Seeing Several Stars: A Rating Inference Task for a Document Containing Several Evaluation Criteria

Kazutaka Shimada and Tsutomu Endo

Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Iizuka Fukuoka 820-8502 Japan
{shimada, endo}@pluto.ai.kyutech.ac.jp

Abstract. In this paper we address a novel sentiment analysis task of rating inference. Previous rating inference tasks, which are sometimes referred to as "seeing stars", estimate only one rating in a document. However reviewers judge not only the overall polarity for a product but also details for it. A document in this new task contains several ratings for a product. Furthermore the range of the ratings is zero to six points (i.e., stars). In other words this task denotes "seeing several stars in a document". If significant words or phrases for evaluation criteria and their strength as positive or negative opinions are extracted, a system with the knowledge can recommend products for users appropriately. For example, the system can output a detailed summary from review documents. In this paper we compare several methods to infer the ratings in a document and discuss a feature selection approach for the methods. The experimental results are useful for new researchers who try this new task.

Keywords: Sentiment analysis, Rating inference, Review mining.

1 Introduction

As the World Wide Web rapidly grows, a huge number of online documents are easily accessible on the Web. Finding information relevant to user needs has become increasingly important. The most important information on the Web is usually contained in the text. We obtain a huge number of review documents that include user's opinions for products. Buying products, users usually survey the product reviews. More precise and effective methods for evaluating the products are useful for users. Many researchers have recently studied extraction and classification of opinions [6, 10, 11, 12, 14, 15].

There are many research areas for sentiment analysis; extraction of sentiment expressions, identification of sentiment polarity of sentences, classification of review documents and so on. In this paper we address a new sentiment analysis task of review documents. Most of existing studies for classification of review documents have handled two polarities: positive and negative opinions [10, 12]. On the other hand, several researchers have challenged not only p/n classification but also rating inference, namely seeing stars in a review document [8, 9]. We also handle a rating inference task in this paper.

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 1006–1014, 2008.

The previous studies, p/n classification and rating inference, contain a problem; a document includes only one polarity (or stars). They did not discuss a task handling several polarities in a document. However, reviewers judge not only the overall polarity for a product but also details for it. For example, they are "performance", "user-friendliness" and "portability" for laptop PCs and "script", "casting" and "music" for movies.

In this paper we deal with a document containing several sentiment polarities. It is a new task for sentiment analysis: seeing ***several*** stars in a document. This is a primary experiment for the task. To estimate several ratings in a document is beneficial for users. Furthermore it is important for sentiment analysis tasks to extract words or phrases that relate to each polarity (evaluation criteria). Zhuang et al. have reported a method of movie review mining and summarization using the discovered p/n information [15]. If significant words or phrases for an evaluation criteria and their strength as positive or negative opinions are extracted, a system with knowledge that consists of them can recommend products for users appropriately. For example, the system can output a detailed summary from review documents: it generates not only a simple summary "This movie is good", but also a more detailed summary "The story of this movie is excellent (five stars), but the music might be substandard (two stars)".

In this paper we compare several methods for the rating inference task. Also we compare some feature sets for SVR in this task and discuss solutions for the improvement of accuracy. The experimental results are useful for new researchers who try this new task.

2 Task

There are many review documents of various products on the Web. In this paper we handle review documents about game softwares. Figure 1 shows an example of a review document. The review documents consist of evaluation criteria, their ratings, positive opinions, negative opinions and comments for a product. The number of evaluation criteria is 7: "Originality", "Graphics", "Music", "Addiction", "Satisfaction", "Comfort", and "Difficulty". The range of the ratings, e.g. stars, is zero to six points.

