Promote Product Reviews of High Quality on E-commerce Sites

Shen Huang
eBay Research Labs
Shanghai, China
shenhuang@ebay.com

Dan Shen
eBay Research Labs
dashen@ebay.com

Wei Feng
Department of Computer Science
University of Toronto, Canada
weifeng@cs.toronto.edu

Catherine Baudin
eBay Research Labs
San Jose, CA, USA
cabaudin@ebay.com

Yongzheng Zhang
eBay Research Labs
San Jose, CA, USA
ytzhang@ebay.com

Abstract

With the community of online reviewers growing rapidly, we find it increasingly difficult to digest all the information within a limited time. Users' requirements raise an interesting problem not well studied yet: how to discover the high quality product reviews? We believe a good solution will provide at least two types of benefit: 1) Rank reviews in terms of their quality. This could improve user experience by enabling them to learn more with a few detailed high-quality reviews instead of review outlines of irrelevant content and spam. 2) Automatically summarize user opinions. Researchers have studied this problem for years and are trying to assist users in getting the main products information concepts more efficiently. With this respect, low-quality content will definitely degrade the accuracy performance of any algorithm on this task. For the purpose of quality prediction, previous research thoroughly examined various properties of product reviews based on their content. Although some promising results have been obtained, we believe there is still room for improvement. Overall, we explore the topic of review quality from two aspects: 1) to filter out noisy data. Here we leverage classification techniques to differentiate real product reviews from other types of reviews and spam. Indeed many articles that fall under the label "product reviews" really belong to three groups: product reviews, feedback for retailers, and commercial spam. The empirical results show that this research could be put into practice with sufficient training data. 2) To assess the quality of a review we also take into consideration another information resource: the behavior of a review author in an e-commerce community. Our requirement is that after the noise filtering step, all product reviews must be ranked according to their quality. The common methods for this type of task are usually based solely on the *analysis of the text of the review. By contrast, we performed a high-level analysis on two kinds of data: product reviews and deal transactions. An interesting finding reveals that review quality is not only related to their content, but can also be derived from the behavior of the review author. Therefore, in order to inspect review quality from the perspectives of human credibility and expertise, we consider the following three features: the author personal reputation, the "seller degree" that reflects if the author is also a seller, and the "expertise degree". Our experiments show that the addition of these features increase the performance of the review quality ranking. Furthermore, we propose an evolving model given the above observations. The model is able to generate the basic characteristics of the review community, especially when the above three features are taken into consideration. In addition, the model could help us make more reasonable predictions for concerning the evolution of the review community.*

Keywords: E-commerce, Product review quality, User behavior modeling

Introduction

In the Web 2.0 era, the e-commerce industry is becoming is increasingly energized because of the content contributed by Web users. Statistical marketing reports from 2006 state that at least 70% of online shoppers read reviews[1] and that 60% of shoppers are more likely to purchase from a site containing product ratings and reviews[2]. Meanwhile, retailers are more willing to advertise on these sites. In such business environments, research on opinion mining for product reviews has drawn considerable attention (Pang et al., 2002; Barabasi and Albert, 1999; Pang and Lee, 2005; Hu and Liu, 2004; Mishne and Glance, 2006). The major topics in this area include product features identification, user sentiment classification and opinion summarization, all technologies that are geared toward helping users estimate products reputation more efficiently.

Unfortunately, people sometime leave unrelated or even spammed content on internet. Reviews of products are not exceptions. Let's take eBay Reviews & Guides as a sample (http://reviews.ebay.com). EBay is a client-to-client e-commerce site and customers often get confused between the product review portal and seller/buyer feedback profiles. In fact, feedback for sellers is often found among product reviews. For instance, "Camera and service were as described and seller treated us fine but we just paid too much...shipping is fast…". Even worse, some commercial spam can also be found among product reviews, for example, "This is a great gadget but very expensive, I'm glad i found www.ultimate-free-gifts.co.uk where i got it for free…". If we attempt to organize information using opinion mining techniques on such a noisy dataset, we believe that no approach achieve any high level of accuracy. More important, high quality information will become much more precious to users if they have to dig through huge amount of articles. Therefore product review quality can be seen as another contributing factor if we go beyond the scope of opinion summarization or the analysis of brand reputation influence.

Initial study confirms the necessity of this research. The first branch in this area takes care of the noise issue. Jindal and Liu (2007) showed the existence of duplicated and spammed reviews by reporting some statistics in several product domains. They extracted various features, such as rating and title length, to detect low quality articles. We follow their idea of detecting duplicated content. Besides that, we also try to detect the feedback for seller on eBay site in this paper.

Another branch of the quality study addresses the ranking issue. Here removing unrelated content is not sufficient. Systems should tell users which articles contain valuable information, from the product review angle. Kim (Kim et al, 2006) and Zhang (Zhang and Varadarajan, 2006) tackled the quality issue by leveraging the review "helpfulness" voting from readers. They viewed the task as a ranking problem and tested regression models on it. Liu et al. (2007) initiated a series of experiments by incorporating content information with some other review features. After observing extensive e-commerce data, we found that substantial clues can also be acquired from the behaviors of the authors behind the documents. For example, a person who has used many types of digital cameras will have more insights on the performance of different cameras. Thus, his opinions will be more reliable and valuable for readers to make purchase decisions. We attempt to infer such information from review author's behavior on e-commerce sites. To the best of our knowledge, it is the first attempt of systematic study in this direction. We expect that this exploration can help improve the performance of review quality assessment.

Based on above observations, we think the research on review quality could be improved from at least the following two aspects: 1) Discover more obvious noisy data like feedback for retailer and commercial spam. 2). Assess product review quality from more aspects with the involvement of user online activity. For the first aspect, a new kind of noise: feedback for seller is introduced. We follow the normal practice and adopt a classification technique – Support Vector Machine (SVM). The three groups for classification include product review, feedback for seller and commercial spam. We conducted 5-folder cross validation on the test dataset and the empirical results is really convincible.

For the second aspect, we define two concepts from user behavior: credibility and expertise. Credibility is motivated by the observation that only reliable information can be trusted by readers, while expertise is based on the common practice that senior users with more domain knowledge will contribute more valuable information to a community, especially the review sites. In another word, we assume that a review has a satisfactory quality only if it is full of real and professional opinions. In the scope of the two aspects, we design three features including personal reputation, seller degree and expertise degree, and attempt to extract them from user behavior accordingly. A reputation system is indispensible for any e-commerce site. Personal reputation of one user could be represented by using a feedback score rated by others on such sites. Seller degree means to indicate the role that the user performs in transactions, i.e. more like a seller or a buyer. The reason to involve this feature is that we wonder if sellers have any advertising behaviors. It becomes necessary to assess the credibility criterion from another perspective. Finally, expertise degree is devised to measure how good domain knowledge one user has on a specific product. After proposing above features, we need a series test to investigate the significance of each feature for estimating review quality. Now a new question comes into our mind: Is there any objective evaluation metric exist for this task? One of the reasonable answers, we suppose, is human wisdom. Most review sites provide a feedback system for readers to vote for an article with “helpful” or “not helpful”. Similar to the work introduced in (Kim et al., 2006; Zhang and Varadarajan, 2006), we leverage the information of “helpful” votes. The correlation between each feature and the portion of “helpful” votes is examined with Spearman rank coefficients and linear regression analysis. Experimental results show that the three features have strong relationship with review quality when we employ “helpful” voting as a kind of quality metric. Moreover, we prove the effectiveness of those features by integrating them into previous research efforts. Finally, we model user behaviors on e-commerce sites and their corresponding influences on product reviews by adapting a Web graph modeling method (Kleinberg et al, 1999; Kumar et al, 2000; Aiello et al., 2001; Laura et al., 2002; Eppstein and Wang, 2002). The model further proves the necessity of involving user behaviors for better simulation results.

Overall this paper makes three main contributions:

1. It refines noisy filtering task in product review by introducing feedback for seller.
2. It frames the problem of assessing review quality by considering behaviors of review authors which hasn’t been well studied so far. Three features are devised in terms of two factors of a review.
3. It offers an evolving model to measure the effect of the three features in a simulation method.

The rest of the paper is organized as follows: We start with the brief review on the related work. Next, the experiment of noise filtering is reported. After that, we study some basic properties of transactions and reviews on an e-commerce site. Then three features are described and certain correlation analysis is performed. We also propose and evaluate a model for the relationship between author behaviors and the reviews s/he left. Finally, we draw conclusions in the end of this paper.

