Incorporating Expectations as a Basis for Business Service Selection

Adel M. ElMessiry, Xibin Gao, and Munindar P. Singh

North Carolina State University, Raleigh NC 27695, USA
{ammessir,xgao2,singh}@ncsu.edu

Abstract. The collaborative creation of value is the central tenet of services science. In particular, then, the quality of a service encounter would depend on the mutual expectations of the participants. Specifically, the quality of experience that a consumer derives from a service encounter would depend on how the consumer's expectations are refined and how well they are met by the provider during the encounter. We postulate that incorporating expectations ought therefore be a crucial element of business service selection.

Unfortunately, today's technical approaches to service selection disregard the above. They emphasize reputation measured via numeric ratings that consumers provide about service providers. Such ratings are easy to process computationally, but beg the question as to what the raters' frames of reference, i.e., expectations. When the frames of reference are not modeled, the resulting reputation scores are often not sufficiently predictive of a consumer's satisfaction.

We investigate the notion of expectations from a computational perspective. We claim that (1) expectations, despite being subjective, are a well-formed, reliably computable notion and (2) we can compute expectations and use them as a basis for improving the effectiveness of service selection. Our approach is as follows. First, we mine textual assessments of service encounters given by consumers to build a model of each consumer's expectations along with a model of each provider's ability to satisfy such expectations. Second, we apply expectations to predict a consumer's satisfaction for engaging a particular provider. We validate our claims based on real data obtained from eBay.

1 Introduction

This paper investigates the problem of business service selection based on an expanded notion of reputation and trust. It is widely recognized now that the collaborative creation of value is the central tenet of services science [13]. Specifically, the importance of understanding human behavior as a basis for service science is well-recognized, but is not necessarily reflected in the technical approaches developed by computer scientists. In particular, then, the quality of a service encounter would depend on the mutual expectations of the participants. Specifically, the quality of experience that a consumer derives from a service encounter would depend on how the consumer's expectations are refined and met by the provider during the encounter. Indeed, this is well-known in marketing theory as the service quality GAPS model as a basis of customer satisfaction [16]. This model, however, is traditionally applied from the perspective of the

P.P. Maglio et al. (Eds.): ICSOC 2010, LNCS 6470, pp. 486–500, 2010.

service provider in terms of its marketing and operations. In contrast, we postulate that incorporating expectations is a crucial element of business service selection as well.

Unfortunately, today's technical approaches to service selection rely upon combining numeric ratings without regard to what the raters' frames of reference, i.e., expectations, might have been. When the frames of reference are not modeled, the resulting reputation scores are often not sufficiently predictive of a consumer's satisfaction. Accordingly, the main claim of this paper is that reputation scores produced merely by an aggregation of context-free numeric ratings are not significantly effective in producing trust. Specifically, this paper proposes to explicitly consider the *expectations* of the parties involved in order to arrive at a finessed notion of reputation that an agent may use as a basis for trusting others.

Both to show the practicality of the above claim and to evaluate it rigorously, this paper considers the important setting of e-commerce interactions, such as the marketplaces of eBay and Amazon. E-commerce settings provide an immediate and widespread application for research into service selection. Further, they provide a source for independent, real-life data with which to objectively evaluate research claims. Such real-life evaluations are generally not prominent in the services literature.

To further motivate the problem, consider a buyer who is faced with a decision to select a seller from a group of sellers. Other things being equal, a buyer would rationally decide based on the experiences of previous buyers with the various sellers. For this reason, e-commerce sites include reputation systems whereby buyers can state a numeric rating of a seller with whom they interact (and sellers can rate buyers, but we do not consider those here). A subsequent buyer can use those ratings to select a suitable seller. This buyer too would rate the seller he chose, thus helping maintain the information in the reputation system. Current reputation systems aggregate numeric ratings and present a simple measure of a potential seller's quality.

In general, the better the reputation a seller accrues the more trustworthy it becomes. The fundamental deficiency of this approach lies in its presumption that we can simply combine ratings by different users. Doing so assumes that the different users have the same frame of reference. Such naïve aggregation may be acceptable in some cases, e.g., where a seller has obtained a large number of ratings from homogeneous buyers, but is not valid for many practical settings involving smaller sets of ratings, especially when the ratings differ in a way that can matter to a prospective buyer. Although reputation aggregated solely from numeric ratings can be useful, it often misses the point of what a buyer seeks. This is because the various ratings are given by different buyers based on their respective frames of reference. It would be surprising if simply aggregating such ratings would yield the most valuable information for a prospective buyer.