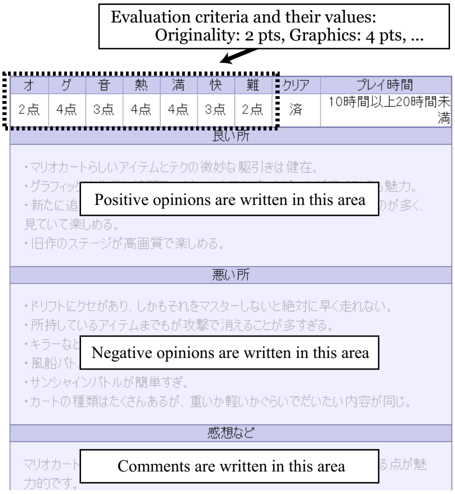
We extract review documents from a Web site[1]. The site establishes a guideline for contributions of reviews. In addition, the reviews are checked by the administrator of the site. As a result, the reviews unfitting for the guideline are rejected. Therefore the documents on the site are good quality reviews.

3 The Methods and Features

3.1 The Methods

In this section we describe 4 methods, which are SVM, SVR, Maximum entropy and a similarity based method, for inferring the ratings in a document.

[1] `http://ndsmk2.net`


Evaluation criteria and their values:
Originality: 2 pts, Graphics: 4 pts, ...
Positive opinions are written in this area
Negative opinions are written in this area
Comments are written in this area

オ	グ	音	熱	満	快	難	クリア	プレイ時間
2点	4点	3点	4点	4点	3点	2点	済	10時間以上20時間未満

良い所

・マリオカートらしいアイテムとテクの微妙な駆引きは健在。
・グラフィッ も魅力。
・新たに追 のが多く、
見ていて楽しめる。
・旧作のステージが高画質で楽しめる。

悪い所

・ドリフトにクセがあり、しかもそれをマスターしないと絶対に早く走れない。
・所持しているアイテムまでもが攻撃で消えることが多すぎる。
・キラーなど
・風船バト
・サンシャインバトルが簡単すぎ。
・カートの種類はたくさんあるが、重いか軽いかぐらいでだいたい内容が同じ。

感想など

マリオカート る点が魅
力的です。

Fig. 1. An example of a review document

SVM and SVR. SVMs are a machine learning algorithm that was introduced by [13]. We expand the binary SVMs into a multi-classifier by using one-versus-one methods. Also we employ linear support vector regression (SVR). This is one of straightforward methods for this task. Related studies also used SVR for the rating inference task. We use the SVM^{light} package[2] for training and testing, with all parameters set to their default values [4].

ME. Maximum entropy modeling (ME) is one of the best techniques for natural language processing [1]. In this paper we use Amis[3], which is a parameter estimator for maximum entropy models. We estimate parameters by using the generalized iterative scaling algorithm.

SIM. The 4th method is based on a similarity measure. We use the cos measure for the similarity calculation as follows:

$$sim(tr_x, te_y) = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \times \sum_{i=1}^{N} y_i^2}} \tag{1}$$

where tr and te are a document in training data and a document in test data respectively. x_i and y_i are the value of a word i in tr and tr respectively. Next we extract documents of which the similarity exceeds a threshold. For the extracted documents, we compare the average values of each evaluation criterion. Finally we output the values as the result of the method.

[2] http://svmlight.joachims.org

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/amis/index.html

3.2 Feature Selection

For the features of the methods, we use words appearing in positive and negative opinions in review documents. We do not use words in comment areas because the accuracy with them in a preliminary experiment was lower than that without them. Here we distinguish words in the positive opinion areas and the negative opinion areas. In other words, for a word w_i, the word in the positive opinion areas is w_i^p and the word in the negative opinion areas is w_i^n. A vector of an evaluation criterion y for a document d_x is as follows:

$$d_{xy} = \{w_1^p, w_2^p, ..., w_j^p, w_1^n, w_2^n, ..., w_j^n\}$$

where j is the number of words appearing in review documents. We select words belonging to "noun", "verb", "adjective" and "adverb". We use ChaSen for the morphological analysis[4]. The value of the features is based on the word frequency.

