

Improving Amazon-like Review Systems by Considering the Credibility and Time-Decay of Public Reviews

Bo-Chun Wang

Department of Computer Science, University of Southern California, USA
bochunwa@usc.edu

Wen-Yuan Zhu

Department of Computer Science, National Chiao Tung University, Taiwan
wyzhu@cs.nctu.edu.tw

Ling-Jyh Chen

Institute of Information Science, Academia Sinica, Taiwan
ccljj@iis.sinica.edu.tw

Keywords: Review systems, credibility, time-decay, e-commerce

Received:

In this study, we investigate the review system of Amazon.com, which is regarded as one of the most successful e-commerce websites in the world. We believe that the review results provided by Amazon's review system may not be representative of the advertised products because the system does not consider two essential factors, namely the credibility and the time-decay of public reviews. Using a dataset downloaded from Amazon.com, we demonstrate that although the credibility and time-decay issues are very common, they are not handled well by current public review systems. To address the situation, we 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 reviews posted by the public. We evaluate the proposed scheme against the current Amazon scheme. The results demonstrate that the RTBR scheme is superior to the Amazon scheme because it is more credible and it 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

Amazon.com is regarded as one of the most successful online vendors in the world. In addition to spear-

heading online retail sales of a variety of products (such as books, music CDs, DVDs, software, and consumer electronics) and providing various useful web services, Amazon features review and recommendation systems that provide candid comments and recommendations for customers. Recent surveys have reported that 50% of online shoppers spend at least ten minutes reading reviews before making a decision about a purchase, and 26% of online shoppers read reviews on Amazon prior to making a purchase [1].

Review systems have been implemented on a number of popular Web 2.0-based e-commerce websites (e.g., Amazon¹ and eBay²), product comparison

⁰A preliminary version of this study was published in the Proceedings of International Workshop on Web Personalization, Reputation and Recommender Systems, 2008 [28]. In this extended version paper, we adopt the *Shallow Syntactic Features (ShallowSyn)* method [33] to estimate the credibility of public reviews that have not received a sufficient number of reviews. Moreover, we apply the *Kendall* test [16] to evaluate the correlation between the ranking results of our approach and the Amazon review system. We evaluate our proposed approach using a rich set of datasets, and discuss the deployment issue of the proposed approach in real systems. Hence, this manuscript is a much more thorough and authoritative presentation of our study on the Amazon review system.

¹Amazon. <http://www.amazon.com/>

²eBay. <http://www.ebay.com/>

websites (e.g., BizRate³ and Epinions⁴), and news websites (e.g., MSNBC⁵ and SlashDot⁶). Generally, a review system is a kind of reputation system that facilitates the development of trust in Internet interactions [24]. Unlike recommendation systems, which seek to personalize each user’s web experience by exploiting item-to-item and user-to-user correlations [20, 26], review systems give an average rating for an item based on other customers’ opinions about the item.

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 (either the real name or a nickname), several lines of comment, a rating score (ranging from one to five stars), and the timestamp of the review. All reviews are archived in the system, and the aggregated result, derived by averaging all the received ratings, 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 [6, 9].

However, since the Amazon review system is an open forum, the anonymity of web reviewers increases the chances of abuse, such as unfair/biased ratings, ballot stuffing, and bad mouthing [7]. As a result, the review results may be misleading and untrustworthy [15, 21]. 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 because 1) the review results have the tendency to be skewed toward high scores [6]; 2) the *aging* issue of the reviews is not considered [32]; and 3) it has no means to assess the reviews’ helpfulness if the reviews are not evaluated by a sufficiently large number of users (unless additional machine learning techniques could be applied [17, 33]).

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 system by exploiting the *credibility* and *time-decay* of public reviews. Using data downloaded from the

bookstore department of Amazon.com, we compare the proposed scheme with the current Amazon scheme, and show that it is superior because it is more credible and provides timely ranking results in all test cases. Moreover, the proposed scheme 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, using the real data downloaded from Amazon.com, we compare the proposed scheme with the current Amazon scheme and analyze the results. We also discuss the feasibility issue of the proposed scheme. We then summarize our conclusions in Section 5.

2 Related Work

A review system provides an average rating for each item based on other users’ opinions of the item; hence, it is a kind of reputation system that facilitates the development of trust in Internet interactions [4, 8, 24, 30]. Review systems are used by a number of Web 2.0 sites (such as Amazon, BizRate, eBay, Epinions, and SlashDot). Though the systems differ in how they aggregate users’ opinions and present the results, recent studies have shown that such systems have a strong impact on cybershoppers’ purchase decisions [1, 6, 9].

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. Resnick et al. [23, 25] evaluated eBay’s review system via controlled experiments and empirical analysis. In [23], they found more than half users were willing to provide feedback and it was almost all positive. In addition, [23] suggested that the users may reciprocate and retaliate. [25] found that eBay’s reputation system had significant effect in the market, and sellers who had high reputation scores would sell their goods with higher prices. [15, 21] suggested that the eBay review system is likely to mislead users be-

³BizRate. <http://www.bizrate.com/>

⁴Epinions. <http://www.epinions.com/>

⁵MSNBC News. <http://www.msnba.msn.com/>

⁶SlashDot. <http://slashdot.org/>

cause it lacks a discriminating capability (for instance, the eBay review system has difficulty distinguishing between a user who receives 50 positive reviews and a user who receives 100 positives and 50 negatives, as the aggregated ratings of the two users are equal to +50). Moreover, it has been observed that ballot stuffing is common in the eBay review system; hence, this issue also needs to be resolved [3].