Related Work

Product review analysis is a major branch of opinion mining research (Pang et al., 2002; Barabasi and Albert, 1999; Pang and Lee, 2005; Hu and Liu, 2004; Mishne and Glance, 2006). Pang et al. (2002) used machine learning techniques for sentiment classification. After applying Naive Bayes, Maximum Entropy, and Support Vector Machines on movie review data, the authors found that for sentiment classification, standard machine learning techniques definitively outperform human-produced baselines. Dave et al. (2003) exploited information retrieval methods for feature extraction and classification of positive or negative electronics product reviews. They concluded that although the performance is not so good due to noise and ambiguity, the results are qualitatively quite useful. In further investigations, Pang and Lee (2005) presented how to predicate the review ratings, which was formulated as a problem of multiclass text classification. They showed that significant improvements can be achieved over both multi-class and regression versions of SVMs when the proposed similarity measure is employed. Hu and Liu (2004) proposed a feature-based opinion summarization system, which focuses on mining and summarizing customer reviews of products posted on Web sites. The authors pointed out that the task is different from traditional text summarization and demonstrated the effectiveness of the techniques using a number of review samples. In the domain of movies, Mishne and Glance (2006) studied whether applying sentiment analysis methods to Web log data will result in better correlation than volume only. The major finding is that positive sentiment is a better predictor for a movie's success when related comments are posted prior to its release.

In another branch of study, researchers investigated the problem of review quality. The results reported by Jindal and Liu (2007) show that review spam and duplication are very popular. They proposed to perform spam detection via duplicate detection and classification. Spam detection was treated as a binary classification problem, spam and nonspam. Logistic regression was also used to learn a predictive model. In the process of model training and testing, Kim et al. (2006), Zhang and Varadarajan (2006) used the ground-truth derived from users' votes of helpfulness provided by Amazon. The results in (Kim et al. 2006) show that the most informative features include the length of the review, its unigrams, and its product rating. Zhang and Varadarajan (2006) viewed the problem as a regression task, and built regression models by combining a diverse set of features. They found that the perceived utility of a product review highly depends on its linguistic style. Liu et al. (2007) conducted similar experiments by using more linguistic and product-related features, such as the sentence length, the number of product features and so on. After that, they also attempted to improve the opinion summarization by detecting and filtering out low quality reviews. However, these research efforts did not study filtering out noise like seller feedback, and focused on the metadata and plaintext of reviews. We try to tackle the quality issue in another way, by considering both noise issue and the author's behaviors on e-commerce sites.

Feedback or reputation systems have been studied for years due to their importance for e-commerce sites. Recently, Khopkar et al. (2005) explored the usage history for a large panel of eBay sellers. Their analysis shows that behaviors of both sellers and buyers change in response to the changes in a seller's feedback profile. Resnick et al. (2006) performed a randomized controlled field experiment of eBay reputation mechanism. They found that one or two negative feedbacks for new sellers did not affect buyers' willingness-to-pay, which does not agree with our intuition. Our work differs from it in that we do not focus on the personal reputation impact on seller/buyer's future transactions, but rather on the impact on the reviews s/he composited.

Previous work also analyzes the structure of the Web and seeks to model Web evolution (Kleinberg et al., 1999; Kumar et al., 2000; Aiello et al., 2001; Laura et al., 2002; Eppstein and Wang, 2002). The Web is viewed as a directed graph in which a vertex represents a Web page and an edge represents a hyperlink from one Web page to another. Kleinberg et al. (1999) measured a set of properties of a Web graph and proposed a new family of random graph models. Kumar et al. (2000) presented an evolving copying model which could generate more bipartite cliques. In ACL models (Aiello et al., 2001), Aiello et al. expand the graph at discrete time steps with at least one vertex and at least one edge. Laura et al. (2002) proposed multi-layer model to capture the fractal structure of the Web generated by the presence of multiple regions within independent stochastic processes. Eppstein and Wang (2002) showed a steady-state model which results in a power law distribution without incremental growth. It has been found that we could model the user behaviors on e-commerce sites in a similar way, including make transactions, leave reviews and vote on reviews. Our work differs from Web graph modeling in that there are more than one kind of vertices and edges in a graph model. Basically, the vertices can be subdivided into users and reviews. The edges can be subdivided into three types of activities: transaction, review writing and review voting. This makes the modeling more complicated. The correlation analysis helps us create additional parameters to simulate the evolution better.

Noise Filtering

Dataset Building

We adopt eBay (http://www.ebay.com) as a test bed since it is one of the most matured e-commerce sites. Many types of e-commerce activities and review information are available on this platform. We study reviews in four popular product domains on eBay Reviews & Guides[3], i.e., cell phones, DVD, digital cameras and MP3 players. For each sampled review, we gather the title, content, number of votes, number of "helpful" votes, reviewer ID and timestamp. Moreover, for each review author, we collect all transactions he/she get involved as a buyer or a seller. Considering that a user's interests may change as time goes by, we focus on the ones that occurred either within 3 months before or within 3 months after the occurrence of the review. The transaction information gathered includes item title, item category, related feedbacks, feedback scores of sellers and buyers. Finally, we sample more than 155,000 reviews for 6,000 products, 37,000 reviewers and 285,000 transactions during October, 2008.

However, the 155,000 reviews cover not only reviews for product but also some noisy data like feedback for eBay seller and commercial spam. The first and most important step is to detect all non-review stuff. We treat this as a classification problem and divide all articles into three groups including product review, feedback and spam. The classification algorithm we utilize is SVM (Support Vector Machine). In order to verify the effectiveness of our approach, we took the following steps:

1. Randomly pick up 5,485 samples from all of the 155,000 articles and invited three human judgers to tag the data with Spam, Feedback and Product Review. We define a guideline and try to obtain the most objective results. More details of the guideline could be found in the appendix. Meanwhile we also found some complicated cases like a mixture of feedback and review. The samples are not expected to belong to more than one group here. We point out how to handle such cases in the guidelines.
2. After the human labeling, we could see the basic information of all groups in Table 1. As it can be seen, a high percentage (~23%) of the "product reviews" on eBay site is actually feedback for seller.
3. We did a multi-class classification through SVM-light4. The methodology is 5-folder cross validation. For evaluation metrics, we adopt micro-precision and micro-recall in context of classification (Yang, 1999). As it can be seen from Table 2, the accuracy performance for all the three categories is high and the research technique could be put into practice easily.

Table 1 - Noise filtering dataset

Product Review	Feedback for Seller	Commercial Spam	Total
4,060	1,286	139	5,485

Table 2 - Precision and Recall for the three groups

Product Review	Feedback for Seller	Commercial Spam	Total
Precision	94%	91%	98%
Recall	97%	84%	90%
Overall Accuracy	93%		

Quality Measurements and Feature Design

We found that noise is just one issue for nowadays product review during the whole process of human labeling. Reviews have various qualities even if they are from the same site and same product. In this section, we attempt to get a rough understanding of review authors' activities on e-commerce sites, which has not been fully studied in previous research work. First, we collect both reviews and transactions from the eBay platform. Then, we perform a high-level study on the two kinds of data. From the statistics, we draw two conclusions: 1) reviews have different levels of quality according to readers' voting; 2) authors' behaviors are very diverse across the whole community. After that, we explore whether there is any correlation exist between these two. Certain features are designed and extracted from transactions for the purpose of review quality estimation. A set of experimental data shows the effectiveness of user behavior in this task. To avoid confusion, the product review data we mention in the rest of the paper is the one with noise data removed.

