Exploiting Shopping and Reviewing Behavior to Re-score Online Evaluations

Rong Zhang #1, ChaoFeng Sha *2, Minqi Zhou #1, Aoying Zhou #1
#East China Normal University, Shang Hai, China
*Fudan University, Shang Hai, China
{rzhang, mqzhou, ayzhou}@sei.ecnu.edu.cn, cfsha@fudan.edu.cn

ABSTRACT

Analysis to product reviews has attracted great attention from both academia and industry. Generally the evaluation scores of reviews are used to generate the average scores of products and shops for future potential users. However, in the real world, there is the inconsistency problem between the evaluation scores and review content, and some customers do not give out fair reviews. In this work, we focus on detecting the credibility of customers by analyzing online shopping and review behaviors, and then we re-score the reviews for products and shops. In the end, we evaluate our algorithm based on the real data set from Taobao, the biggest E-commerce site in China.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: [Data Mining]

Keywords

Review inconsistency, customer credibility

1. INTRODUCTION

A new user usually relies on the existing reviews given by the previous genuine customers before performing a transaction. Generally reviews include both review content and its corresponding real number evaluation (rating-score in short). However, although these reviews are not written by spammers, they still have two problems. Firstly there is the inconsistency between review content and rating-score. Checking the reviews from Taobao(http://www.taobao.com) and Gmarket(http://english.gmarket.co.kr), inconsistency between the review content and rating-score generally exists. For example, some customers give high rating-scores but write negative review content. Secondly some customers do not give out fair reviews. The reviews are affected by the customers' characteristics.

In this work, we define *Customer Credibility* that refers to the *reliability* to customer reviews which are considered in calculating the average scores of products and shops. We focus on detecting the credibility of the genuine customers so as to help evaluate the real quality of both product and shop service, which has not been studied so far. Our solution is based on the following observations. First, non-good-rated products is spammed with less probability than good-rated products. In Taobao-like E-Commerce sites, customers are only allowed to give reviews to their bought products, which makes it different from Amazon or Ebay in the review systems. If there is spammed non-good-rating to the competitor's product, it will cost at least the price of the corresponding product. It will be expensive. Second, many products with high rating-scores have unsatisfied review content which may be caused by subjective reasons and shall be used to revise the final scores. Third, customers, products and shops can construct complex business networks represented as a twin-bipartite-graph. By take full usage of these observations, 1) we construct a train data to train a classifier for scoring the review content to replace the rating-score and 2) we design an algorithm to calculate the customer credibility by defining a mutual reinforcement relationship on this graph, which is inspired by HITS[2]. Our work is different with spammer detection[4] or review quality assessment[3]. Spammers can be taken as the special case of our work, who have very low credibility; review quality assessment is generally to retrieve reviews with clear descriptions to product's quality instead of "good" or "bad" ratings.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. THE PROPOSED TECHNIQUE

The start point of our work is: if individual scores are consistent (inconsistent) with the scores of most customers, we can increase (decrease) individual credibilities.

The proposed method contains three modules:

◇ **Module 1 Review Content Scoring:** We take the Maximum Entropy(**ME**) model[1] to classify reviews (content) into two categories: negative and positive ones. Two steps are taken to complete the task. **Step-1** is to define feature templates for ME model, which are classified into three types. **Type-1** is the BASE Feature describing the basic information of review content, such as token number; **Type-2** is the Part-of-Speech Feature, e.g. NN, ADJ, and so on; **Type-3** is the Emotion Feature, such as emotion words. **Step-2** is to construct positive and negative training data for the ME model. For negative data, we select the reviews with low rating-scores randomly, which are supposed to be credible with high probability because of the high spamming cost. For positive ones, we can not select randomly from good reviews considering the inconsistency problem. However, we find, for the same shop, the returned customers are likely to assign high rating-scores and write positive review content. We select their good reviews as positive examples. We can then train the ME model and use it to assign the score for review content (content-score in short).

◇ **Module 2 Customer Credibility Analysis:** We model customer transactions by a twin bipartite graph called **PCS** (Product-Customer-Shop) defined as:

DEFINITION 1 (PCS GRAPH). *: A PCS graph is defined as* $PCS = (V, E, \Upsilon, \mu, \nu)$. $V = \{P \cup C \cup S\}$ *with* P, C *and* S *representing products, customers and shops;* $E \subseteq P \times C \cup C \times S$ *representing transactions;* Υ *is the credibility vector for customers;* μ *and* ν *are the functions for computing the customer credibility from PC sub-graph and CS sub-graph respectively.*

Based on this graph, we design an iteration algorithm to evaluate the customer credibility. And our algorithm relies on the following assumptions:

- **Product /shop:** If credible customers give high evaluations to the (shop's) product, we think the product (shop) is good. Good products (shop) will own high percentage of good customers.
- **Customer:** If good (bad) products belonged to good (bad) shops are given low (high) scores by a customer, it supposes to reduce the credibility of the customer. If good(bad) products/shops are given high(low) scores by a customer, it supposes to improve the credibility of the customer.