In contrast to eBay, the Amazon review system aggregates users' rating scores by averaging, instead of summing. As mentioned in the previous section, the Amazon review system allows users to submit their reviews to the web page of each product. It has been shown that the results of the Amazon review system are highly correlated to the prices of the corresponding products [6], and about 25% of online shoppers read reviews on Amazon before they make a purchase [1]. However, the shortcomings of the system are that it does not consider the aging issue of the reviews [32], and the review results are generally skewed toward high scores [6]. In addition, [11] shows the average score of 53% of the products does not reveal the true quality of product and may provide misleading recommendations. As a result, the discriminating capability of the Amazon review system is limited.

In addition to summing and averaging approaches, a number of other schemes have been proposed to improve the discriminating capability of review systems [13, 14, 22, 27, 29, 31]. For instance, [14, 22] propose Bayesian-based review systems that rate each product according to the feedback received. Specifically, each item of feedback is given either a positive (+1) or a negative (-1) rating. The Bayesian-based systems have been extended to filter out bad mouthing reviews [31]. However, the disadvantage of the Bayesian-based system is that it can not provide ratings with graded levels because it is a binomial model. Therefore, [13] proposes a generalization of the Bayesian-based systems, called *Dirichlet reputation systems*, which can support multiple value ratings. Finally, [27, 29] propose personalizing review results based on the *Personalized Similarity Measure* and users' preferences. However, these approaches are rarely implemented in reality because the computation and storage overheads are prohibitive.

Since 'helpful' reviews have stronger impacts on consumers' purchase decisions than other reviews [5], several studies have investigated how to assess reviews' helpfulness recently [17, 33]. For instance,

[33] presents a *utility scoring* approach that computes three features of a given product review (namely the *Lexical Similarity Features*, *Shallow Syntactic Features*, and *Lexical Subjectivity Clues*) and then feeds the calculation results into a regression algorithm to measure the review's helpfulness. Similarly, [17] assesses review helpfulness using a SVM-based regression approach that considers five types of features, namely *Structural*, *Lexical*, *Syntactic*, *Semantic*, and *Meta-data*.

Finally, [12] focuses on the analysis and detection of review spam. In [12], Jindal and Liu use a supervised learning and classification model to detect three types of review spam: *False Opinions*, *Reviews on brands only*, and *Non-reviews*. In [10], Hu and Liu propose a data mining and natural language processing based approach to facilitate mining and summarizing product reviews from a large number of customer reviews of a particular product. Since it is difficult and tedious for consumers to read hundreds or thousands of reviews for each product, the feature-based summary results provide consumers more concise information for purchase decisions.

3 The Proposed Approach: RTBR

In this section, we present the proposed review system, called *Review-credibility and Time-decay Based Ranking* (RTBR), for emerging Web 2.0-based applications. Unlike the current Amazon review system, the proposed scheme is expected to better represent the public's opinions about reviewed items, because it considers two additional factors of each review in the system, namely, 1) the quality of being convincing and believable, i.e., the *review-credibility* factor; and 2) the timeliness of being representative, i.e., the *time-decay* factor⁷.

More precisely, we assume that 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. For the j -th review of N_i , we define the review-credibility factor as $\omega_{i,j}$ and the time-decay factor as $\phi_{i,j}$, which we will discuss further in

⁷A product may become popular (e.g., due to advertising, promotion, or marketing) or outdated (e.g., due to the release of a newer version) over time. Note that the time-decay factor should be weighted in accordance with the properties of the product types. For instance, it should be weighted higher for reviews of electronic products than that of books. We defer a detailed discussion and evaluation of this issue to a future work.

the following subsection.

Then, the proposed RTBR scheme calculates the score value of N_i (i.e., \mathcal{S}_i) by combining the review-credibility factor, the time-decay factor, and the review score of the received r_i reviews, as shown in Equation 1.

$$\mathcal{S}_i = \frac{\sum_{j=1}^{r_i} \omega_{i,j} \phi_{i,j} s_{i,j}}{\sum_{j=1}^{r_i} \omega_{i,j} \phi_{i,j}}. \quad (1)$$

Suppose $\Delta(\mathcal{S}_i, \mathcal{S}_j)$ is a comparison function that returns 1 when $\mathcal{S}_i \geq \mathcal{S}_j$, and it returns 0 otherwise. The RTSB scheme then reports the ranking of N_i by taking the complementary cumulative distribution function (CCDF) of \mathcal{S}_i . As shown in Equation 2, the ranking result indicates that N_i is in the top \mathcal{R}_i^{RTBR} of all the compared products.

$$\mathcal{R}_i^{RTBR} = 1 - \frac{\sum_{j=1}^n \Delta(\mathcal{S}_i, \mathcal{S}_j)}{n}. \quad (2)$$