Review Data Analysis

We first analyzed certain properties of reviews. On the eBay review site, readers are allowed to vote for any article with "helpful" or "not helpful". The distributions of the number of total votes and the number of "helpful" votes are taken into consideration for all the reviews in the four product domains mentioned above. Figure 1 and 2 show the following two relations respectively:

1. Number of reviews vs. number of votes
2. Number of reviews vs. number of "helpful" votes

As it can be seen from these two charts, both two kinds of votes have a power law distribution. A power law relation between two variables can be defined as follows:

$$y = \alpha x^k$$

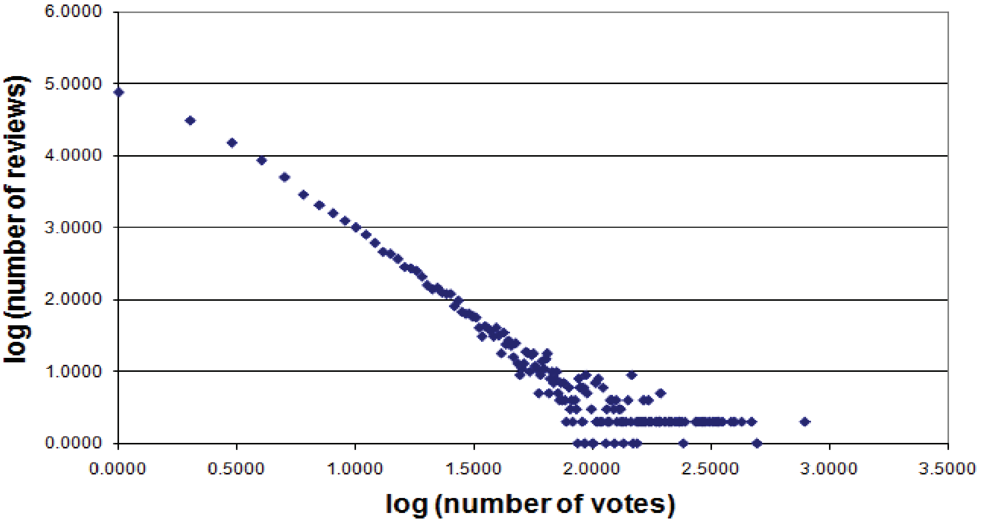

Figure 1 - Log-log plot of number of votes versus number of reviews

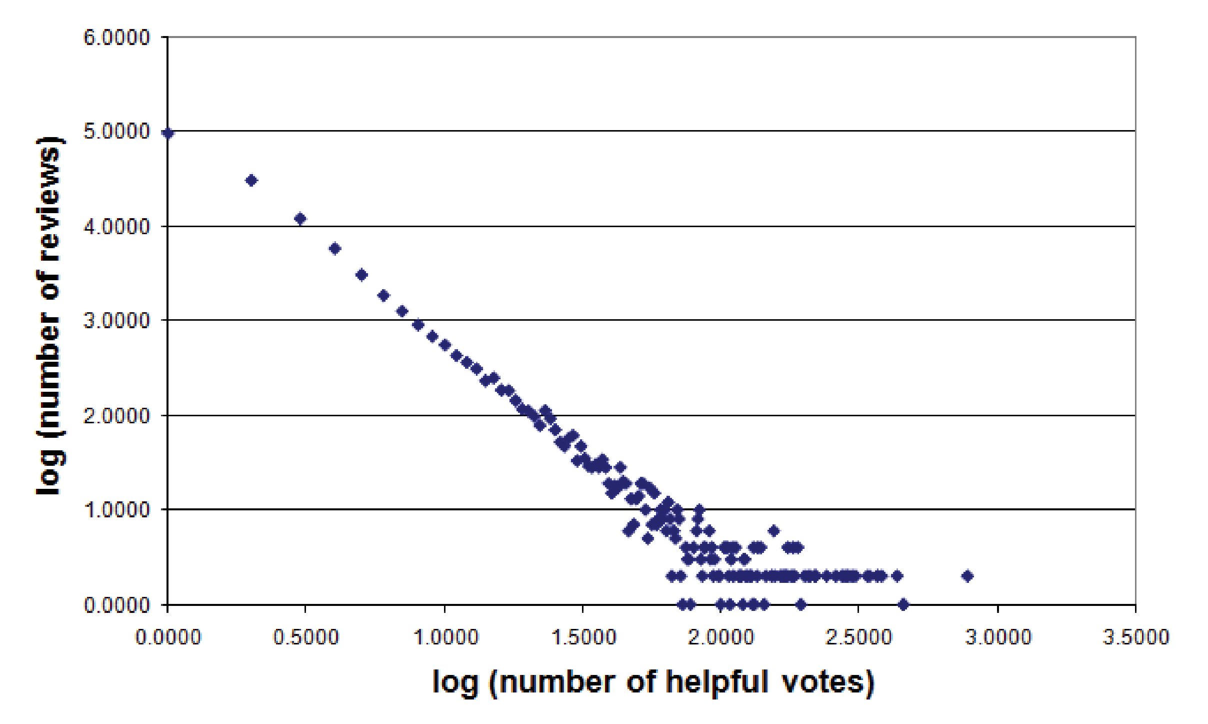

Figure 2 - Log-log plot of number of "helpful" votes versus number of reviews

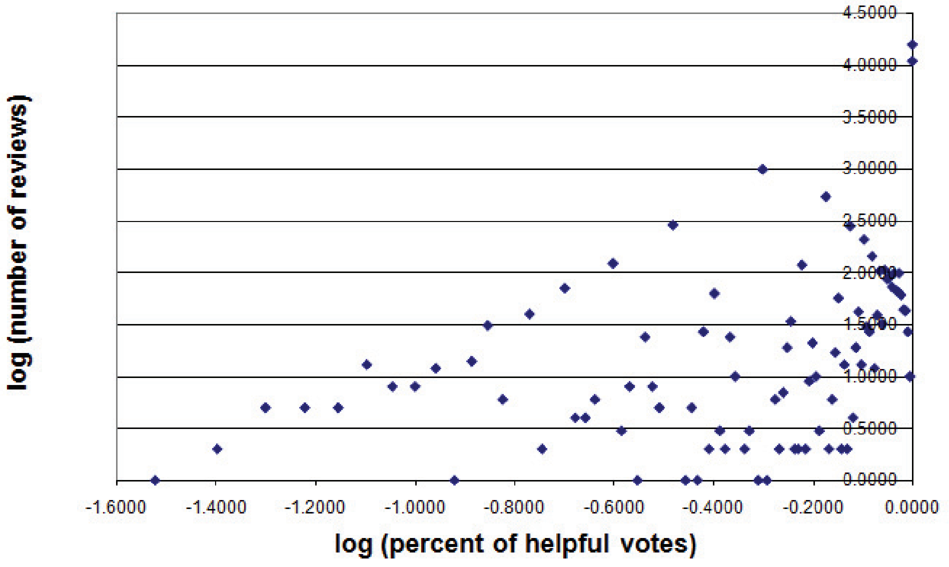

Figure 3 - Log-log plot of percent of "helpful" votes versus number of reviews

where and are constants. Figure 1 shows an almost straight line by taking logarithm scale on both axes. We can see that a small portion of the reviews attracts much more attention while a large number of reviews attract only little. As shown in Figure 2, the relation between the number of useful votes and the number of reviews also closely follows the power law distribution. In these two plots, is very close to 0.94 and is very close to -1.29. Figure 3 presents the relationship between number of reviews and percent of “helpful” votes. Although this distribution does not follow power law, it confirms the diversity of review quality and the necessary of our study.

Transaction Data Analysis

In our study, eBay users play a pivot role since they connect e-commerce and review communities. We conducted a user centric analysis using all transactions available. More specially, we center on two relationships:

1. Number of users vs. feedback score
2. Number of users vs. number of transactions

From Figure 4 and 5, we observe that distributions are also close to power law. In the first plot, is about 1.38 and is about -1.1. In the second plot, is about 5.61 and is about -1.81. Only a small portion of the users has high feedback score and high involvement in the business. We wonder if the behavior diversity will result in different review quality.

Another interesting phenomenon is related to the roles of sellers and buyers in transactions. To minimize the effect of noisy data on influence analysis, we first identify relevant transactions for each review from all the ones the review author participated in. The relevance between reviews and transactions will be introduced in the next section. In our experiments, a transaction and a review will be treated as related if the similarity between them is bigger than zero. The overall statistics on four product categories show that for a seller and one of his/her reviews, 67% of the related transactions occurred after the review. On the other hand, for a buyer and one of his/her reviews, 76% of the related transactions occurred before the review. In other words, sellers would like to write reviews before making transactions while buyers would like to leave reviews after transactions. One intuition behind this is that sellers usually promote their items via reviews and buyers usually make purchase decisions through reading reviews. The two percentage numbers indicate that the role of sellers/buyers in e-commerce may also impact product reviews.

