Improving the Amazon Review System by Exploiting the Credibility and Time-Decay of Public Reviews*

Bo-Chun Wang, Wen-Yuan Zhu, and Ling-Jyh Chen
Institute of Information Science, Academia Sinica
{bcwang, juwyx, cclljj}@iis.sinica.edu.tw

Abstract

In this study, we investigate the review system of Amazon.com and propose a Review-credibility and Time-decay Based Ranking (RTBR) approach, which improves the Amazon review system by exploiting the credibility and time-decay of public reviews. Using a dataset downloaded from Amazon.com, we evaluate the proposed scheme on the current Amazon scheme. The results demonstrate that the RTBR scheme is superior to the Amazon scheme because it is more trustworthy and provides timely review results. Moreover, the scheme is simple and applicable to other Amazon-like review systems in which the reviews are time-stamped and can be evaluated by other users.

1. Introduction

Review systems have been implemented on a number of popular Web 2.0-based websites. Generally, a review system is a kind of reputation system that facilitates the development of trust in Internet interactions [9]. Unlike recommendation systems that seek to personalize each user's web experience by exploiting item-to-item and user-to-user correlations [6], review systems give an aggregated rating for an item based on other customers' opinions about the item.

Being one of the most successful online vendors in the world, `Amazon.com` allows users to submit their reviews to the web page of each product, and the reviews can be accessed by all users. Each review consists of the reviewer's name, several lines of comment, a rating score, and the timestamp. All reviews are archived, and the averaged result is reported on the web page of each product. It has been shown that such reviews provide basic ideas about the popularity and dependability of the corresponding items; hence they have a substantial impact on cybershoppers' behavior [2]. However, since it is an open forum, the anonymity of web reviewers increases the chances of abuse [3], and the review results may be misleading and trustworthy [5, 7]. To mitigate the problem, `Amazon` incorporates a feature that allows users to evaluate other users' product reviews by stating whether they think a review is useful or not; however, the *discriminating capability* of the `Amazon` review system is generally considered limited.

To improve the *discriminating capability* of the `Amazon` review system, we propose a *Review-credibility and Time-decay Based Ranking* (RTBR) approach. Specifically, RTBR enhances the `Amazon` scheme by exploiting the credibility and time-decay of public reviews. Using data downloaded from `Amazon.com`, we compare the proposed scheme with the current `Amazon` scheme, and show that it is more trustworthy and provides timely results in all test cases. Moreover, RTBR is simple and applicable to other `Amazon`-like rating systems, as long as each product's review is time-stamped and it can be evaluated by other users.

The remainder of this paper is organized as follows. In Section 2, we discuss related works on review systems. In Section 3, we present the proposed RTBR approach. In Section 4, we compare the proposed scheme with the current `Amazon` scheme and analyze the results. We then summarize our conclusions in Section 5.

2. Related Work

`Amazon` and `eBay`, two of the most successful Web 2.0 e-commerce stores, pioneered the use of review systems by aggregating user-contributed content. On the `eBay` website, buyers and sellers are allowed to post reviews about each other after a transaction has been completed. A review can be positive (1), neutral (0), or negative (-1). The system aggregates the reviews of each user by summing all of his/her received ratings, and details the results on the user's profile page. However, [5, 7] have suggested that the `eBay` review system is likely to mislead users because it lacks a discriminating capability, and [1] has observed that ballot stuffing is common and needs to be resolved.

*This research was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC 96-3113-H-001-012.

In contrast to eBay, the Amazon review system aggregates users' rating scores by averaging, instead of summing. It has been shown that the results of the Amazon review system are highly correlated to the prices of the corresponding products [2]. However, since it does not consider the aging issue of the reviews [11], and the review results are generally skewed toward high scores [2], the discriminating capability of the Amazon review system is limited.

In addition to summing and averaging approaches, several schemes have been proposed to improve the discriminating capability of review systems. For instance, [8] proposes a Bayesian-based review system that rates each product according to the feedback received. The Bayesian-based systems have been extended to filter out bad mouthing reviews [10] and support multiple value ratings [4]. However, these approaches are rarely implemented in reality because the computation and storage overheads are prohibitive.