This paper is based on the idea that a key aspect of the frame of reference of a buyer is captured in the buyer's expectations. When a buyer's expectations are met, his experience is pleasant, and one would assume his rating of the seller is positive. More importantly, to predict the buyer's quality of a buyer's experience and his ultimate rating, we need to look beyond simply the ratings given by other buyers, and also incorporate the expectations that underlie those ratings. When we relate the expectations of a buyer with the expectations of previous raters, we would produce a more accurate recommendation and a more justifiable basis for the buyer to select a seller.

A natural challenge is how to estimate the expectations of a buyer. Fortunately, e-commerce settings provide a clue as to their users' expectations through text comments (termed *feedbacks*) that a user may produce in addition to a numeric rating. A user's feedback often describes the user's experience from a specific transaction and gives reasons for the associated rating. Although feedbacks are free form, we find that their vocabulary is generally quite restricted. Therefore, we can mine text feedback reasonably effectively to understand its author's expectations for the given interaction.

Contributions. We begin from the prima facie reasonable assumption that users with shared expectations would share a similar degree of satisfaction from their respective encounters with the same business services. Our main contribution is to refine and validate this assumption. We show that applying expectations in a common e-commerce setting yields better predictions of ratings than otherwise possible. Further, we show that the expectations of buyers can be reliably and effectively mined from the text feedbacks they produce. Additionally, through the use of abstract expectations, this approach can help match buyers and sellers even if there is no direct relationship between them. This is crucial in overcoming the sparsity of data, e.g., with respect to new buyers and sellers.

Organization. Section 2 introduces expectations, a representation for them, and our approach. Section 3 describes our evaluation methodology and presents our results. Section 4 discusses some relevant literature and some future research challenges.

2 Understanding Expectations

There is fairly strong support in the literature on consumer behavior for the notion of expectations. Kim et al. [11] observe that the fulfillment of a consumer's expectation is a key factor in the consumer's satisfaction, and may indirectly influence the consumer's intention to repurchase from the same seller. The approach of this paper reflects the intuitions of *Expectation-Confirmation Theory* due to Bhattacherjee [4], which is a leading model of consumer satisfaction. To understand the relation among expectations, satisfaction, and ratings, we consider a three-phase model.

Formulate expectations. The customer identifies his requirements and expectations.
Transact. The customer selects a seller and carries out the interaction.
Evaluate. The customer compares his expectations with his experience. The customer' expectations being met or *confirmed* correspond to greater satisfaction, and thus a higher rating of the seller. The customer' expectations not being met correspond to (partial or total) dissatisfaction, and thus a lower rating of the seller.

2.1 Expectation and Reputation Profiles

To realize the above approach computationally, we need to express expectations, automatically infer expectations, and compare them. A simple representation proves quite effective. We can think of each expectation as a name-value pair: the value describes the strength of the corresponding expectation as a real number in the interval $[0, 1]$.

It is convenient to write the expectation profile of a consumer as a row vector whose columns are interpreted as the expectation attributes and whose cells are the corresponding values. For example, in a two dimensional setting, we may interpret $\langle 0.9, 0.1 \rangle$ as the profile of a consumer who expects a high *Level of Service* and is relatively unconcerned with *Shipping Time*. The order in which the expectations are written is irrelevant, but we require that the order is (arbitrarily) fixed so we can perform sound calculations on the vectors. We use vector such as the above as the main representation in our approach:

Buyer's expectation profile based on the buyer's previous interactions. This represents the buyer's typical expectations as the buyer enters into an encounter.

Seller's reputation profile based on the previous interactions of buyers with this seller. This represents the typical preexisting expectations of a buyer who enters into an encounter with this seller.

We observe that feedbacks associated with negative ratings from a buyer yield a more meaningful estimation of the buyer's expectations. When buyers give negative ratings, they often elaborate on why. By contrast, with positive ratings, they often merely state that the experience is good. Some studies [19] also show that eBay auctions are mildly influenced by positive ratings, however, negative ratings emerged as highly influential and detrimental. Thus, in this paper, we focus exclusively on negative feedbacks to induce expectation and reputation vectors.

For each buyer, we create an expectation profile based on the buyer's comments. For each seller, we create a reputation profile based on the comments posted by the buyers who have interacted with that seller and given it negative feedbacks. Thus, the seller's reputation profile is negative: it captures expectations that the seller does not meet well. From a match between a prospective buyer and a target seller we can estimate how unsuccessful the buyer's experience with that seller will be. That is, *the stronger the match the greater the chances of the buyer's expectations not being met*.