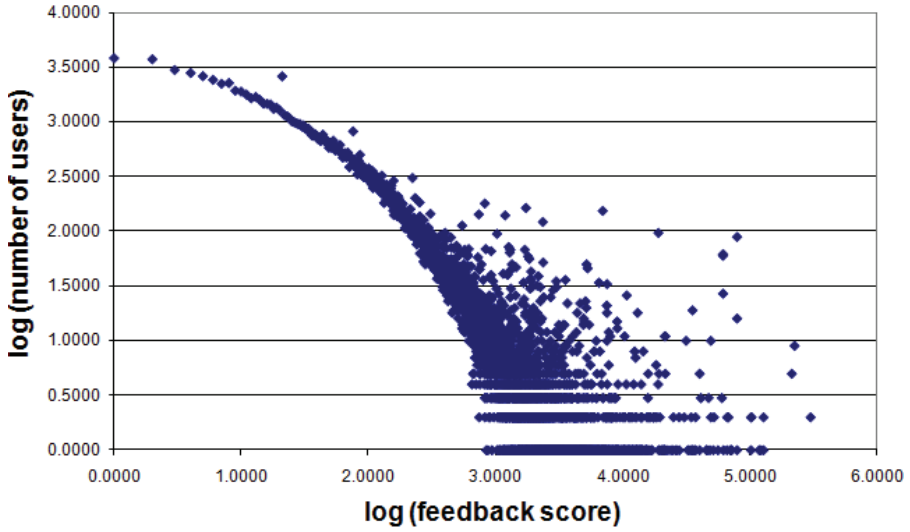

Figure 4 - Log-log plot of feedback score versus number of users

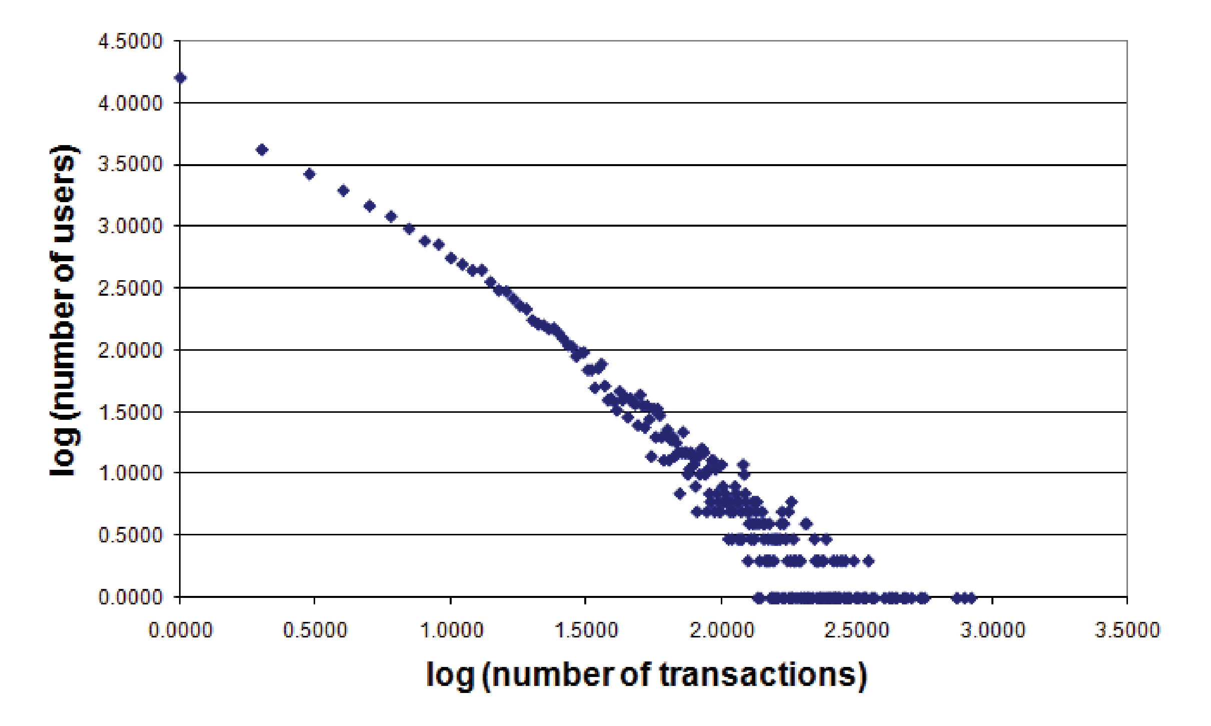

Figure 5 - Log-log plot of number of transactions versus number of users

Feature Design and Analysis

It is not easy to identify features directly from user behavior for the quality of product reviews. We attempt to evaluate the review quality from human understanding: Credibility and Expertise. Credibility indicates whether a review is trustable. Expertise indicates whether a review is full of professional knowledge in a specific domain. It is a challenging task to figure out credibility and expertise only from plain-text itself. To attack this issue we first introduce the concept of Transaction-Review relevance. Then, with this definition, several features are devised in terms of the two factors we conceive for review quality. Last, we study the correlation between the features and the percentage of helpful votes, which we use as a standard metric for review quality.

Transaction-Review Relevance

Personal credibility can be easily represented by feedback score on eBay site. However, gathering the information about user's role and expertise is not an easy task. Thus, we introduce an important concept: transaction-review relevance (abbreviated as t-r relevance in the rest of this paper). Let's take a case of digital camera. Two titles are selected: "Nikon D80 - The Camera You've Been Waiting For," the title of a review from eBay Reviews & Guides[5], and "NEW!~NIKON D80 SLR DIGITAL CAMERA KIT 10.2MP," the title of an eBay transaction[6]. We think these two are very relevant because they talk about the same type of Nikon camera. We therefore apply the VSM (Vector Space Model) (Salton and Buckley, 1988) on a corpus composed of all the titles of reviews and transactions. The assumption behind is that the higher content similarity of such two titles, the higher relevance between a review and a transaction. Feature vectors are built using the TF-IDF weighting scheme. Cosine function is used to measure the similarity between a review and a transaction. Finally, we achieve a similar score what we call t-r relevance.

Features Design

We propose features include personal reputation score, seller degree and expert degree by assuming that they can potentially reflect credibility and expertise of a person.

User Credibility

Here we propose two for user credibility: personal reputation and seller degree

1. Personal reputation score

Reputation systems are crucial for e-commerce sites. At eBay.com, users are allowed to leave others positive, neutral or negative feedbacks for deals. A good reputation score indicates one's honesty in transactions. As Figure 4 shows, the users with different levels of reputation participated in review writing activities. We wonder if the information provided by a person with a higher feedback score is more helpful for readers.

2. Seller Degree

This feature is devised based on user's deal behavior. As mentioned previously, sellers and buyers handle transactions and reviews differently. It becomes more necessary to

identify the user's role in a transaction. As Figure 6 illustrates, we calculate a value named seller degree for a given review in three steps:

- Find all the transactions, including both buying and selling, from the author of a review
- Accumulate the t-r relevance between each transaction and the review. For all the transactions in which the author plays a seller role, the t-r scores are counted as as str, otherwise the scores are counted as btr
- Seller degree sd is defined as

$$sd = \frac{\sum_{i=1,n} str_i}{\sum_{i=1,n} str_i + \sum_{j=1,m} btr_j}$$

The seller degree indicates how much a user contributed as a seller in all the deals. The transaction that is more related with the given review will have a higher weight. The value 1.0 means 100% seller role and 0.0 means 100% buyer role.

Expertise Degree

For a user, we also attempt to infer his/her degree of expertise from the goods s/he sold or bought. The assumption is that the users having many related transactions will be more experienced and are more likely to compose reviews with professional knowledge accordingly. Expertise degree is defined as follows where the t-r relevance is not grouped by sellers and buyers anymore.

$$ed = \sum_{i=1,n} tr_i$$

Correlation Analysis

To verify the significance of the above three features, we need to choose objective metrics of quality evaluation. One intuitive way to measure the quality is to leverage the percentage of "helpful" votes from review readers, which is similar to the approach used in (Kim et al., 2006; Zhang and Varadarajan, 2006). However, we will face two issues if using the percentage straightforwardly.

1. Voting sparseness. When the voting is sporadic, the percentage will not be reliable enough. For example, "100% helpful of 1 votes" doest not mean an article must be better than the one with "90% helpful of 100 votes".
2. Voting spam. Similar to the spam mentioned in (Jindal and Liu, 2007), "helpful" voting can be spammed too. The reviews on the eBay site are less vulnerable to cheating since a reader needs to register as a eBay user and the same person cannot comment on his/her own articles. Nevertheless, the potential spam behavior is unavoidable.