3. The Proposed Approach: RTBR

We assume there are n items in the system, and the i-th item has been reviewed r_i times. Let N_i denote the i-th item, $s_{i,j}$ denote the j-th rating score of n_i, and $t_{i,j}$ denote the length of time since $s_{i,j}$ was rated. As each product review may also be reviewed by other users, we use $k_{i,j}$ to denote the number of users that have reviewed $s_{i,j}$, $u_{i,j}$ to denote the number of users (out of $k_{i,j}$) that think $s_{i,j}$ is useful, and λ to denote the aging factor ($0 < \lambda \leq 1$) that we discuss further in the next subsection. Then, for the j-th review of N_i, we define the review-credibility factor ($\omega_{i,j}$) and the time-decay factor ($\phi_{i,j}$) as follows:

$$\omega_{i,j} = \begin{cases} \frac{u_{i,j}}{k_{i,j}} & \text{, if } k_{i,j} \neq 0 \\ 0.5 & \text{, if } k_{i,j} = 0 \end{cases} \quad (1)$$

$$\phi_{i,j} = \lambda^{t_{i,j}}. \quad (2)$$

Finally, by combining the review-credibility factor, the time-decay factor, and the review scores, the aggregated score value for N_i is derived by $\mathcal{S}_i = \frac{\sum_{j=1}^{r_i} \omega_{i,j}\phi_{i,j}s_{i,j}}{\sum_{j=1}^{r_i} \omega_{i,j}\phi_{i,j}}$.

Suppose $\Delta(\mathcal{S}_i, \mathcal{S}_j)$ is a comparison function that returns 1 when $\mathcal{S}_i \geq \mathcal{S}_j$, and 0 otherwise. The proposed RTSB scheme then reports the ranking of N_i by taking the complementary cumulative distribution function (CCDF) of $\mathcal{S}_i$. As shown in Eq. 3, the ranking result indicates that N_i is in the top $\mathcal{R}_i^{RTBR}$ of all the compared products.

$$\mathcal{R}_i^{RTBR} = 1 - \frac{\sum_{j=1}^{n} \Delta(\mathcal{S}_i, \mathcal{S}_j)}{n} \quad (3)$$

3.1. The Aging Factor: λ

Here, we present the algorithm for selecting the aging factor λ in the RBTR scheme. Note that, as shown in Eq. 2, the smaller the value of λ, the more emphasis we put on the time-decay factor. Since each type of item has different sensitivity to the time-decay of reviews, the algorithm tries to determine the value of λ that will ensure the results of the RTBR scheme more representative and timely.

Algorithm 1 The algorithm for determining the value of the aging factor, λ, in the RBTR scheme.

```
Function Aging_Factor
i ⇐ −1; α1 ⇐ 1 − 10^i; δ1 ⇐ ϒ(α1)
while true do
    α2 ⇐ 1 − 10^(i−1); δ2 ⇐ ϒ(α2)
    if |δ1 − δ2| / δ1 ≤ 0.5 then
        return α1
    end if
    α1 ⇐ α2; δ1 ⇐ δ2; i ⇐ i − 1
end while
```

We denote the *ranking distance* of N_i by υ_i and $\upsilon_i = \mathcal{R}_i^{RTBR} - \mathcal{R}_i^{Amazon}$, where $\mathcal{R}_i^{Amazon}$ is derived in a similar manner to Eq. 3, except that $\Delta(\mathcal{S}_i, \mathcal{S}_j)$ is replaced by $\Delta(\overline{\mathcal{S}}_i, \overline{\mathcal{S}}_j)$ and $\overline{\mathcal{S}}_i = \frac{\sum_{j=1}^{r_i} s_{i,j}}{r_i}$. Suppose $\Upsilon(\alpha)$ is a comparison function that returns the average ranking distance of all the items when the value of λ is set to α in the RTBR scheme. Algorithm 1 shows the decision algorithm used to calculate the aging factor in the scheme.

4. Evaluation

We wrote a crawler program to download data from the bookstore department of Amazon.com at the end of April 2008. The downloaded data relates to books tagged as Programming, Animation, or Business. For each book, the collected data contains the book's title, the author's name, and the reviews received. Moreover, each downloaded review contains the rating score, the reviewer's name, the timestamp, the number of times the book has been evaluated, and the number of evaluations that deemed it useful. In this study, we only consider the books that have received more than five reviews, and the dataset contains 5,286 books and 309,140 reviews. Table 1 lists the properties of the dataset, and Figure 1 plots the mean score (for each book) distribution on a cumulative distribution function (CDF) curve. We find that about 70% of the books have a mean score higher than 4, and only 5% have a mean score lower than 3. The results confirm the previous findings [2] that the mean score distribution on the Amazon website is skewed towards higher scores.

We use Eq. 1 to calculate the credibility of each review, and plot the CDF distribution of the credibility scores and ages of the downloaded reviews in Figure 2 and 3 respectively. From Figure 2, we observe that there is a sharp increase (from 0.27 to 0.52) when the credibility value is 0.5.

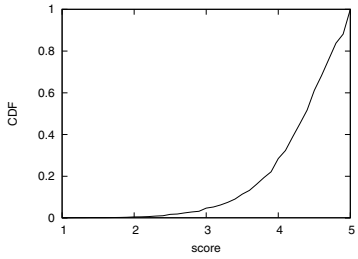


Figure 1. The CDF of the mean scores of the Amazon dataset.