Next we consider two extensions for the feature selection. One approach is to use more complex information. In this paper, we use a word sequential pattern between two words in each sentence, namely cooccurrence. In the pattern extraction, we allow a skip between words. We extract word pairs within a length that we define. For example, we obtain the patterns "Fighting::WiFi, Fighting::excited, Fighting::me, WiFi::excited, WiFi::me, excited::me" from a sentence "Fighting with WiFi excited me."

Another approach for improvement of the accuracy of a classifier is to select effective and significant features for the feature space. Furthermore it seems unlikely that all words in a document contribute to all evaluation criteria. In other words some words that are significant to estimate the rating of an evaluation criterion exist in a review document. To extract the words, we compute a confidence measure of each word. The confidence measure in this paper is variance of words concerning each evaluation criterion. We measure whether a word appears frequently with the same point regarding an evaluation criterion. It is computed as follows:

$$var(w_{c_j}) = \frac{1}{m} \sum_{i=0, w \in d_i}^{n} (real(d_i, c_j) - ave(w_{c_j}))^2 \qquad (2)$$

where c_j is an evaluation criterion. m and n are the document frequency (df) of a word w (or a word pair) and the number of documents respectively. $real(d_i, c_j)$ and $ave(w_{c_j})$ are the actual rating of c_j in d_i and the average score of w for c_j. We use w of which the var is a threshold or less.

Furthermore we apply two conditions to the feature selection.

Frequency (F). The frequency of a word is n times or more.

Evaluation value (E). If a word w appears in "positive opinion area", the actual rating of the evaluation criterion have to be 3 points or more. If a word w appears in "negative opinion area", the actual rating of the evaluation criterion have to be 3 points or less.

[4] http://chasen.naist.jp/hiki/ChaSen/

4 Experiment

In this section, we explain datasets and criteria for the experiment first. Then we evaluate our method with a dataset and discuss the experimental results.

4.1 Dataset and Criteria for the Experiment

We evaluated this new sentiment analysis task with a dataset that consists of 1114 review documents that consist of different kinds of game softwares such as RPGs and action games of Nintendo DS, namely a mixed dataset. In this experiment we evaluated the dataset with 5-fold cross-validation.

In this experiment, we evaluated the outputs of each method with the following criteria: the mean squared error (MSE) between actual ratings and outputs of each method, the standard deviation (SD) of the MSE, and the accuracy . The mean squared error (MSE) is computed as follows:

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}(out(d_{ij}) - real(d_{ij}))^2 \qquad (3)$$

where i and j denote a review document and an evaluation criterion in the document respectively. *out* and *real* are the output of a method and the actual rating in a document respectively. We converted the outputs of the SVR and the similarity based method into integral value with half adjust because it was continuous. The MSE is one of important criteria for the rating inference task because not all mistakes of estimation with the methods are equal. For example, assume that the actual rating of a criterion is 4. In this situation, the mistake of estimating it as 3 is better than the mistake of estimating it as 1.

In this experiment, we used two types of accuracy. The first accuracy is simple accuracy, that is to say the correspondence between real ratings and outputs. The second one is PNN accuracy (Positive-Neutral-Negative). For the PNN accuracy, we defined 4 and 5 points as “Positive”, 3 points as “Neutral” and 0, 1, 2 points as “Negative”.

4.2 Results

First we compared the methods with bag-of-words (Bows) features only. We ran the SVR and SVM with all default parameters in this experiment. For the Maximum entropy we estimate parameters by using the generalized iterative scaling algorithm.

Table 1 shows the result. “All-3” in the table is the MSE in the assumption that the ratings of all criteria are 3. “Ave” is the MSE computed from actual ratings and average values of each evaluation criterion in the training data. The average values are discretized for the MSE computation. These MSEs are baselines for this task. As you can see, all methods outperformed the baselines[5].

[5] We evaluated the Naive Bayes classifier and C4.5 with the same dataset. However, the MSEs of them were larger than the average-based baseline.

