed from Ranking SVM and SVM regression in the Fig.1 and the Fig.2. The description of features is shown in Table 1.

From Fig.1 and Fig.2, we can find that the effectiveness of each feature is similar for Ranking SVM and SVM regression. The performance drops significantly when feature 1, 4, 6 and 7 is removed. So unigram features (feature 1) are very important for the ranking task. Comparing with unigram features, we didn’t add bigram features since they are suffered from the data sparsity of short reviews. We can also find that the adjective feature (Feature 4) is an active feature, but the feature of sentiment word (feature 5) and the adverb feature (Feature 3) are negative features. The more adjectives a review has the more likely the review is good. On the other hand, negative effect of the sentiment word feature likely comes from adverb because the sentiment words include the adjectives and adverbs. The product feature (Feature 6 and 7) is an important semantic feature and their occurrence in a review correlate with review quality. In the high quality user-generated product review, the role of product feature becomes more effective.

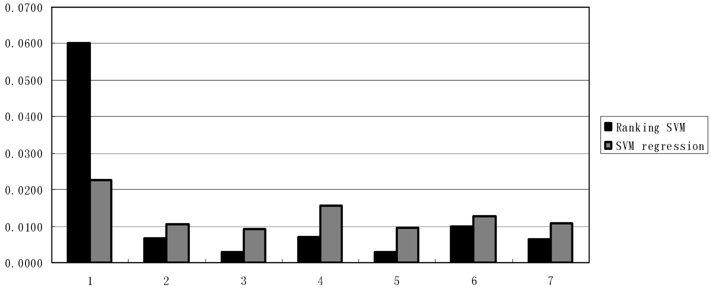

**Fig. 1.** The Results of R-precision_E&G Drop

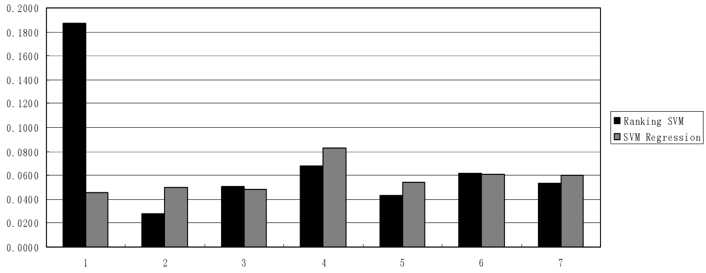

**Fig. 2.** The Results of R-precision_B Drop

Finally, we analyze the error results of Ranking SVM and SVM regression. We find that few subjective reviews sorted at the top are not associated with the book discussed. These reviews describe another book and were recommended to customers. In fact, these reviews should be labeled as "bad reviews". But this problem can not be solved using our proposed approach. Moreover, we also find that limitation of the features and annotation errors deteriorate the performance of ranking reviews.

## 5 Conclusion

For many online websites of product reviews, ranking review according to review quality is a key issue. But ranking of voting according to user helpfulness can not list these new and good reviews on the top in a short time. This paper proposed a new approach to automatically evaluate quality and rank review. Firstly, the specification of review quality is defined and a corresponding score is labeled. We trained a Ranking SVM system using the several effective features and then applied them to rank unlabeled reviews. For the task of ranking reviews, we use NDCG, three modified R-precision measures and proposed error rate to effectively assess the ranking results. Experimental results indicate that Ranking SVM outperforms the voting method significantly and is better than the SVM regression in the five measures. We also give a detailed analysis of effectiveness of each feature and conclude that unigrams, adjectives and product features are most useful.

In the future work, we hope to propose better discriminating standard for different reviews and validate the effectiveness of ranking reviews in other applications.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China (No. 61073130) and the National High-Tech Development 863 Program of China (No. 2006AA010108) and MOE-MS Key Laboratory Opening Funding of Natural Language Processing and Speech (HIT.KLOF.2009019).

## References

1. Kim, S.M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automacically Assessing Review Helpfulness. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 423–430 (2006)

2. Liu, J.J., Cao, Y.B., Li, C.Y., Huang, Y.L., Zhou, M.: Low-Quality Product Review Detection in Opinion Summarization. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 334–342 (2007)
3. Zhang, R.C., Thomas, T.: Helpful or Unhelpful: A Linear Approach for Ranking Product Reviews. Journal of Electronic Commerce Research 11(3), 220–230 (2010)
4. Zhang, Z., Varadarajan, B.: Utility Scoring of Product Reviews. In: CIKM 2006, pp. 52–57 (2006)
5. Liu, Y., Huang, X., An, A., Yu, X.: Modeling and Predicting the Helpfulness of Online Reviews. In: Proceedings of the 8th International Conference on Data Mining, pp. 443–452 (2008)
6. Lim, E.-P., Nguyen, V.-A., Jindal, N., et al.: Detecting product review spammers using rating behaviors. In: Proceedings of the International Conference on information and Knowledge Management (2010)
7. Lei, Z., Bing, L., Hwan, L.S., O'Brien-Strain, E.: Extacting and Ranking Product Features in Opinion Documents. In: Coling 2010, pp. 1462–1470 (2010)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: EMNLP 2002, pp. 79–86 (2002)
9. Turney, P.: Thumbs up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: ACL 2002, pp. 417–424 (2002)
10. Liu, B., Hu, M., Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web. In: Proc. International World Wide Web Conference, pp. 342–351 (May 2005)
11. Popescu, A.M., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of Empirical Methods in Natural Language Processing, pp. 339–346 (2005)
12. Xu, J., Cao, Y.-B., Li, H., Zhao, M., Huang, Y.-L.: A supervised Learning Approach to Search of Definations. Journal of Computer Science and Technology 21(3), 439–449 (2006)
13. Liu, T.-Y.: Learning to Rank for Information Retrieval. Foundations and Trends in Information Retrieval 3(3), 225–331 (2009)
14. Herbrich, R., Graepel, T., Obermayer, K.: Suppert vector learning for ordinal regression. In: Proceedings of 9th International Conference Artificial Neural Networks, pp. 97–102 (1999)
15. Joachims, T.: Optimizing Search Engines Using Click-through Data. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142 (2002)