2.2 Analyzing Feedback to Infer Expectations

We consider the following expectation attributes specialized for e-commerce services: *Item (is as described)*, *Communications (are effective)*, *Shipping time (is small)*, *Shipping (and handling) charges (are appropriate)*, *(Level of) service (is high)*. For brevity, below, we omit the parenthesized parts of the names of each attribute. Notice that *Item*, *Communications*, and *Service* are subjective qualities.

As we remarked above, often in practical settings, the set of text feedbacks given by a user is the only source of knowledge we have of the user's expectations. We adopt the techniques of sentiment and affect analysis of text to infer a user's expectations. Sentiment analysis assesses the directionality of a text fragment and asserts if it is positively or negatively oriented [15]. Affect analysis [1] seeks to identify the emotions or affect classes indicated in a text fragment.

We analyze expectations in analogy with affect, and abstract the expectation vector construction process as a multiclass, multilabel text classification problem. An expectation vector has five dimensions corresponding to the above attributes. The value of each attribute represents its strength. For example, $\langle 0.1, 0.9, 0.0, 0.0, 0.0 \rangle$ means that the user has a strong concern with communication, and does not care about shipping

time, shipping charges, or service. The expectation vector for a buyer is constructed by aggregating the class labels for all the feedbacks the buyer left.

For each textual feedback, the vector is assigned by a text classifier. The class labels are the above attributes plus *Others* because some feedbacks fall outside the five attributes. For example, some feedbacks are in Spanish, and some only contain symbols. Multiple class labels can be assigned to each feedback because multiple concerns can be expressed in each feedback. For example, "Dirty console. Did not respond. Non Working Console" alludes to *Item* ("Dirty console" and "Non Working Console") and to *Communication* ("Did not respond").

We apply text processing techniques to analyze the feedbacks and induce an expectation vector from these feedbacks.

Clean up the text using the Google Spell Checker service [8] to replace wrongly spelled words and thus reduce the noise in the input. The checker mostly suggests correct words, mapping "recieved" to "received" and "emials" to "emails." However, this step is not perfect. For example, it maps "wii" to "WI."

Remove stop words (such as "a," "the," and "all" [12]) because they carry little meaning. This process simplifies further text processing without sacrificing quality.

Reduce dimensionality of the data by stemming using Porter's algorithm [17]. Stemming maps several forms of a word to their common stem. For example, "receive," "received," and "receiving" are reduced to "receiv." Although "receiv" is not a dictionary word, it suffices for the purpose of classifying feedbacks as similar words are reduced to the same form.

Represent text computationally via two alternatives for representing text: unigram (bag of words) and bigram (bag of pairs of adjacent words).

Assign class labels to the textual feedback using a text classification module [20]. We evaluated two popular classification algorithms: Naïve Bayes and Support Vector Machine (SVM). We found that SVM outperforms Naïve Bayes. Therefore, we applied SVM over a combination of the unigram and bigram models.

Compute expectation profiles of the buyers using the results of the classification. For example, suppose a buyer has left three feedbacks that are assigned the class labels (1) *Item*, *Communication*; (2) *Others*; and (3) *Communication*. We disregard the *Others* label because it is outside our five main concerns. Then we aggregate the class labels from the other two feedbacks to obtain the initial vector $\langle 1.0, 2.0, 0.0, 0.0, 0.0\rangle$. We divide this vector by the number of aggregated feedbacks to normalize it. The final expectation vector is $\langle 0.5, 1.0, 0.0, 0.0, 0.0\rangle$.

2.3 Buyer-Buyer Profile Match

Our approach reflects the intuition that if two buyers have closely related expectation profiles, then each buyer is more predictive of the other's ratings of a seller. Consider a prospective buyer interested in purchasing a product offered by more then one seller. We collect the feedback and ratings given by previous buyers to the same seller. We analyze the feedback to extract the buyers' expectations.

We calculate the prospective buyer's predicted rating as the weighted sum of the previous buyers ratings for the same seller. The weight used for each previous buyer

is calculated based on the Pearson correlation between the prospective buyer's expectation profile and the previous buyer's expectation profile, as used in conventional recommender systems [5]. The rating of each buyer is then weighted by how close his expectation profile matches the prospective buyer's expectation profile.

2.4 Buyer-Seller Profile Match

In addition to the above, we use the seller's reputation profile to predict the prospective buyer's rating. Since the seller's reputation profile indicates the expectations of an average buyer of this seller, comparing them with the prospective buyer's expectations helps us determine if the prospective buyer and the seller match.