Table 1. Comparison with baselines

		All-3	Ave	SVR	SVM	ME	SIM
MSE	Originality	1.26	1.54	**0.88**	0.91	0.98	1.03
	Graphics	1.03	0.85	**0.74**	0.78	0.82	0.84
	Music	1.21	0.79	0.70	**0.69**	0.75	0.77
	Addiction	1.89	1.89	**1.21**	1.54	1.44	1.45
	Satisfaction	1.97	1.77	**1.22**	1.54	1.57	1.42
	Comfort	1.29	1.29	**1.13**	1.24	1.35	1.27
	Difficulty	1.74	**1.17**	1.22	1.23	1.35	1.18
	Average	1.48	1.33	**1.02**	1.13	1.18	1.14
	SD	0.17	0.24	**0.12**	0.19	0.19	0.20
Accuracy		26.60	37.69	41.37	**41.76**	40.23	39.47
PNN Accuracy		26.60	51.98	57.41	**58.43**	57.05	55.71

Table 2. The effectiveness of *var*

var	0.25	0.5	0.75
MSE (Ave)	0.99	0.99	0.99
MSE (SD)	0.12	0.12	0.12
Accuracy	41.40	41.59	41.56
PNN Acc	57.49	57.49	57.46

In this experiment, the SVMs produced the best accuracy. However the MSE of the SVR was the smallest of them. The SD of the SVR was also small. As a result, we arrived at the conclusion that the SVR was the most suitable in this experiment because the MSE is the most important criterion in this task.

Next we compared the results concerning the extensions for the feature selection, namely word sequential patterns and a confidence measure *var* based on the variance. In this experiment, we used the SVR only for the evaluation. Here we applied the extension with *var* to word sequential patterns only. Table 2 shows the comparison of the value of *var*. In this experiment, the length for the pattern extraction was 4. The value of the condition of the frequency (F) in Section 3.2 was 1[6]. Table 3 shows the comparison of the length for the pattern extraction. The value of the *var* was 0.5. As you can see, there is no difference in the MSE and the accuracy.

Here we need to discuss a problem for this task. In this task, there is a possibility that humans even can not infer a rating in a document because a document contains many evaluation criteria. In other words, words or phrases for an evaluation criterion do not exist in a document occasionally. Therefore we inquired into 30 documents selected from review documents randomly. We judged whether we could infer each criterion in the documents or not. The criterion of the judgment was whether the document contained words or phrases for an evaluation

[6] Although we compared several conditions of the frequency (F) in this experiment, there is no difference in the MSE and the accuracy.

Table 3. The effectiveness of the patterns

Length	1	2	6
MSE (Ave)	1.02	1.00	0.99
MSE (SD)	0.12	0.12	0.13
Accuracy	41.39	41.39	41.46
PNN Acc	57.08	57.34	57.26

criterion or not[7]. As a result, approximately 75% of all criteria could be inferred by humans. We think that this is one reason that the accuracy was low. However, the judgment of the possibility of inference was examined by one test subject only. We need to discuss the reliability of the judgment process with some test subjects by using a concordance rate such as the Kappa coefficient [2].

4.3 Discussion

In this section we discuss this task on the basis of the experimental results. The accuracy in the experiment was insufficient; approximately 41% for the 5-fold cross-validation. These results show the difficulty of this "seeing several stars" task (6 grades for 7 criteria). We need to discuss the improvement of the accuracy and the MSE. We think that dictionaries obtained from opinion extraction or word polarity estimation tasks [5, 6, 14] are useful to infer the ratings in our task.

In this experiment, we used SVR to estimate the ratings in a document. The SVR is often utilized in rating inference tasks [8, 9]. However Koppel and Schler [7] have discussed a problem of use of regression for multi-class classification tasks and proposed a method based on optimal stacks of binary classifiers. Pang et al. [9] have proposed a method based on a metric labeling formulation for the rating inference problem. The results of these studies denote that SVR is not always the best classifier for this task. We need to consider other methods for the improvement of the accuracy. We have proposed high accuracy classifiers for a p/n classification task [11]. The method incorporated three classifiers: SVMs, Maximum Entropy and score calculation. In the movie review classification task [10], this multiple classifier improved the accuracy as compared with the single classifiers. Applying this method to this task is one of our future work.