We assume that bigger number of total votes means higher significance of statistics and lower risk of being spammed. Therefore, only the reviews that won votes above average are adopted in the experiments. After that, we study the relationships between the percen-

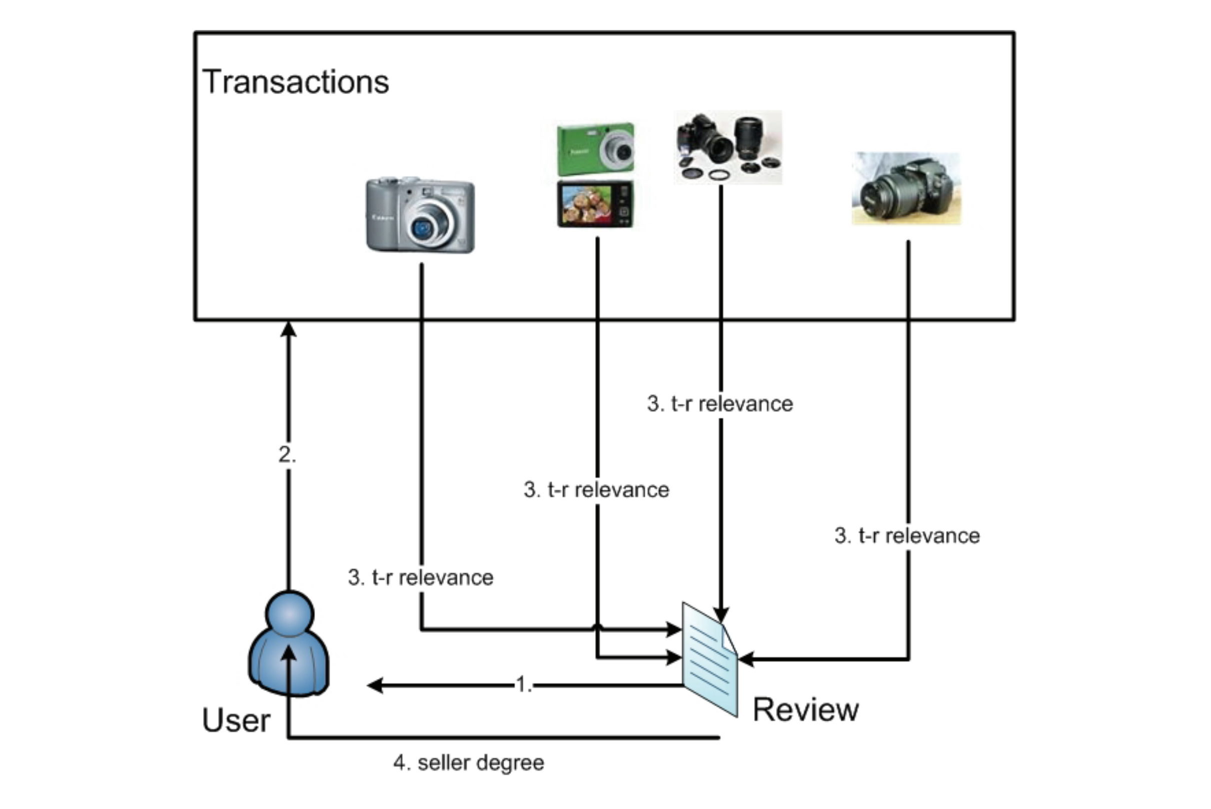

Figure 6 - Relevance between Transaction and Review

tage of “helpful” votes and the three features: user reputation, user role and expertise.

First of all, the Spearman rank correlation coefficient[7] is used to measure the relationship between variables that are placed on an ordinal or categorical scale of measurements. Table 3 shows the rank correlations (rho) as well as the p-values of the features on all categories. As we can see, all the three features are strongly correlated with the percentage of “helpful” votes across all product categories. One interesting and expected observation is that seller degree has a negative impact on the helpfulness based on reader’s voting. We examined the data and found that a possible reason is that some sellers make too much advertisement for their products. This observation is consistent with the conclusion in (Jindal and Liu, 2007) that review spam is becoming more and more common on review sites.

From the above results, we can see that the several features are positively or negatively correlated with the percentage of helpful votes. Here we also utilize a linear regression model to automatically learn the weight of each feature. Regression is used to determine the relationships between two random variables x = (x1, x2, ..., xp) and y. Linear regression (Hastie et al. 2001) attempts to explain the relationship of x and y with a straight line that fits to the data. The linear regression model postulates that:

$$y = b_0 + \sum_{j=1}^{p} b_j x_j + e$$

where the "residual" e is a random variable with a mean of zero. The coefficients bj (0 j p) are determined by the condition that the sum of the square residuals is as small as possible. Therefore the linear combination with bj’s should be better than those with any other coefficients. In our case, the indepen-

Table 3 - Spearman rank coefficients for the three features proposed

Feature	Cell Phone		DVD		Digital Camera		MP3 Player		All	
	rho	p-value	rho	p-value	rho	p-value	rho	p-value	rho	p-value
Personal Reputation	0.3704	<2.3E-5	0.4618	<2.2E-16	0.4401	<2.2E-16	0.4208	<2.2E-16	0.4491	<2.2E-16
Seller Degree	-0.7741	<2.2E-16	-0.9361	<2.2E-16	-0.8609	<2.2E-16	-0.8447	<2.2E-16	-0.8980	<2.2E-16
Expertise Degree	0.8199	<2.2E-16	0.8037	<2.2E-16	0.6357	<2.2E-16	0.4988	<2.2E-16	0.7198	<2.2E-16

dent variable x can be the values of the 3 features, x = (Feedback Score, Seller Degree, User Expertise), and the dependent variable y can be the combined score, which represents the predicted percentage of helpful votes. In the above equation, each single feature is normalized by the corresponding maximal value. In that way, we could find out which feature plays a more important role in the linear combination. The learned coefficients for all four categories are listed in Table 4. It confirms the conclusion that we made based on Spearman rank coefficients, i.e., feedback score and expertise degree are positively correlated with percentage of "helpful" votes while seller degree is negatively correlated with the percentage of "helpful" votes. The coefficients are also used in the experiments of model validation in the next section. As the simulation results show, the characteristics are closer to the real dataset if we specially consider such features.

Effectiveness of User Behavior

We discuss a lot how the relationship between human behavior and review quality. A straightforward experiment is necessary to show whether we could improve this task by involving behavior feature. We selected one recent study (Liu et al. 2007), which is closely related to our work, and followed the experimental approach in that paper to build a baseline. The major steps include:

1. Annotation of quality. Human judgers will subdivide the review articles into several groups including "best", "good", "fair" and "bad".
2. Feature development such as informativeness, readability and subjectiveness
3. SVM model based classification and evaluation.

More details could be found at Liu's publication (Liu et al. 2007). We also made two big changes to make the experiment fit out test environment better.

1. Product feature, which is one of the informativeness features in previous work. Since our review data is from eBay, we directly use the catalog information of eBay business site. The catalog has been build for decades in eBay to offer product feature level information. It's definitely enough for the purpose of feature develop in the electronics domains we used.
2. Evaluation method. We did not adopt annotation method because we found the boundaries among "best", "good", "fair" and "bad" are not so clear. It's really hard to get consistent results from different judgers. Therefore the voting of "helpfulness" is still used here. We firstly sort all reviews in terms of the "helpfulness" percentage mentioned. Then the top 25%, the second 25%, the third 25% and the last 25% will be marked as "best", "good", "fair" and "bad" respectively.
3. Linear combination of different features. Liu et al. test several individual features separately but did not combine all of them to see if better accuracy could be achieved. We adopt a basic linear combination for this step, i.e. each feature has equal weight. This is named as "Combination A" in Table 5.