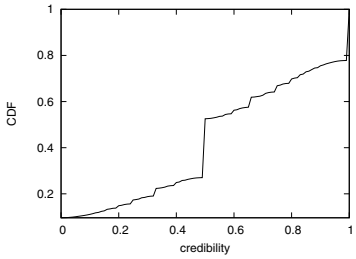


Figure 2. The CDF of the review credibility of the Amazon dataset.

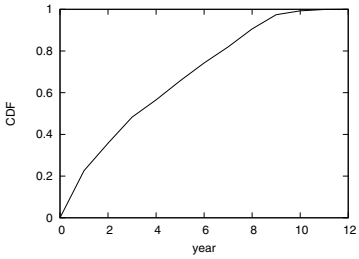
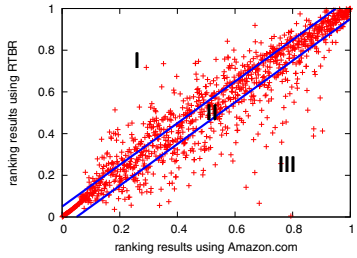
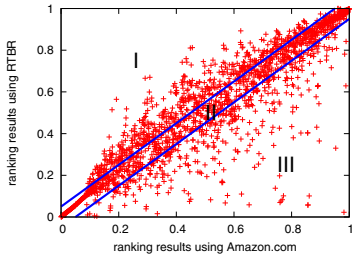


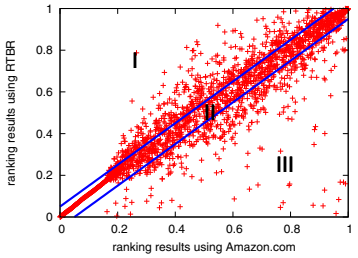
Figure 3. The CDF of the ages of reviews in the Amazon dataset.



(a) Programming



(b) Animation



(c) Business

Figure 4. Comparison of the review results derived by the proposed RTBR scheme and the Amazon scheme on the downloaded dataset. The sample points in areas I, II, and III are considered to be overestimated, consistent (within $\pm 5\%$ error), and underestimated respectively.

This is because the default credibility value is set to 0.5 if a review has not been evaluated by other users, as defined in Eq. 1. We also find that about 20% of the reviews are not trustworthy (i.e., the credibility value is less than 0.5), and only 25% of the reviews are highly credible (i.e., the credibility value is higher than 0.9). It seems that a substantial number of reviews on the Amazon website are either unreliable or malicious. Moreover, from Figure 3, we find that only 22% of the reviews were posted within the previous year, whereas more than 50% were posted at least four years earlier. The results confirm that the aging of reviews is a significant issue and must be carefully managed [11].

Next, we compare the review results of the proposed RTBR approach and those of the Amazon approach (i.e., by taking the mean of all the received rating scores). Using the dataset downloaded from the Amazon website, Figure 4 shows the comparison results, where each point represents a product with its corresponding ranking using the two schemes. Each sub-figure is divided into three areas: area I contains over-estimated products (i.e., the review results of the RTBR scheme are far lower than those of the Amazon scheme); area II contains consistently estimated products (i.e., the review results derived by the RTBR scheme and the Amazon scheme are within $\pm 5\%$ of each other); and area III contains under-estimated products (i.e., the review results of the RTBR scheme are far higher than those derived by the Amazon scheme). Table 2 summarizes the comparison results for the three categories of books, and the results show that only about 50% of the products have consistent review results in both schemes, while the others are either over-estimated or under-estimated. To investigate the causes of the inconsistent results, we design two tests, namely a *credibility test* and a *time-decay test*.

Credibility Test: Suppose that $\delta(s_{i,j}, x)$ is a comparison function that returns 1 if $s_{i,j} = x$, and 0 otherwise. For the i-th product N_i, we calculate the credibility factor $D_c(i, x)$ of each score value x using Eq. 4. Then, we apply linear regression to analyze the relationship between x and $D_c(i, x)$, and obtain the slope $L_c(i)$ of the regression line. Based on the value of $L_c(i)$, the *Credibility Test* reports *YES* (i.e., the inconsistency is caused by review credibility) if 1) $L_c(i) < 0$ and the corresponding point of N_i is in the Area I, or 2) $L_c(i) > 0$ and the corresponding point of N_i is in area III.

$$D_c(i,x) = \frac{\frac{\sum_{j=1}^{r_i} \omega_{i,j} s_{i,j} \delta(s_{i,j},x)}{\sum_{j=1}^{r_i} \omega_{i,j} \delta(s_{i,j},x)} - \sum_{j=1}^{r_i} s_{i,j} \delta(s_{i,j},x)}{\sum_{j=1}^{r_i} s_{i,j} \delta(s_{i,j},x)} \quad (4)$$

Tag	No. of products	Avg. no. of reviews
Programming	1159	31
Animation	1915	93
Business	2212	43

Table 1. The properties of the dataset downloaded from the Amazon.com website.