To predict a prospective buyer's experience with a particular seller, collect the feedback and ratings given by previous buyers to this seller. Analyze the feedbacks to extract the buyers' expectations (for reasons motivated earlier, consider only buyers giving the seller negative ratings). Finally, use those profiles to generate the seller's reputation profile, which represents the average expectations of the buyers for that seller. Compare the seller's reputation profile to the prospective buyer's expectation profile, to predict what the prospective buyer's rating would be.

We develop a seller's reputation profile that reflects the feedbacks *received* by the seller from previous buyers. This is the seller's negative reputation profile from the standpoint of the expectations of the previous buyers. It represents the expectations most strongly arising in the buyers' negative feedbacks for this seller.

The seller's reputation profile represents the average buyer's expectation profile as the average buyer interacts with this seller. Intuitively, if the prospective buyer's expectation profile is close to the seller's reputation profile, the prospective buyer will have similar reaction. We apply this to the negative feedbacks, from which we can construct the (negative) reputation profile of the seller and the prospective buyer.

We determine the similarity between the profiles in terms of the cosine of the angle between them [18]. Below $\otimes$ refers to the inner product of two equal-length vectors, namely, the sum of their element-wise products. Then $\cos(V_1, V_2) = \frac{V_1 \otimes V_2}{\|V_1\| \times \|V_1\|}$. In order to convert similarity into a categorical value, we check if the cosine is larger than 0.87 (which corresponds to an angle of 30 degrees or less) to determine that the profiles are in agreement. Since we are focusing on the negative profiles, if those profiles are in agreement, we conjecture the buyer is likely to give the seller a negative rating. But if the buyer's main complaints from past purchases indicate different expectation attributes than what the seller's previous buyers have complained about, the buyer and seller's expectation profiles would not agree. Consequently, the buyer would be more likely to give a positive rating.

3 Evaluation

We conduct our evaluation using eBay, because it is one of the most popular online reputation systems for e-commerce, and because we can retrieve the ratings and feedbacks left by buyers after their actual transactions on eBay. On eBay, each party involved in a transaction can leave a feedback and a rating on the other. A rating can be of one of three values: $\{-1, 0, 1\}$. The numeric ratings help ground our approach.

3.1 Dataset

We explain below some important decisions necessary for the development of our dataset from the large amount of information available through eBay. Some of these decisions are pragmatic—to make the effort tractable. And, some decisions are necessary for the theme of our experiments.

Selecting a Category. We select data pertaining to a particular sales category so that we can find enough overlap among the buyers and sellers to conduct our experiments. We choose a category based on the following criteria.

- *Common*. The category needs to be for common items. This is so we can find sufficiently many buyers and find buyers with broad characteristics, so our results are not biased by any tight community we might happen to select. For example, if we chose a niche category, then there might be well-developed communities of interest with established patterns of expectations.
- *Affordable*. The category must be affordable to allow for repeat purchases. For example, not many buyers will be using the "Automotive" category repeatedly.

Thus, we focus our research on the categories *Music CDs* and *Cell Phones*.

Selecting the Sellers. Not all sellers would have meaningful data. We conjecture this is due to the positive feedback being often quite vague and not containing sufficient useful information. Thus we have followed the following criteria in selecting the sellers:

- *Not perfect*. The seller's score should be less than 100%, and preferably in the 95% to 99% range, so there is sufficient negative feedback to analyze. We remark in passing that the average feedback on eBay is high, about 95% positive: thus we identify sellers who are about average, not those who are unusually positive.
- *Adequate feedback*. In general, it is difficult to draw strong conclusions with only sporadic data. We select sellers who have received at least 40 negative feedbacks.

Selecting the Buyers. We seek buyers for whom meaningful data is available. We need buyer data for two experiments, and we select the buyers appropriately to suit the needs of each experiment. In each experiment, the previous buyers are extracted from eBay data. These buyers would give negative ratings in order for the feedback to be meaningful. The prospective buyer is treated differently in each experiment.

- *For Enhancing Ratings*. The prospective buyers' expectation profiles are determined in such a manner as to show their impact on the seller's rating.
- *Predicting the Buyer's Rating*. The prospective buyer can be positive or negative. We withhold the buyer's rating and use it as the ground truth to evaluate the predictions of our algorithm. The main criterion in selecting these buyers is that they would have given an adequate amount of negative feedback to sellers *other* than the one under consideration. The reason for this choice is that we need buyers with an adequate track record to be able to infer their expectation profiles.