Note that, in the Amazon review system, the score value of N_i is obtained by averaging the received rating scores of r_i reviews (i.e., $\bar{\mathcal{S}}_i$), as shown in Equation 3, and the ranking results are derived in a similar manner to Equation 2, except that $\Delta(\mathcal{S}_i, \mathcal{S}_j)$ is replaced by $\Delta(\bar{\mathcal{S}}_i, \bar{\mathcal{S}}_j)$, as shown in Equation 4.

$$\bar{\mathcal{S}}_i = \frac{\sum_{j=1}^{r_i} s_{i,j}}{r_i}. \quad (3)$$

$$\mathcal{R}_i^{Amazon} = 1 - \frac{\sum_{j=1}^n \Delta(\bar{\mathcal{S}}_i, \bar{\mathcal{S}}_j)}{n}. \quad (4)$$

3.1 The Review-Credibility Factor

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 the j -th review of N_i (i.e., $s_{i,j}$), and $u_{i,j}$ to denote the number of users (out of $k_{i,j}$) that think $s_{i,j}$ is useful. There are two cases for the definition of the review-credibility factor ($\omega_{i,j}$) for $s_{i,j}$, as shown in Equation 5:

- *Case 1:* If the j -th review of N_i has been reviewed by a sufficient number of people, i.e., $k_{i,j} \geq \gamma$, we define $\omega_{i,j}$ as the ratio of $u_{i,j}$ to $k_{i,j}$.
- *Case 2:* If $k_{i,j} < \gamma$, there may be a strong bias in the $k_{i,j}$ reviews when $k_{i,j}$ is small, or the value of $\frac{u_{i,j}}{k_{i,j}}$ cannot be calculated when $k_{i,j} = 0$. Thus, we define the value of $\omega_{i,j}$ using the *Shallow Syntactic Features* (ShallowSyn) method [33].

$$\omega_{i,j} = \begin{cases} \frac{u_{i,j}}{k_{i,j}} & , \text{ if } k_{i,j} \geq \gamma \\ \text{ShallowSyn}(\text{the } j\text{-th review of } N_i) & , \text{ if } k_{i,j} < \gamma \end{cases} \quad (5)$$

Specifically, the *ShallowSyn* method employs the Support Vector Machine (SVM) approach to estimate the review-credibility of a give product review. The training dataset contains a sufficiently large number of reviews that have been reviewed by at least γ users. Using the training dataset, a SVM model is built in the off-line phase by considering a variety of features of each review, including the number of words, the number of sentences, and the number of the words of *shallow syntactic features* (i.e., proper nouns, numbers, modal verbs, interjections, comparative and superlative adjectives, comparative and superlative adverbs, wh-determiners/adverbs/pronouns and possessive wh-pronouns). Then, the SVM model is used to estimate the review-credibility of the reviews that are reviewed by less than γ users.

3.2 The Time-Decay Factor

For the j -th review of N_i , we define the time-decay factor ($\phi_{i,j}$) by

$$\phi_{i,j} = \lambda^{t_{i,j}}, \quad (6)$$

where λ is an aging factor ($0 < \lambda \leq 1$). The value of λ is calculated using the decision algorithm, as shown in Algorithm 1, where $\Upsilon(\alpha)$ is the 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, i.e., the mean of $|\mathcal{R}_i^{RTBR} - \mathcal{R}_i^{Amazon}|$ for all i . Note that, as shown in Equation 6, the smaller the value of λ , the more emphasis we put on the time-decay factor of public reviews. 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 proposed RTBR scheme are more representative and timely.

4 Evaluation

In this section, we evaluate the proposed RTBR scheme and compare it with the current Amazon review system. We present the properties of the dataset downloaded from the bookstore department of Amazon.com in subsection 4.1, and show the evaluation results in subsection 4.2. Moreover, we discuss

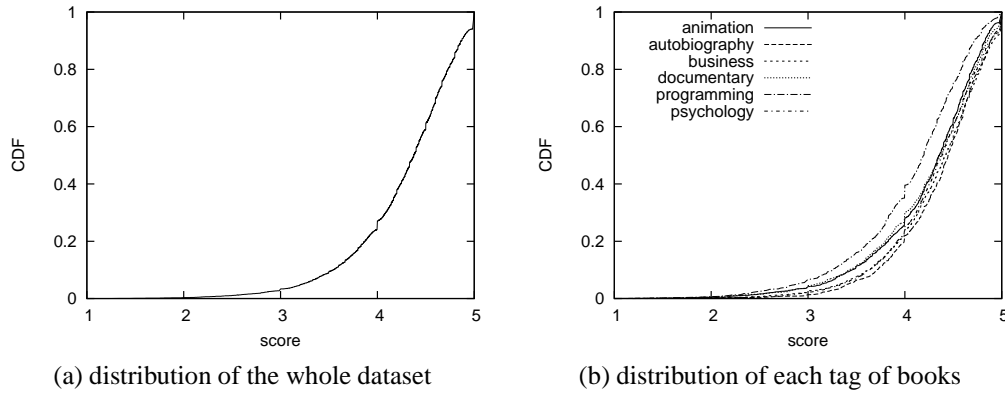


Figure 1: The CDF distribution of the mean scores of the downloaded Amazon dataset.