Finally we get the accuracy as follows:

Table 4 - Linear regression analysis for the three features proposed

Feature	Cell Phone		DVD		Digital Camera		MP3 Player		All	
Interception	0.9788	R^2	0.9841	R^2	0.9867	R^2	0.9398	R^2	0.9382	R^2
Personal Reputation	0.0032	0.384	0.0644	0.391	0.0512	0.229	0.0902	0.263	0.0352	0.326
Seller Degree	-0.0278		-0.0072		-0.0149		-0.0082		-0.0130	
Expertise Degree	0.1089		0.0041		0.0628		0.0503		0.0368	

Table 5 - Accuracy of quality classification – baseline

Feature Category		Accuracy
Informativeness	SL	75.2%
	WL	81.7%
	PFL	76.3%
Readability		81%
Subjectiveness		75.3%
Combination A (each weight 0.2)		82.2%

Table 6 - Accuracy of quality classification – user activity incorporated

Feature Category	Accuracy
Combination A + personal reputation (each weight 0.5)	0.842
Combination A + seller degree (each weight 0.5, seller degree score has been transformed due to negative correlation)	0.823
Combination A + expert degree (each weight 0.5)	0.836
Combination + all three (each weight 0.25)	0.859

From above two tables, we found that combination of different features, especially user behavior will improve the accuracy of quality classification. Results are promising and we could expect better ones achieved with fine tune of the weights.

Behavior Modeling

Evolving Model

In this section we try to model the sociology behavior related to online business and review composition. The motivations of the modeling include but are not limited to:

1. It allows us to make predictions about review quality before any human voting
2. It allows us to estimate the impact on product sales from reviews of low or high quality
3. It allows us to conduct the e-commerce behavior study in a simulated environment

Our model seeks to capture the two intuitions: (1) Some users would like to leave online re-

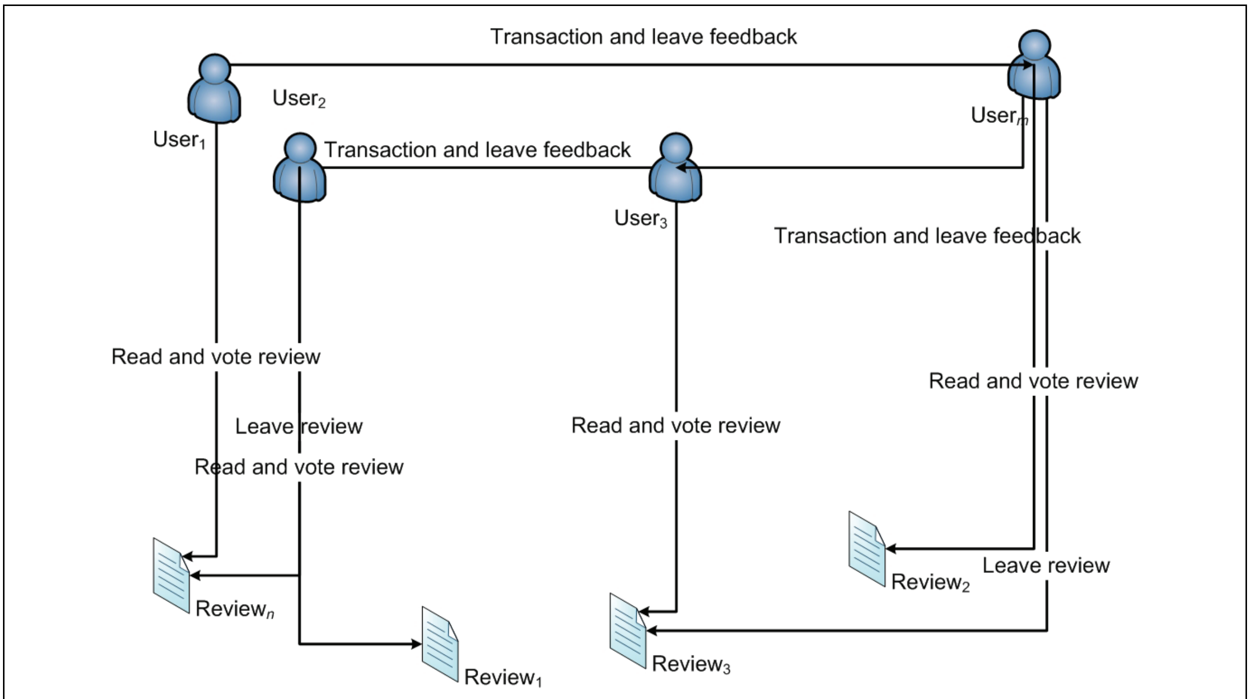

Figure 7 - Graph Model for Activities in Community

views during the e-commerce activities; (2) Some users would like to read and vote on online reviews before buying or selling on the Web. Therefore the model is characterized by three stochastic processes:

1. A user buys or sells products online. Then, a user will leave feedbacks to others, or receive feedbacks from others on each transaction. This will cause more generations of feedbacks and reviews.
2. A user writes a review. Therefore more user generated content is available for the whole community to consume.
3. A user reads and votes on a review. Since we evaluate the quality of reviews according to user's votes, the reading without any votes is not considered.

This model does not take into consideration the impact of reviews on user's shopping behaviors. We use Figure 7 to illustrate the relations among different entities. Users and reviews are treated as vertices in a graph. The transactions and votes are considered as edges between vertices.

Based on the above statement, we propose a graph model that evolves over discrete time steps t = 1, 2, …. At time t, let the graph be G_t = $<U_t, R_t, T_t, L_t, V_t>$, where U and R denotes user and review vertices respectively; T, L and V denote three types of edges respectively: generating a transaction, leaving a review and voting on a review. In order to simplify this model, activities such as registration of new users, deletion of expired users, and release of new products are not considered. Three functions are required to characterize the evolution of the graph. We describe the growth of transaction edges using function $f_e(f_{uu}, G_t, t)$. It is a probabilistic process that returns a set of transactions to be added and positive/neutral/negative feedbacks associated with the transactions. The growth of review voting edges is captured by $f_e(f_{urv}, G_t, t)$, which returns a set of votes for views. The growth of review authoring edges is captured by $f_e(f_{url}, G_t, t)$, which returns a set of reviews to be added. Compared to $f_e(f_{urv}, G_t, t)$, $f_e(f_{url}, G_t, t)$ also needs to generate new vertices of reviews. As discussed previously, three features are required for a review author in our model: feedback score, seller degree and expertise degree. Feedback score and customer satisfaction can be inferred directly from the information of users. We need to randomly generate *t-r* relevance since no product related entities are consolidated in the model. In that way we are able to estimate the seller degree and the expertise degree.

Supposing our model is composed by only three stochastic processes, we design the following parameters to control the model.

1. $\gamma_1(0<\gamma_1<1)$: the probability of generating a transaction between any two users.
   a) α_1, α_2 and α_3 ($\alpha_1+\alpha_2+\alpha_3 = 1.0$): the probabilities that a user leaves a positive, neutral or negative feedback, respectively.
2. $\gamma_2(0<\gamma_2<1)$: the probability that a user leaves a review.
   a) δ: the possibility that a transaction and a review are both related to the same product.
3. $(1-\gamma_1-\gamma_2)$: the probability that a user reads and votes on a review.
   a) θ: the probability to gain a "helpful" vote. As the empirical results in rest section will show, we could optimize this parameter by using feedback score, seller degree and expertise degree.

We call this model ERM (E-commerce and Review Model). In fact some key elements are not involved in this model. For example, a transaction will be more likely to happen for the sellers with higher reputation; the review with more "helpful" votes will get more chance to be browsed etc. We ignore such phenomena to keep the model simple and try to concentrate on the three features related to the review quality. Investigation of such factors is a direction of future research.

Model Validation

Basically two approaches can be used for model validation: theoretical analysis and experimental simulation. We employ a simulation experiment due to the complexity of the model consolidating commercial and review behaviors. This approach requires special attention to the selection of experimentation parameters and output analysis. The experiment attempt to seek answers for the following questions:

- Whether the model will create a power law distribution for all votes for a review?
- Whether the model will create a power law distribution for "helpful" votes for a review?
- Are the three features able to help the model to create the results of simulation closer to the reality?

The operation of ERM is simulated as follows: we set the initial value of each parameter according to the statistics of our dataset used in previous experiment, including the probabilities γ_1, γ_2, α_1, α_2, α_3, δ and θ. Since no user registration and expiration are considered in the model, we fix the number of users to 10,000 in the experiments. In each step, three actions are simulated:

1. A transaction is generated between a pair of users. The action of leaving feedback is also involved in this process.
2. A review is generated after each transaction.
3. A vote on each review is generated by a user.