Collecting the Data. For each of the selected sellers, we collect the feedback and rating left by each buyer. We consider all ratings and feedbacks received by a seller in a particular category. For each buyer, we collect the feedback and rating associated with each previous transaction. (Typically, such transactions are with different sellers.) This leaves us with a dataset that has sellers and buyers with sufficient history about their interactions with each other and with additional parties. The following are the instructions given to the human study participants.

- Consider the transactions with negative and neutral comments for each seller.
- For each transaction, determine which expectations of the buyer were not met—these would be the ones that the buyer complained about.

From the above data, we calculate the ratio of the expectations not met with the number of negative or neutral feedbacks.

Summary of the Data. Table 1 show a quick summary of our collected data.

Table 1. Statistics regarding feedbacks analyzed

Item	Count
Sellers	1,794
Buyers	147
Number of feedbacks	2,242
Unique buyer-seller interactions	2,048
Feedbacks left by buyers for a joint seller	1,195

The first important test is whether the buyers leaving feedback for a joint seller have matching profiles. If the expectation profile is not predictive, then buyers complaining about the same seller would not share the same profile. We compare each pair of buyers for the same seller across the entire data set, to find a relatively high percentage, 54%, of profile matching between buyers complaining about a joint seller in 649 cases.

3.2 Result: Robustness of Expectations as a Well-Formed Concept

We now show that even though expectations are subjective, they are a robust concept in that humans can extract buyers' expectations from feedbacks, and do so in a reliable manner. We show this by computing the *interrater agreement* among humans regarding the expectations that can be inferred from text feedbacks.

The Kappa measure of interrater agreement captures whether agreement among raters exceeds chance levels [9]. Given $P(a)$ as the relative observed agreement among raters, and $P(e)$ as the (hypothetical) probability of chance agreement, the Kappa measure is defined as $\frac{P(a)-P(e)}{1-P(e)}$, and ranges from 0 (complete disagreement) to 1 (complete agreement). For a setting involving multiple classifications and raters, the appropriate variants of Kappa have intricate definitions and we omit them for brevity.

Three raters independently assessed buyers' expectations from various text feedbacks. We selected 361 feedbacks that have three out of the six categories selected by the raters. We obtained a high level of interrater agreement with overall Kappa of 85.5% and fixed and free marginal Kappa of 80.0% and 82.6%, respectively.

3.3 Result: Effectiveness of Negative Feedback in Indicating Expectations

We selected five sellers with high ratings (over 95%), compiled their recent 16 positive and 16 negative feedbacks, thereby forming a sample of size 160 feedbacks. We submitted these 160 feedbacks to a human rater to rate each feedback's usefulness. The rater would assign the value of 1.0 if the feedback is useful in capturing the buyer's expectation associated with the feedback; otherwise, the rater would assign the value of 0.0. For each seller, the average is computed for both the positive and the negative feedbacks. From the summary Table 2, we infer that a negative feedback is more than twice as indicative of the expectations associated with the transaction than is a positive feedback.

Table 2. Relative effectiveness of positive and negative feedback in indicating expectations

Seller	Average Rating	Positive Feedback Usefulness	Negative Feedback Usefulness
Seller 1	99.7%	40.00%	73.33%
Seller 2	99.8%	33.33%	80.00%
Seller 3	99.5%	26.67%	60.00%
Seller 4	99.6%	26.67%	66.67%
Seller 5	98.6%	26.67%	60.00%
Average	**99.6%**	**30.67%**	**68.00%**

3.4 Result: Effectiveness of Automatically Computing Expectations

As remarked above, we apply supervised machine learning using Naïve Bayes (NB) and SVM techniques. Traditional text classification is often evaluated in terms of precision, recall, and F-measure. Because our problem involves multiple labels, we report these metrics as *macro-averaged* (calculating the average of a metric for each class) and *micro-averaged* (using a global contingency table for each class) [21].

Using 500 annotated feedbacks, we apply ten-fold cross validation. That is, we train our classifier on 90% of the data and test it on the remaining 10%, using a different 10% for testing each of ten times. Table 3 shows the results from different experimental settings. In particular, SVM classification on a combination of unigram and bigram yields the best performance. So we use that setting for the subsequent evaluation.

Notice that general text classification can yield better metrics than we obtained, but our approach proves quite effective in demonstrating the power of expectations. We conjecture that our classification results can be improved by using a larger training set, a better spelling checker, and considering those domains of business services where the feedbacks given are more complete.

3.5 Result: Buyer-Buyer Profile Match

In order to apply the buyer's expectations, we need to construct the expectation profile associated with the ratings.