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

```

1: Function Aging_Factor
2:  $i \leftarrow -1$ ;  $\alpha_1 \leftarrow 1 - 10^i$ ;  $\delta_1 \leftarrow \Upsilon(\alpha_1)$ 
3: while true do
4:    $\alpha_2 \leftarrow 1 - 10^{i-1}$ ;  $\delta_2 \leftarrow \Upsilon(\alpha_2)$ 
5:   if  $\frac{|\delta_1 - \delta_2|}{\delta_1} \leq 0.1$  then
6:     return  $\alpha_1$ 
7:   end if
8:    $\alpha_1 \leftarrow \alpha_2$ ;  $\delta_1 \leftarrow \delta_2$ ;  $i \leftarrow i - 1$ 
9: end while

```

the feasibility and implementation issues of the proposed scheme in subsection 4.3.

4.1 Data Collection and Analysis

We wrote a crawler program to download data from the bookstore department of Amazon.com at the end of June 2011⁸. The downloaded data relates to books tagged as Animation, Autobiography, Business, Documentary, Programming, or Psychology. For each book, the collected data contains the book’s title, the author’s name, and the reviews received. Moreover, each 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. For simplicity, in this study, we only consider the books that

⁸We note that the proposed approach is also applicable to the other product reviews on Amazon.com. However, we do not include the evaluation using the other products on Amazon.com in this study because the number of the items differs greatly among different product categories, and the number of the reviews per item varies a lot even within the same product type. We defer a detailed discussion of this issue to a future work.

Table 1: The properties of the dataset downloaded from the bookstore department of Amazon.com

Tag	No. of products	Avg. no. of reviews
Animation	4,511	86
Autobiography	2,410	92
Business	6,780	52
Documentary	3,331	52
Programming	2,355	34
Psychology	5,780	63

have received more than five reviews. The dataset contains 25,167 books and 1,644,871 reviews. Table 1 lists the properties of the dataset.

Like the Amazon review system, we first calculate the mean of the rating scores for each book in the dataset. Then, we plot the mean score distribution on cumulative distribution function (CDF) curves, as shown in Figure 1. We find that 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 findings of previous studies that the mean score distribution on the Amazon website is skewed towards higher scores [6, 28]. Thus, the current Amazon review system cannot be considered representative because it lacks a discriminating capability.

Next, following [33], we set the value of γ to 10, and used Equation 5 to calculate the credibility of each review. We used the *Stanford Parser* [18, 19] to parse every review and compute its features. Then, for each category of books, we collected the features from the reviews that have been reviewed by at least 10 users as the training dataset, and applied the ϵ -Support Vector Regression (ϵ -SVR) implemented in LIBSVM [2] to build the SVM models⁹, which were used to esti-

⁹To be accurate, the SVM models should be updated periodically, for each category of books, after real-world deployment.

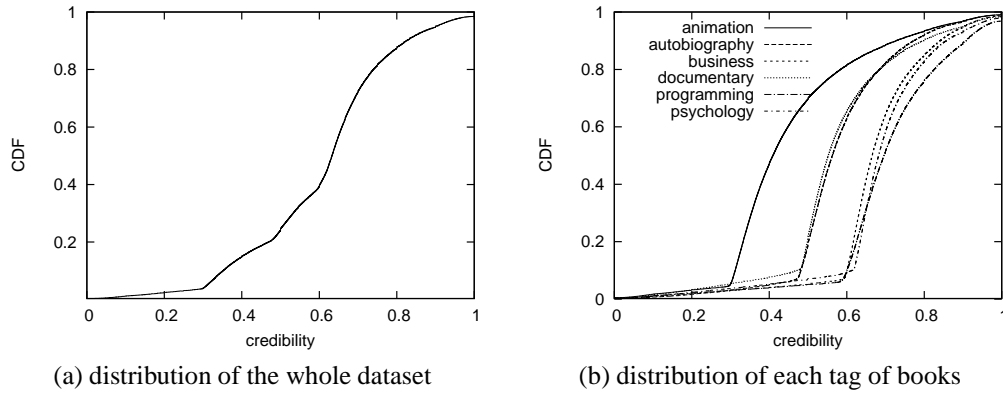


Figure 2: The CDF distribution of the review credibility of the downloaded Amazon dataset.

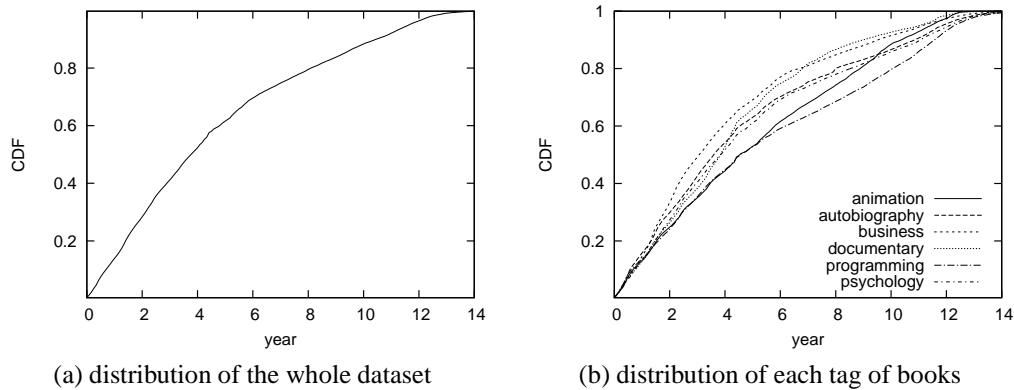


Figure 3: The CDF distribution of the ages of reviews in the downloaded Amazon dataset.