Finally, after running 1,000 time steps, we generated more than 600,000 reviews and 775,000 transactions. We gathered statistics of users and reviews. Figure 8 is the log-log plot of number of votes versus number of reviews, and Figure 9 is the log-log plot of number of "helpful" votes versus number of reviews. As shown in Figures 8 and 9, the two are not very close to a power law distribution. There are two obvious curves in the middle of the two lines. Therefore we investigated whether the three features introduced will improve the situation. In another trial, the possibility of obtaining a "helpful vote" is not controlled by a predefined parameter θ anymore. For the voting process, we first calculate the author's feedback score, seller degree and expertise degree with regards to

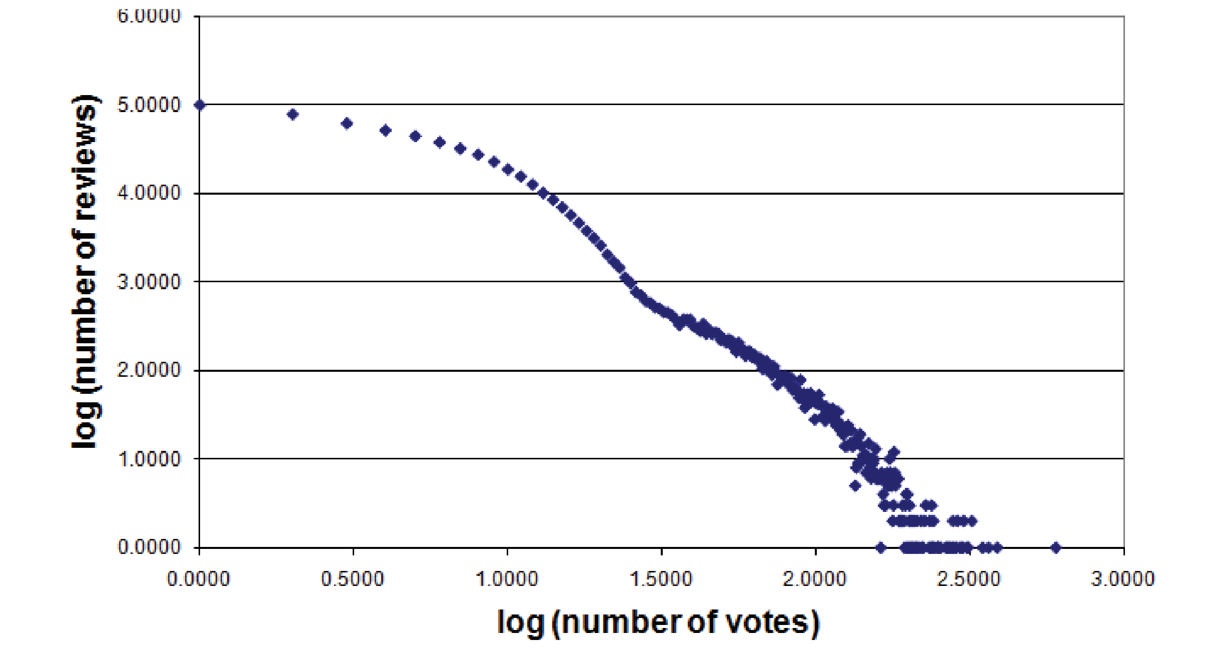

Figure 8 - Log-log plot of number of votes versus number of reviews on the dataset generated by the ERM model

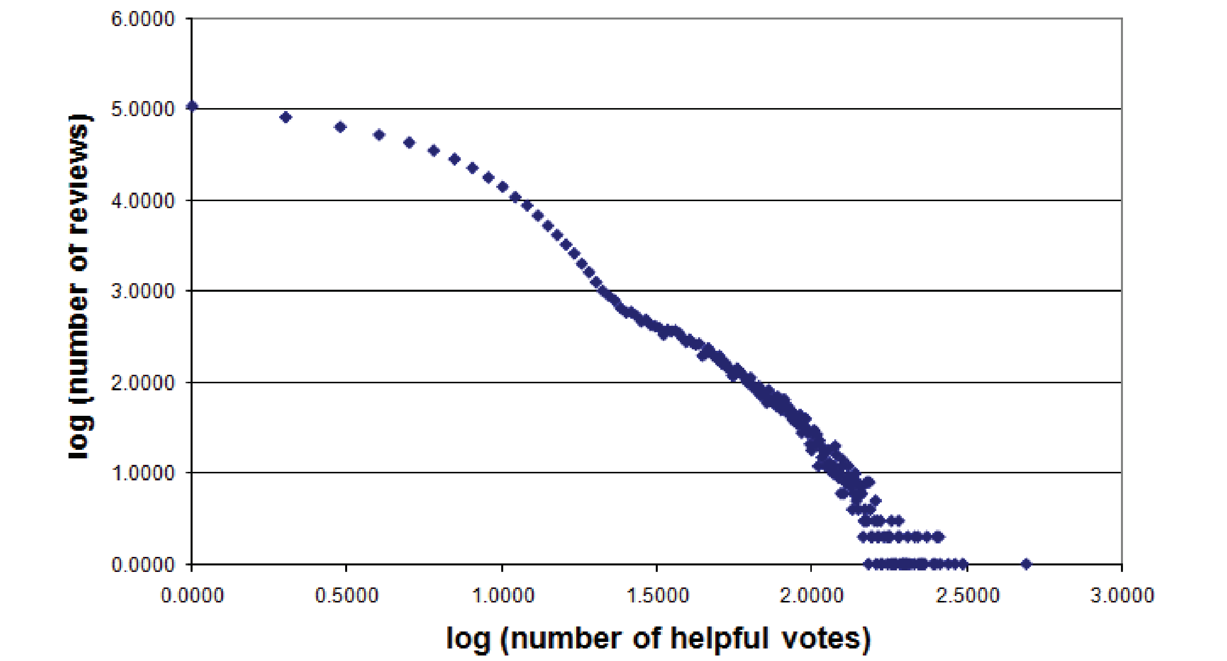

Figure 9 - Log-log plot of number of "helpful" votes versus number of reviews on the dataset generated by the ERM model

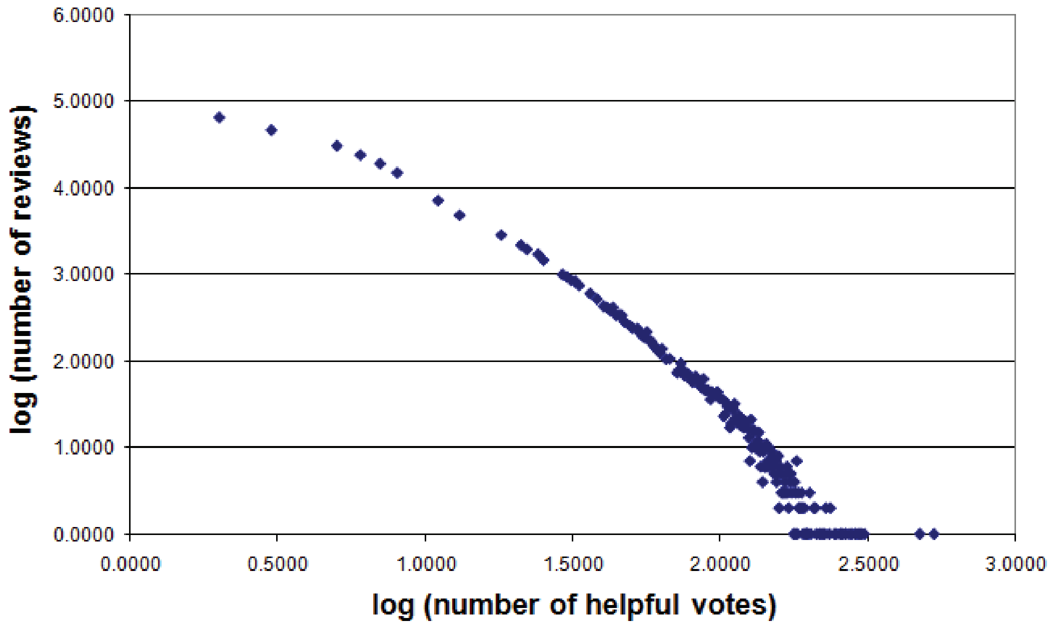

Figure 10 - Log-log plot of number of "helpful" votes versus number of reviews on the dataset generated by the refined ERM model

current transactions. After that, the probability of "helpful" votes is estimated by linearly combining the above scores. The weights in "All" column of Table 4 are adopted. With such a refinement, we get a smooth distribution for log-log plot of number as shown in Figure 10, which is much closer to a straight line. For this plot, α is about 1.77 and k is about -1.85. The two numbers are also close to the observed values from the real dataset (α ~ 0.94, k ~ -1.29). This result demonstrates the effectiveness of considering author behaviors from another perspective.