Specifically, we select two sellers who sell under the *Cell Phones and PDAs, Bluetooth Wireless Accessories, Headsets-Wireless* category. Both sellers have a high Positive

Table 3. Feedback classifier performance in different experimental settings. All settings use stop words removal and stemming. They vary in using unigram (U), bigram: B, Naïve Bayes (NB), and Support Vector Machine (SVM). Each value is the mean of a ten-fold cross validation.

Setting	Micro P	Micro R	Micro F	Macro P	Macro R	Macro F	Error
B+U+SVM	0.67	0.76	0.71	0.66	0.76	0.67	0.14
B+SVM	0.72	0.53	0.61	0.70	0.45	0.50	0.16
U+SVM	0.67	0.73	0.70	0.68	0.73	0.66	0.15
B+U+NB	0.59	0.79	0.68	0.61	0.66	0.58	0.18
B+NB	0.40	0.82	0.54	0.39	0.65	0.42	0.32
U+NB	0.57	0.79	0.66	0.54	0.64	0.54	0.19

Table 4. Expectation profiles for prospective buyers

	Buyer 57	Buyer 119
Item	0.11	0.48
Communication	0.67	0.28
Shipping time	0.56	0.32
Shipping charges	0.11	0.04
Service	0.00	0.00

Feedback Percentage of 97.5%. In other words, the traditional eBay reputation cannot be used to distinguish between them. We claim that our approach can help a prospective buyer distinguish between such sellers.

From the negative feedback left for each seller by the previous buyers, we first determine the expectation profile of each of the previous buyers. Let us consider two prospective buyers with expectation profiles as shown in Table 4.

Table 5. Sellers' reputation profiles computed by mining feedbacks

	Item	Communication	Shipping time	Shipping charges	Service
Seller 1606	0.42	0.17	0.34	0.08	0.0
Seller 1321	0.25	0.75	0.50	0.25	0.0

To evaluate this approach, we select two sellers, collect the feedback left for them by previous buyers. Using the approach of Section 2.2, we compute the expectation profile for each feedback. We then average the profiles for all feedback received by each seller to compute the seller's reputation profile. Table 5 shows these results, normalized based on the total number of values. It is evident that whereas both sellers have relatively close results with respect to *Item* and *Communication*, they vary widely with respect to *Shipping time*, *Shipping charges*, and *Service*.

Next we calculate the predicted ratings for each buyer for each seller. Our approach yields the predicted ratings for the two sellers for each of the prospective buyers. It shows a clear distinction between the two sellers: Seller 1321 has a lower performance for *communication*, which is an important expectation attribute for Buyer 57. Thus its

predicted rating is reduced, leading us to identify Seller 1606 as a better match than Seller 1321 for Buyer 57. Similarly, Buyer 119 would prefer Seller 1606.

3.6 Result: Buyer-Seller Profile Match

To validate the effectiveness of using expectation profiles, we select the top fifteen sellers with the most feedback (more than five negative feedbacks each). Their minimum rating is 97.2% with a mean of 98.4% and a standard deviation of 0.5 percentage points. We isolate the distinct buyers for those sellers to eliminate repeated seller-buyer interactions. We then apply the method of Section 2.2 to generate the buyers' expectation profiles and the reputation profile of each seller. Finally, we subject the resulting profiles to our buyer-seller matching. Our results show that in 73 cases out of the 116 available feedbacks, the buyer-seller profile matched, indicating a negative buyer experience. This is a hit ratio of 63% for our approach. Clearly, these 73 buyers would not have considered purchasing from the seller if our expectations-based approach were used. They interacted only because the traditional metric is not as effective in predicting outcomes in such cases.

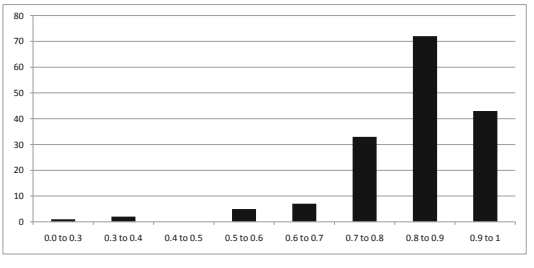

Fig. 1. Distribution of strength of matches between buyers and sellers for the top 15 sellers

Figure 1 shows that using the reputation profile of the seller and comparing it to the prospective buyer's expectation profile yields a significant advantage over using the traditional approach. Typically, the benefit of our approach is greatest when the profile match suggests a negative rating. This is because on eBay the average rating is overwhelmingly positive. For this reason, a true negative rating is highly indicative of an unsatisfied user, and an ability to predict accurately in cases of negative ratings is of extremely high value. We think of this as a major result from our study.