Table 2: No. of training data and mean squared error

Tag	No. of training data	σ^2
Animation	53,850	0.0899
Autobiography	44,799	0.0618
Business	84,063	0.0708
Documentary	45,424	0.0775
Programming	17,156	0.0734
Psychology	95,050	0.0741

mate the credibility values of the reviews which are reviewed by less than 10 users. Table 2 lists the properties of the training dataset and the regression performance (in terms of the mean squared error σ^2). All results are based on 10-fold cross validation.

Figure 2 shows the CDF distribution of the credibility scores of the downloaded reviews. We observe that about 24% of the reviews are not credible (i.e., the credibility value is less than 0.5), and only 12% of the reviews are highly credible (i.e., the credibility value is higher than 0.8). It seems that a substantial number of reviews on the Amazon website are either unreliable (e.g., due to individual preferences or unintentional bias) or malicious (e.g., due to ballot stuffing,

bad mouthing, or intentional bias). Again, the results confirm the findings of previous studies as reported in [7, 15, 21, 28].

Moreover, in Figure 3, we plot the CDF distribution of the ages of the 1,644,871 downloaded reviews (i.e., the time since each review was posted until the data was collected). Interestingly, only 13% 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 [28, 32]. For instance, over time, a book may become popular (i.e., due to advertising, promotion, or marketing) or outdated (i.e., due to the release of a newer version of the product). Hence, the aging factor must be carefully managed in order to improve the discriminating capability of the review system.

4.2 Evaluation of the RTBR Scheme and the Current Amazon Review System

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 Kendall test [16]. Table 3 shows the

Table 4: The example of two programming books on Amazon.com that are over-estimated and under-estimated respectively

Book A				Book B			
Score	Data	Review Title	Ratio of people felt helpful	Score	Data	Review Title	Ratio of people felt helpful
5	2007/09/02	“Best book on QT 4”	12 of 16	4	2003/01/17	“Good OOP Book”	8 of 11
5	2007/10/14	“Arrived in good order”	0 of 34	5	2003/03/25	“Extremely well written and ENJOYABLE Book”	4 of 4
5	2007/12/04	“It’s an excellent guide for any QT programmer”	6 of 7	5	2003/05/17	“Which C++ Book To Read First?”	39 of 40
3	2008/07/02	“A Mixed Bag”	14 of 14	5	2003/09/16	“C++ enthusiast”	9 of 9
3	2009/03/21	“UI files are incompatible with Qt 4.5”	3 of 3	5	2003/12/13	“Pure C++ Tutorial”	6 of 6
4	2009/04/02	“Very Good Book”	1 of 2	4	2004/02/15	“GOOD BOOK, BUT...”	1 of 27
4	2009/10/15	“Pretty Good”	1 of 1	5	2008/01/11	“Good Start Point for Professionals”	2 of 2
4	2009/11/26	“Full of useful informaton”	0 of 0	5	2008/07/09	“Well written, good examples”	0 of 0
Ranking by Amazon.com: 50%				Ranking by Amazon.com: 7%			
Ranking by RTBR: 68%				Ranking by RTBR: 2%			

Table 3: Kendall’s rank correlation (τ)

Tag	Kendall’s rank correlation (τ)
Animation	0.8507614
Autobiography	0.8617765
Business	0.8876215
Documentary	0.8533894
Programming	0.8666127
Psychology	0.8351172

Kendall’s rank correlation (τ) results. The Kendall’s rank correlation (τ) is effective in evaluating the degree of similarity between two rankings given to the same set of objects. From the evaluation results, we observe that the ranking results of the two approaches have a high correspondence, which is encouraging since the goal of the RTBR approach is to adjust the Amazon approach by considering the credibility and time-decay factor, it should not change the order of ranking results of the Amazon approach drastically.

Then, using the dataset downloaded from the Amazon website, Figure 4 shows the comparison results, where each point represents a product with its corresponding ranking (as a percentage) using the RTBR scheme and the Amazon scheme. Each sub-figure is divided into three areas: 1) area I contains *over-estimated* products (i.e., the review results of the RTBR scheme are far lower than those of the Amazon scheme); 2) area II contains consistently estimated products (i.e., the review results derived by the RTBR scheme and the ordinary Amazon scheme are within $\pm 5\%$ of each other); and 3) 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 4 shows an example of two programming books¹⁰ that are over-estimated and under-estimated respectively.

The results in Table 4 show that the RTBR scheme can improve the Amazon approach because it considers the credibility and time-decay factors. Specifically, in Table 4, *Book A* is considered over-estimated under the Amazon approach because most of the high-score reviews are either outdated (e.g., all 5-star reviews were made in 2007) or not creditable (e.g., one of the 5-star reviews were regarded not helpful by 34 reviewers). In contrast, *Book B* is regarded under-estimated under the Amazon approach, as its 5-star reviews are more creditable than 4-star reviews (e.g., one of the 4-star reviews was accepted by one of 27 reviewers).