The key is to design functions μ *and* ν for the iteration algorithm. Each function takes three steps: **Scoring**, **Feedback**, **Normalization**. We iterate the three steps until the credibility values keep unchanged. Initially, we assign uniform credibility values to customers. We only describe μ for PC sub-graph here due to the space limitation. And ν is similar to μ.

Scoring: In each iteration, we compute a combined-score for each product by considering the rating-scores or content-scores of the reviews it receives and the credibility values of its customers obtained in the previous iteration.

Feedback: There is a consensus: if there shall be good and bad products, the most highly scored products are good and the most lowly scored are bad. So after we get the product combined-scores, we will order them decreasingly. Then we select the top m and bottom m items in the list as the representative products for good and bad ones. The feedback has two main steps.

The first step is to decide feedback type: positive feedback or negative feedback. As we have declared in previous section, if the evaluation of a customer is consistent with most of others', we shall increase the credibility of the customer (positive feedback), or else, we shall reduce the credibility (negative feedback).

The second step is to decide feedback quantity. The rule is that if a customer's score is more consistent with the classification, the customer wins more feedback value.

Normalization. After each calculation, we perform the normalization as HITS does.

We can perform the iteration based on PC or CS sub-graph. Also, we can perform the iteration on PCS twin graph by running μ *and* ν interactively.

◇ **Module 3 Product Quality Re-Scoring:**

After we obtain the credibility vector, we compute the combined-scores for the products and shops as the final scores for ranking.

3. EXPERIMENTAL EVALUATION

Our experiment is conducted using the real data collected from Taobao. From the collected data, We extract the products and shops which have the returned customers as the test data.

Evaluation method: We sort the products (shops) in decreasing order by the scores and calculate the number of products (shops) N_k having returned customers among the TOP k%. The evaluation metric is: $t_k@TOPK = N_k/N_{all}$, where N_{all} is the total number of products (shops) having returned customers.

We build four systems including (1) Baseline: ranking products or shops with the average rating-scores. (2) SingleGraph: using the rating-scores in step **Scoring** based on PC/CS sub-graph. (3) TwinGraph: using the rating-scores in step **Scoring** based on PCS twin graph. (4) TwinGraph+ME: using the content-scores from ME in step **Scoring** based on PCS twin graph.

Table 1 and 2 shows the results. From the tables, we find that our proposed approach outperforms the Baseline and TwinGraph+ME achieves the best performance.

TOPK / System	10%	20%	40%	60%	80%	100%
Basline	10.01	17.43	35.75	51.47	65.77	100.00
SingleGraph	80.97	93.21	94.01	94.64	94.73	100.00
TwinGraph	78.03	94.30	95.10	95.64	95.91	100.00
TwinGraph+ME	77.42	94.48	95.51	95.83	97.17	100.00

Table 1: Results for products

TOPK / System	10%	20%	40%	60%	80%	100%
Basline	7.78	15.73	34.14	55.23	78.16	100.00
SingleGraph	39.67	65.61	78.16	87.20	93.98	100.00
TwinGraph	41.51	65.52	78.74	88.03	93.89	100.00
TwinGraph+ME	42.09	65.94	82.68	92.13	97.32	100.00

Table 2: Results for shops

4. CONCLUSION

As the online existing reviews are important for future potential users, it is critical to evaluate the reliability of the reviews. This paper proposes an effective technique to detect the customer credibility and use it to re-score the reviews for products and shops. Our experimental results show that the technique is promising.

5. ACKNOWLEDGMENTS

This work is partially supported by National Science Foundation of China under grant No.61103039 and No.60925008, National Basic Research Program of China under grant No. 2012CB316200, Joint Research Program with MSRA(Microsoft Research Asia).

6. REFERENCES

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

[2] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 1999.

[3] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *Proc. of ICDM*, pages 443–452, 2008.

[4] G. Wang, S. Xie, B. Liu, and P. S. Yu. Review graph based online store review spammer detection. In *Proc. of ICDM*, 2011.