We next take a closer look at how a buyer may fare using the traditional approach compared to our approach. We selected a buyer (Buyer 60) who had already interacted with five sellers, apparently because the sellers had high overall ratings. Table 6 shows the profile we construct for each of these sellers in the usual way based on the feedback left by buyers other than Buyer 60.

All of the above sellers match Buyer 60's profile strongly (the cosines of the vectors are 0.87, 0.92, 0.97, 0.92, and 0.92, respectively), indicating that the buyer's expectations were not met. This result emphasizes our findings above that matching a buyer's expectations profile with a seller's (negative) expectations profile is an effective predictor of the buyer's expectations being unmet.

Table 6. Sellers' reputation profiles computed by mining feedbacks

Seller ID	Item	Communication	Shipping time	Shipping charges	Service	Other
235	0.5	0	0.5	0	0	0
805	1	1	1	0	0	0
838	0.50	0.17	0.58	0.08	0.25	0
1035	1	0	1	0	0	0
1620	1	0	1	0	0	0

4 Conclusions and Discussion

To summarize, our approach produces three important results. First, we show that incorporating expectations leads to improved predictions in ratings. These improvements arise precisely where they are the most valuable, which is when the prospective buyer would otherwise be likely to produce a negative rating. The key observation is that even if two sellers obtain similar numeric ratings, when they are viewed from a buyer's perspective, they may exhibit unique deficiencies and strengths with respect to that buyer's expectations. Capturing such variations is crucial for service selection.

Second, we show that even minimal text fragments can carry useful clues about a human's expectations that go beyond mere numeric ratings; it is possible to mine such text effectively to help bring cognitive models of trust to a new level.

Third, because the approach works at a level of abstraction, it can avoid the problem of sparsity of data, which plagues traditional approaches (and which might be the reason why content-free numeric ratings have become as popular as they have). We can predict the ratings of a buyer based on the feedbacks of other buyers even if prospective buyer has never shared a seller with them. This is important because finding adequate overlaps in the transactions of different pairs of parties can be incredibly difficult, especially as we approach the era of The Long Tail in e-commerce [2].

4.1 Threats to Validity

In general, it appears that the key elements of our approach are generalizable: most service settings involve rich notions of expectations and text feedbacks from users may often be our best path to access knowledge of such expectations. However, proving the above claim presupposes access to sufficiently large amounts of data from other settings. Such data is not readily available in vetted form with sufficient controls. It would be valuable if the research community were to develop curated datasets by collecting information from social websites regarding service interactions.

Our empirical study carries some inherent biases. First, we focus on negative ratings because our limited evaluation in Section 3.3 shows they are more useful than positive ratings. Second, the amount of data we consider is apparently fairly small in the scheme of things. We hope to scale up our approach in future studies. However, two important aspects of our study are human-intensive. One is to obtain human annotations of the ratings to judge the stability (interrater agreement) of the concept of expectations as reflected in text and the other is to use human annotations as a basis for supervised learning. Third, we have considered buyers who generally gave multiple feedbacks and

sellers who received multiple feedbacks. We expect that a data mining approach such as ours is inherently limited to such cases: without adequate data, it would not get far.

4.2 Relevant Literature

Many efforts on the theme of trust and reputation in e-commerce address challenges such as malicious ratings or a seller employing others to provide false high ratings, or to provide false low ratings of another seller in order to distort the other seller's reputation [10]. Other efforts concentrate on the propagation of trust through several second-hand sources. Few have explored what other factors can influence the rating of a user and thus, influence the final computation of trust. To our belief, not much research has considered on the study of expectations and their relationship to reputation and trust.

Singh and Maximilien [14] introduce a trust model that is centered on a shared conceptualization for QoS (ontology) and a QoS preference model that considers consumer's tradeoffs among qualities as well as relationships between qualities. Their work could be combined with the present approach by modeling the users' expectations with regard to the various qualities.

At a practical level, an interesting direction for future work is to expand the techniques for sentiment and affect analysis that we have employed. We have considered the five most common domain-independent expectations attributes, four of which are now supported by eBay. However, the space of expectations is extremely broad. We would like to expand this work to accommodate more sophisticated expectations especially those that arise in specific domains. For example, the expectations of coin collector may be quite different from those of a business-woman purchasing a printer. When we broaden the scope of expectations, the problem naturally calls for sophisticated text processing and machine learning techniques.

The notion of expectations is central to trust. Bernard Barber [3] defines trust essentially in terms of expectations—regarding general social structures as well as the technical competence and intentions to meet obligations and responsibilities. The present paper has focused on the lower end of this scale of complexity and subtlety so as to demonstrate the effectiveness of several apparently simple techniques. However, the scope of the work could be naturally expanded, and we hope that the success of the approach and its results lead to greater interest in the study of expectations.