We summarize the distribution of the comparison results for the six categories of books in Table 5, and the results show that only less than a half of the products have consistent review results in both schemes, while the others are dominated by over-estimation and next by under-estimation. Moreover, the results are consistent with our previous findings, which are based on the dataset of the same categories collected in April 2008. To investigate the causes of the inconsistent results, we design two tests, namely a *credibility test* and a *time-decay test*.

1. Credibility Test

¹⁰Book A: *The Book of Qt 4: The Art of Building Qt Applications*, ISBN: 1593271476, <http://www.amazon.com/product-reviews/1593271476>; Book B: *Object-Oriented Programming in C++*, ISBN: 0470843993, <http://www.amazon.com/product-reviews/0470843993>. (Accessed on September 1, 2011)

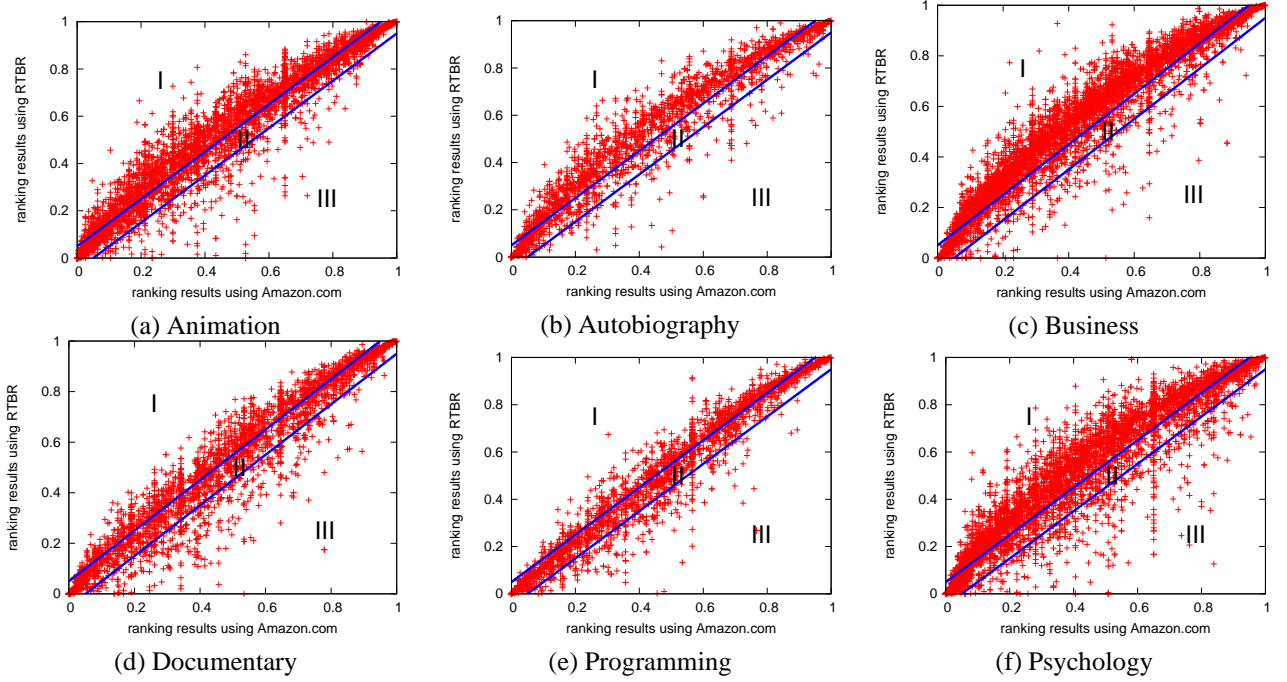


Figure 4: Comparison of the review results derived by the proposed RTBR scheme and the Amazon scheme on the downloaded dataset. The sample points in areas I, II, and III are considered to be overestimated, consistent (within $\pm 5\%$ error), and underestimated respectively.

Table 5: The distribution of the comparison results for the downloaded Amazon dataset.

Tag	Area I	Area II	Area III
Animation	50.14%	43.92%	5.94%
Autobiography	57.93%	37.26%	4.81%
Business	63.60%	33.21%	3.19%
Documentary	37.98%	49.53%	12.49%
Programming	33.89%	55.41%	10.70%
Psychology	58.30%	34.59%	7.11%

This test is designed to determine whether the inconsistency between the RTBR and Amazon schemes is caused by the credibility of reviews. Suppose that $\delta(s_{i,j}, x)$ is a comparison function that returns 1 if $s_{i,j} = x$, and 0 otherwise, as shown in Equation 7. For the i -th product N_i , we calculate the credibility factor $D_c(i, x)$ of each score value x using Equation 8. 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 *TRUE* (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; and it reports

FALSE otherwise.

$$\delta(s_{i,j}, x) = \begin{cases} 1 & , \text{ if } s_{i,j} = x \\ 0 & , \text{ if } s_{i,j} \neq x \end{cases} \quad (7)$$

$$D_c(i, x) = \frac{\sum_{j=1}^{r_i} \omega_{i,j} \delta(s_{i,j}, x)}{\sum_{j=1}^{r_i} \omega_{i,j}} = \frac{\sum_{j=1}^{r_i} \delta(s_{i,j}, x)}{r_i} \quad (8)$$

2. Time-decay Test

This test attempts to determine whether the inconsistency between the RTBR and Amazon schemes is caused by the time-decay of the reviews. 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 (for simplicity, Y is fixed at 10 in this study), and assume that $\sigma(t_{i,j}, y)$ is equal to 1 when $t_{i,j}$ falls in the y -th interval, as shown in Equation 9. For the i -th product N_i , we calculate its time-decay factor $D_t(i, y)$ for each time interval y using Equation 10. More specifically, in Equation 10, $\Delta(s_{i,j}, \bar{s}_i)$ is equal to 1 when $s_{i,j} > \bar{s}_i$, 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

Table 6: The evaluation results of the causes of under-estimations and over-estimations using the designed credibility test and time-decay test.