Tag	Area I	Area II	Area III
Programming	25.37%	53.58%	21.05%
Animation	27.73%	51.49%	20.78%
Business	22.51%	59.27%	18.22%

Table 2. The distribution of the comparison results for the downloaded Amazon dataset.

Subject	Area	Credibility	Time-decay	Union
Programming	I	81.13%	72.17%	96.23%
	III	99.47%	38.50%	99.47%
Animation	I	60.42%	67.55%	90.24%
	III	99.66%	44.56%	100.00%
Business	I	77.64%	80.90%	98.74%
	III	98.55%	32.37%	99.42%

Table 3. The evaluation results of the causes of under-estimations and over-estimations using the credibility test and time-decay test.

Time-decay Test: We denote $t_{i,max}$ and $t_{i,min}$ as the maximum and minimum values of $t_{i,j}$ for $1 \leq j \leq r_i$ respectively. We divide the period between $t_{i,min}$ and $t_{i,max}$ into Y equal intervals (Y is fixed at 10 in this study); and we assume that $\sigma(t_{i,j}, y)$ is equal to 1 when $t_{i,j}$ falls in the y-th interval, and 0 otherwise. For the i-th product N_i, we calculate its time-decay factor $D_t(i, y)$ for each time interval y using Eq. 5, where $\Delta(s_{i,j}, \overline{s}_i)$ is equal to 1 when $s_{i,j} > \overline{s}_i$, and 0 otherwise. Then, we apply linear regression to analyze the relationship between y and $D_t(i, y)$, and obtain the slope $L_t(i)$ of the regression line. Based on the value of $L_t(i)$, the *Time-decay Test* reports *YES* (i.e., the inconsistency is due to the time-decay of the reviews) if 1) $L_t(i) < 0$ and the corresponding point of N_i is in area I, or 2) $L_t(i) > 0$ and the corresponding point of N_i is in area III.

$$D_t(i, y) = \frac{\sum_{j=1}^{r_i} \sigma(t_{i,j}, y)\Delta(s_{i,j}, \overline{s}_i)}{\sum_{j=1}^{r_i} \sigma(t_{i,j}, y)} \qquad (5)$$

We examine the items that fall in area I and III using the two test approaches, and summarize the results in Table 3. We observe that most of the inconsistency is caused by the credibility of reviews. The credibility issue tends to cause more under-estimations, while the time-decay issue causes more over-estimations. Moreover, by combining the credibility and time-decay tests, we find that more than 95% of the inconsistency can be classified.

5. Conclusion

We have discussed the review system of `Amazon.com`. We argue that the results published by the `Amazon` review system are not representative because they do not consider the credibility and time-decay of public reviews. To address this issue, we propose the *Review-credibility and Time-decay Based Ranking* (RTBR) scheme. Using a dataset downloaded from `Amazon.com`, we compare the proposed scheme with the current `Amazon` scheme, and demonstrate that the proposed scheme is superior because it is more trustworthy and it provides timely review results. Moreover, the scheme is simple and applicable to other Web 2.0-based review systems in which the product reviews are time-stamped and they can be evaluated by other users.

References

[1] R. Bhattacharjee and A. Goel. Avoiding ballot stuffing in ebay-like reputation systems. In *ACM SIGCOMM Workshops*, 2005.

[2] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 48(3):345–354, Aug. 2006.

[3] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *ACM Electronic Commerce Conf.*, 2000.

[4] A. Josang and J. Haller. Dirichlet reputation systems. In *Int. Conf. on Availability and Security*, 2007.

[5] A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, Mar. 2007.

[6] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing*, 7(1):76–80, Jan./Feb. 2003.

[7] R. A. Malaga. Web-based reputation management systems: Problems and suggested solutions. *Electronic Commerce Research*, 1(4):403–417, Oct. 2001.

[8] L. Mui, M. Mohtashemi, C. Ang, and P. Szolovits. Ratings in distributed systems: A bayesian approach. In *Info. Tech. and Sys. Workshop*, 2001.

[9] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, Dec. 2000.

[10] A. Whitby, A. Josang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Int. Conf. on Autonomous Agenst Systems*, 2004.

[11] G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms for electronic marketplaces. *Decision Support Systems*, 29(4):371–388, Dec. 2000.