Conclusion and Future Work

In this paper we consider the problem of review quality from a new point of view. In addition to the common noise such as commercial spamming and feedback for retailer, we pay more attention to the activities closely related to the author of a review in an e-commerce environment. The empirical results show that the examined features are correlated with the review quality as assumed and indeed help us on this research task. An evolving model is also proposed and evaluated by the capability to approximate the power law distributions of the review dataset. In future work, we will focus on several aspects: first, we will consider if it possible to identify more features like user behavior as a buyer. Second, we will predict the review quality by combining both author e-commerce behaviors and the review content. Finally, we plan to create a better model. Barabasi and Albert (1999) suggested a better model to simulate power law. Besides, we do not characterize the influence of reviews on product sales so far. A new model is expected to consolidate the bi-directional influence between commerce and review behaviors.

Footnotes

[1] Forrester research: http://www.forrester.com/rb/ research, last viewed on November 1, 2008.

[2] CompUSA: http://www.compusa.com/, last viewed on November 1, 2008.

[3] eBay Reviews and Guides: http://reviews.ebay.com, last viewed on November 1, 2009

[4] SVM-light multi-class: http://svmlight.joachims.org/svm_multiclass.html, last viewed on November 1, 2009.

[5] http://catalog.ebay.com/Nikon-D80-10-2-Megapixel_W0QQ_fclsZ1QQ_pidZ55042806QQ_tabZ3

[6] http://cgi.ebay.com/NEW-NIKON-D80-SLRDIGITAL-CAMERA-KIT-10-2MP_W0QQitemZ190307265250QQcmdZViewItemQQptZDigital_Cameras?hash=item2c4f326ae2&_trksid=p3286.c0.m14&_trkparms=66%3A2|65%3A16|39%3A1|240%3A1318|301%3A0|293%3A1|294%3A50 (This link may be expired after several months because eBay consistently refresh the items on its platform)

[7] Spearman rank correlation coefficient: http://en.wikipedia.org/wiki/Spearman_correlation, last viewed on November 1, 2008.

Acknowledgements

We are thankful to Jean-David Ruvini, Xiaoyuan Wu, Johnson Wang and Neel Sundaresan. Jean-David helped us a lot on the noise filtering approach and experiment design. Xiaoyuan has given us valuable suggestions and helped us polish the paper. Johnson and Neel greatly supported our research projects all the time. Moreover, we want to express our appreciations to every person who contributed with either inspirational or actual work to this paper.

References

Aiello, W., Chung, F. and Lu, L. (2001). “Random Evolution in Massive Graphs,” *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, 510–519.

Barabasi, A.-L., and Albert, R. (1999). “Emergence of scaling in random networks,” *Science*, 286, 509-512.

Dave, K., Lawrence, S. and Pennock, D.M. (2003). “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews,” *Proceedings of the 12th International World Wide Web conference*.

Eppstein, D. and Wang, J. (2002). “A Steady State Model for Graph Power Law,” *Proceedings of the 2nd International Workshop on Web Dynamics*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001) “The Elements of Statistical Learning,” *New York: Springer-Verlag*.

Hu, M., and Liu, B. (2004). “Mining and Summarizing Customer Reviews,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Jindal, N. and Liu, B. (2007). “Product Review Analysis,” *Technical Report*, UIC.

Khopkar, T., Li X. and Resnick P. (2005). “Self-selection, Slipping, Salvaging, Slacking, and Stoning: The Impacts of Negative Feedback at eBay,” *Proceedings of the 6th ACM Conference on Electronic Commerce*.

Kim, S.-M., Pantel, P, Chklovski, T. and Pennacchiotti, M. (2006). “Automatically Assessing Review Helpfulness,” *Proceedings of EMNLP*.

Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. S. (1999). “The Web as a Graph: Measurements, Models and Methods,” *Proceedings of the 5th International Computing and Combinatorics Conference*.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. and Upfal, E. (2000). “Stochastic Models for the Web Graph,” *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 57–65.

Laura, L., Leonardi, S., Caldarelli, G. and De Los Rios, P. (2002). “A Multi-layer Model for the Web Graph,” *Proceeding of the 2nd International Workshop on Web Dynamics*.

Liu J., Cao Y., Lin C.-Y., Huang Y. and Zhou M. (2007). "Low-Quality Product Review Detection in Opinion Summarization," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Mishne, G. and Glance, N. (2006). "Predicting Movie Sales from Blogger Sentiment," *Proceedings ofAAAI-CAAW*, the Spring Symposia on Computational.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Pang, B. and Lee, L. (2005). "Seeing stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales," *Proceedings of the ACL*.

Resnick, P., Zeckhauser, R., Swanson, J. and Lockwood, K. (2006). "The Value of Reputation on eBay: A Controlled Experiment," *Experimental Economics*, 9(2), 79-101.

Salton, G. and Buckley, C. (1988). "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, 24(5), 513-523.

Yang, Y. (1999). "An Evaluation of Statistical Approaches to Text Categorization", *Information Retrieval*, 1(1-2), 69-90.

Zhang, Z. and Varadarajan, B. (2006). "Utility Scoring of Product Reviews," *Proceedings of CIKM*.

About the Authors

Dr. Shen Huang is a researcher from eBay research lab. He received the PhD degree from Shanghai Jiao Tong University, China, in 2000. Then he worked as an associated research at Microsoft Research Asia. After that, he joined eBay to pursue his research interests such as opinion mining and merchandise on E-Commerce sites. Accordingly, he participates some projects aim to improve eBay reviews and business platforms.

Dr. Dan Shen works as a researcher in eBay research lab. She received the PhD degree from Saarland University, Germany in 2007. She is mainly interested in the areas of natural language processing, machine learning and text mining. Now, she is involved in the projects related to opinion mining and merchandising in eBay.

Wei Feng is a second-year master student at Department of Computer Science, University of Toronto, Canada. She received her undergraduate degree from Shanghai Jiao Tong University, China, in 2009. She worked as an intern at eBay Research Lab –China during June, 2008 – Feb, 2009. She is now working with Professor Graeme Hirst in Toronto. Her current research interest is computational linguistics.

Dr. Catherine Baudin is a Principal Research Scientist at the eBay Research Labs. Catherine's expertise is in text mining/knowledge discovery, semantic indexing for vertical search, log analysis and user studies. At eBay her research revolves around session log analysis to study user queries and fuel the design of ecommerce components, finding metrics, sentiment analysis in user blogs and pattern identification for fraud analysis.Prior to joining eBay, Catherine was the Chief Scientist/CTO of Kaidara Inc. Prior to that, Catherine was a senior research scientist at the PriceWaterhouseCoopers technology center in R&D for five years and was a principal investigator at the NASA Ames Research Center for six. Catherine holds a PhD in Artificial Intelligence from the University of Paris VI France. She is the author of numerous peer reviewed papers and the co-author of several patents.

Dr. Yongzheng Zhang obtained his Ph.D. in Computer Science from Dalhousie University, Canada in 2007. Since May 2006, he has been a research scientist at eBay Inc. in San Jose, California, USA. His research interests are in the areas of Information Retrieval, Natural Language Processing, Web Mining, and e-Commerce. He has published several journal and conference articles on Web site summarization and product review mining.

Appendix

Guidelines of Labeling for Noise Filtering

Spam: scams or non-sense

Feedback: response to seller and transaction

Product review: good quality review relevant to a specific product

Features of each level can be summarized as follows:

Spam:

1. how to get the product for free
2. irrelevant comments with the product
3. non-sense words

Feedback:

1. items in good/bad condition
2. low price
3. quality issues
4. seller didn't give response
5. expressing thanks to the seller
6. promise future transaction
7. complaints on eBay service
8. simple reasons to buy + transaction issues
9. Major feedback with few comments on product

Mixed:

1. personal feeling + transaction comments
2. product features + transaction comments

Product Reviews:

1. a little touch of personal feeling + a few impressive review
2. professional reviews but in a short passage
3. containing links to outside review sites which provides great reviews
4. very detailed information about the product, always long and objective
5. Major comments on product with little feedback

After the first round of labeling, we may get some mixed ones. Human judgers will discuss together and re-examine the mixed ones. If major part is review for product then it will finally go to review group. Otherwise go to feedback group. Occasionally we discard some cases hard to judge to ensure the quality of labeling data.