It is also interesting to consider trust, as Castelfranchi and colleagues [7] have argued, as a form of relationship capital that can be accumulated. The present approach could feed into such work. Meeting expectations strengthens the relational capital whereas violating expectations depletes it. We observe that a lot of the cognitively well-motivated research into trust and reputation such as [6] has not had practical applications in broader computational settings. This is because of the difficulty in inferring the cognitive states of users in open settings. The methodology developed here, of inferring expectations as a form of simplified sentiment and affect analysis of text fragments, could possibly develop into a more general approach that could handle the challenges of the cognitive approaches—in settings where some clues to the user's cognitive state are available in text or other media.

4.3 Future Work

We began from a motivation based on the importance of expectations from the services science standpoint, especially as applied to business services. The e-commerce interactions that we study are business as opposed to technical services, and the user experience they offer depends more on subjective expectations than on hard quality of service data such as latency. Therefore, although they are simple, they are a useful surrogate for business services at large. However, we imagine that more complex engagements would offer additional challenges, including the involvement of more than two parties and the evolution of expectations during negotiation. The latter would go beyond the exchange of messages as in the eBay setting.

Formulating a more general model of consumer expectations for service-centric systems along with a method for computationally inferring expectations in such settings are two significant challenges. We imagine that the computational method would again rely upon techniques such as text mining, but perhaps more sophisticated than the present approach. We hope to address some of these conceptual and technical challenges in future work.

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

References

1. Abbasi, A., Chen, H., Thoms, S., Fu, T.: Affect analysis of web forums and blogs using correlation ensembles. IEEE Transactions on Knowledge and Data Engineering 20(9), 1168–1180 (2008)
2. Anderson, C.: The Long Tail: Why the Future of Business is Selling Less of More. Hyperion, New York (2008)
3. Barber, B.: Logic and Limits of Trust. Rutgers University Press, New Brunswick (1986)
4. Bhattacherjee, A.: Understanding information systems continuance: An expectation-confirmation model. MIS Quarterly 25(3), 351–370 (2001)
5. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence, pp. 43–52. American Association for Artificial Intelligence, Menlo Park (1998)
6. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In: Proceedings of the 3rd International Conference on Multiagent Systems, pp. 72–79. IEEE Computer Society Press, Los Alamitos (1998)
7. Castelfranchi, C., Falcone, R., Marzo, F.: Being trusted in a social network: Trust as relational capital. In: Stølen, K., Winsborough, W.H., Martinelli, F., Massacci, F. (eds.) iTrust 2006. LNCS, vol. 3986, pp. 19–32. Springer, Heidelberg (2006)
8. Checker, G.S.: `http://code.google.com/apis/soapsearch/reference.htm#1_3`
9. Eugenio, B.D., Glass, M.: The kappa statistic: a second look. Computational Linguistics 30(1), 95–101 (2004)

10. Kerr, R., Cohen, R.: Smart cheaters do prosper: defeating trust and reputation systems. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 993–1000. IFAAMAS, Budapest (2009)
11. Kim, D.J., Ferrin, D.L., Rao, H.R.: A study of the effect of consumer trust on consumer expectations and satisfaction: the Korean experience. In: Proceedings of the 5th international conference on Electronic commerce (ICEC), pp. 310–315. ACM Press, New York (2003)
12. Project stop words list, S.: `http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop`
13. Maglio, P.P., Spohrer, J.: Fundamentals of service science. Journal of the Academy of Marketing Science 36(1), 18–20 (2008)
14. Maximilien, E.M., Singh, M.P.: Agent-based trust model involving multiple qualities. In: Proceedings of the 4th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS), pp. 519–526. ACM Press, New York (July 2005)
15. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL), pp. 271–278. Association for Computational Linguistics, Barcelona (2004)
16. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: A conceptual model of service quality and its implications for future research. Journal of Marketing 49(4), 41–50 (Fall 1985)
17. Porter, M.F.: An algorithm for suffix stripping. Information Systems 40(3), 211–218 (2006)
18. Salton, G., McGill, M.J.: An Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
19. Standifird, S.: Reputation and e-commerce: ebay auctions and the asymmetrical impact of positive and negative ratings. Journal of Management 27(3), 279–295 (2001)
20. Williams, K.: `http://search.cpan.org/~kwilliams/ai-categorizer-0.09/`
21. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42–49. ACM, New York (1999)