Subject	Area	Credibility	Time-decay	Union
Animation	I	63.31%	56.85%	91.15%
	III	98.88%	66.67%	99.63%
Autobiography	I	36.97%	67.91%	82.34%
	III	96.55%	65.52%	100.00%
Business	I	41.95%	70.67%	86.10%
	III	97.69%	67.59%	100.00%
Documentary	I	51.11%	68.81%	91.19%
	III	99.04%	52.16%	100.00%
Programming	I	58.34%	76.91%	94.35%
	III	98.41%	55.16%	100.00%
Psychology	I	30.77%	72.31%	83.35%
	III	98.78%	63.50%	100.00%

on the value of $L_t(i)$, the *Time-decay Test* reports *TRUE* (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; and it reports *FALSE* otherwise.

$$\sigma(t_{i,j}, y) = \begin{cases} 1 & , \text{if } t_{i,\min} + \frac{(y-1) \times (t_{i,\max} - t_{i,\min})}{Y} \\ & \leq t_{i,j} < t_{i,\min} + \frac{y \times (t_{i,\max} - t_{i,\min})}{Y} \\ 0 & , \text{otherwise} \end{cases} \quad (9)$$

$$D_t(i, y) = \frac{\sum_{j=1}^{r_i} \sigma(t_{i,j}, y) \Delta(s_{i,j}, \bar{s}_i)}{\sum_{j=1}^{r_i} \sigma(t_{i,j}, y)} \quad (10)$$

We examine the items that fall in area I and III using the two test approaches, and summarize the results (i.e., whether they are caused by the credibility or time-decay factors, or a combination of the two) in Table 6. From the results, we observe that most of the inconsistency is caused by the credibility of reviews. Moreover, we observe that the credibility issue tends to cause more *under-estimations*, while the time-decay issue causes more *over-estimations*. We also find that, by combining the credibility and time-decay tests, more than 82% of the inconsistency can be classified. Since the RTBR approach considers the credibility and time-decay issues, it is superior to the Amazon approach because it provides more representative review results.

4.3 Discussion

In this subsection, we discuss the implementation issue of the proposed approach, and we demonstrate

that the score values can be updated in an *incremental* manner in the proposed approach, thereby reducing greatly the computational complexity in real systems.

More specifically, we let N_i denote the i -th item, $s_{i,j}$ denote the j -th rating score of N_i , r_i denote the number of users that have reviewed N_i , and S_i denote the score of N_i at time T . In addition, for the j -th review of N_i , $\omega_{i,j}$ denotes the review-credibility factor, and $\phi_{i,j}$ denotes the time-decay factor. The system has to calculate the numerator ($\mathcal{A}_i = \sum_{j=1}^{r_i} \omega_{i,j} \phi_{i,j} s_{i,j}$) and the denominator ($\mathcal{B}_i = \sum_{j=1}^{r_i} \omega_{i,j} \phi_{i,j}$) respectively in order to obtain the value of S_i (cf. Equation 1), and there are two cases to update the value of S_i :

1. Case 1: a new review for N_i is input at time T'

In this case, we first obtain the review-credibility factor, ω_{i,r_i+1} , and the time-decay factor, ϕ_{i,r_i+1} , using Equations 5 and 6 respectively. Then, the system will update the values of \mathcal{A}_i and \mathcal{B}_i using Equations 11 and 12 (i.e., consider the time decay of the previous values of \mathcal{A}_i and \mathcal{B}_i by multiplying $\lambda^{T'-T}$, and plus the input of the new r_i+1 th review), and derive the updated score value S' using Equation 13.

$$\mathcal{A}'_i = \mathcal{A}_i \lambda^{T'-T} + \omega_{i,r_i+1} \phi_{i,r_i+1} s_{i,r_i+1} \quad (11)$$

$$\mathcal{B}'_i = \mathcal{B}_i \lambda^{T'-T} + \omega_{i,r_i+1} \phi_{i,r_i+1} \quad (12)$$

$$S'_i = \frac{\mathcal{A}'_i}{\mathcal{B}'_i} \quad (13)$$

2. Case 2: the j -th review of N_i is changed at time T'

In this case, we obtain the the new review-credibility factor of the j -th review, ω'_{i,r_j} by Equation 5, and update the values of \mathcal{A}_i and \mathcal{B}_i using Equations 14 and 15 respectively (i.e., plus the offset caused by the update of the j -th review). Then, we obtain the updated score value S'_i using Equation 13.

$$\mathcal{A}'_i = \mathcal{A}_i + (\omega'_{i,j} - \omega_{i,j}) \phi_{i,j} s_{i,j} \quad (14)$$

$$\mathcal{B}'_i = \mathcal{B}_i + (\omega'_{i,j} - \omega_{i,j}) \phi_{i,j} \quad (15)$$

As we can see in the above two cases, the score values of each item can be updated in an incremental manner. Hence, the computational complexity of the proposed approach is moderate, and the proposed approach is feasible to be implemented in real systems.