The size of the dataset in this experiment was not large: 1114 documents. To generate a high accuracy classifier, we need a large amount of training data. Goldberg and Zhu [3] have argued the significance of training data acquisition from unlabeled data. As an additional experiment, we evaluated the SVR-based method with bows and patterns based on the value of *var* computed from 1114

[7] Here we did not consider the correctness of ratings estimated by us. For example, if we could infer an evaluation criterion by reading the positive opinion area in the case that the rating was 4 or 5, we judged that the evaluation criterion could be inferred.

review documents[8]. As a result, the accuracy increased by 11%[9]. We think that one reason for the improvement is the increase of training data for the *var* calculation. Therefore, we need to consider a training data extraction method.

5 Conclusion

In this paper we described a novel sentiment analysis task of rating inference. The documents in this task include 7 evaluation criteria that contain 6 rating points: seeing several stars in a document. As a primary experiment for this task we inferred the ratings in each document and compared some machine learning techniques. As a result, the support vector regression (SVR) produced the best performance. We also explained the feature selection based on variance of words and the use of word sequential patterns. The experimental results show that this is a difficult task of sentiment analysis and we need more training data. Future work includes (1) extraction of more effective features for a classifier, (2) evaluation with other classification methods.

References

[1] Berger, A.L., Della Pietra, S.A., Della Pietra, V.J.: A maximum entropy approach to natural language processing. Computational Linguistics 22(1), 39–71 (1996)

[2] Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)

[3] Goldberg, A.B., Zhu, X.: Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In: HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing (2006)

[4] Joachims, T.: Transductive inference for text classification using support vecor machines. In: Proceedings of the Sixteenth International Conference on Machine Learning, pp. 200–209 (1999)

[5] Kawano, Y., Shimada, K., Endo, T.: Sentence polarity classification based on a scoring method (in Japanese). In: HINOKUNI Symposium 2007 CD-ROM A-3-4 (2007)

[6] Kobayashi, N., Iida, R., Inui, K., Matsumoto, Y.: Opinion extraction using a learning-based anaphora resolution technique. In: Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-2005), pp. 175–180 (2005)

[7] Koppel, M., Schler, J.: The importance of neutral examples in learning sentiment. Computational Intelligence 22(2), 100–109 (2006)

[8] Okanohara, D., Tsujii, J.: Assigning polarity scores to reviews using machine learning techniques. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP), pp. 314–325 (2005)

[8] Conditions: $var = 0.25$, the frequency ≥ 2 for Bows and $var = 0.25$, the frequency ≥ 2, the length$= 4$ for patterns.

[9] However, the method with the conditions could not estimate the ratings for documents of 15% of the test data because zero vectors are often generated owing to the condition of the value of *var*. Moreover, this is a close experiment.

[9] Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 115–124 (2005)

[10] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86 (2002)

[11] Tsutsumi, K., Shimada, K., Endo, T.: Movie review classification based on a multiple classifier. In: The 21th Pacific Asia Conference on Language, Information and Computation (PACLIC) (2007)

[12] Turney, P.D.: Thumbs up? or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)

[13] Vapnik, V.N.: Statistical Learning Theory. Wiley, Chichester (1999)

[14] Wilson, T., Wiebe, J., Hwa, R.: Just how mad are you? finding strong and weak opinion clauses. In: AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (2004)

[15] Zhuang, L., Jing, F., Zhul, X.-Y.: Movie review mining and summarization. In: Proceedings of the ACM 15th Conference on Information and Knowledge Management (CIKM-2006), pp. 43–50 (2006)