5 Conclusion

In this paper, we have discussed the review system of Amazon.com, one of the most popular online vendors in the world. We argue that the results published by the Amazon review system are not representative because they do not consider two essential factors, namely the credibility and time-decay of reviews submitted by the public. To address this issue, we propose the *Review-credibility and Time-decay Based Ranking (RTBR)* scheme. Using a dataset downloaded from the bookstore department of Amazon.com, we compare the proposed scheme with the current Amazon scheme, and demonstrate that the proposed scheme is superior because it is more credible and it provides timely review results. Moreover, we demonstrate that the proposed scheme can update its parameters in an incremental manner, and thus reduce greatly the computational complexity in real world implementation. The scheme is simple, effective, 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.

Acknowledgement

We are grateful to the editors and anonymous reviewers for their insightful comments. This material is based upon work supported by the Taiwan E-learning and Digital Archives Program (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC 99-2631-H-001-020, NSC 100-2631-H-001-013, and NSC 100-2631-S-003-006.

References

- [1] IT Facts. <http://www.itfacts.biz/>.
- [2] LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [3] R. Bhattacharjee and A. Goel. Avoiding ballot stuffing in ebay-like reputation systems. In *ACM SIGCOMM workshop on Economics of peer-to-peer systems*, 2005.
- [4] V. Carchiolo, A. Longheu, M. Malgeri, and G. Mangioni. Context-based Global Expertise in Recommendation Systems. *Informatica*, 34(4):409–417, 2010.
- [5] P.-Y. Chen, S. Dhanasobhon, and M. D. Smith. All reviews are not created equal: The disparate impact of reviews and reviewers at amazon.com. *SSRN eLibrary*, 2008.
- [6] 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, August 2006.
- [7] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *ACM Electronic Commerce Conference*, 2000.
- [8] P. Dondio and S. Barrett. Computational Trust in Web Content Quality: A Comparative Evaluation on the Wikipedia Project. *Informatica*, 31(2):151–160, 2007.
- [9] D. Houser and J. Wooders. Reputation in auctions: Theory, and evidence from ebay. *Journal of Economics & Management Strategy*, 15(2):353–369, June 2006.
- [10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *ACM SIGKDD*, 2004.
- [11] N. Hu, P. A. Pavlou, and J. Zhang. Can online reviews reveal a product’s true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *ACM Electronic Commerce Conference*, 2006.
- [12] N. Jindal and B. Liu. Analyzing and detecting review spam. In *IEEE International Conference on Data Mining*, 2007.
- [13] A. Josang and J. Haller. Dirichlet reputation systems. In *International Conference on Availability, Reliability and Security*, 2007.
- [14] A. Josang and R. Ismail. The beta reputation system. In *15th Bled Conference on Electronic Commerce*, 2002.
- [15] 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, March 2007.
- [16] M. Kendall and J. D. Gibbons. *Rank Correlation Methods*. Charles Griffin & Company Limited, 5 edition, September 1990.

- [17] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *International Conference on Empirical Methods in Natural Language Processing*, 2006.
- [18] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *The Annual Meeting on Association for Computational Linguistics*, 2003.
- [19] D. Klein and C. D. Manning. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, 2003.
- [20] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing*, 7(1):76–80, January/February 2003.
- [21] R. A. Malaga. Web-based reputation management systems: Problems and suggested solutions. *Electronic Commerce Research*, 1(4):403–417, October 2001.
- [22] L. Mui, M. Mohtashemi, C. Ang, and P. Szolovits. Ratings in distributed systems: A bayesian approach. In *Workshop on Information Technologies and Systems*, 2001.
- [23] P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *Advances in Applied Microeconomics*, 11:127–157, 2002.
- [24] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.
- [25] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101, June 2006.
- [26] J. B. Schafer, J. Konstan, and J. Riedi. Recommender systems in e-commerce. In *ACM Electronic Commerce Conference*, 1999.
- [27] M. Srivatsa, L. Xiong, and L. Liu. Trustguard: countering vulnerabilities in reputation management for decentralized overlay networks. In *International World Wide Web Conference*, 2005.
- [28] B.-C. Wang, W.-Y. Zhu, and L.-J. Chen. Improving the amazon review system by exploiting the credibility and time-decay of public reviews. In *International Workshop on Web Personalization, Reputation and Recommender Systems*, 2008.
- [29] Y. Wang and J. Vassileva. Trust and reputation model in peer-to-peer networks. In *International Conference on Peer-to-Peer Computing*, 2003.
- [30] Y. Weng, C. Hu, X. Zhang, and L. Zhao. BREM: A Distributed Blogger Reputation Evaluation Model Based on Opinion Analysis. *Informatica*, 34(4):419–328, 2010.
- [31] A. Whitby, A. Josang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *The International Joint Conference on Autonomous Agent Systems*, 2004.
- [32] G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms for electronic marketplaces. *Decision Support Systems*, 29(4):371–388, December 2000.
- [33] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *ACM International Conference on Information and Knowledge Management*, 2006.