"Helpfulness" in Online Communities: A Measure of Message Quality

Jahna Otterbacher
Computational Linguistics And Information Retrieval (CLAIR)
University of Michigan
Ann Arbor, MI 48109 USA
jahna@umich.edu

ABSTRACT
Online communities displaying textual postings require measures to combat information overload. One popular approach is to ask participants whether or not messages are helpful in order to then guide others to interesting content. Adopting a well-established framework for assessing data quality, we examine the nature of "helpfulness." We study consumer reviews at Amazon.com, deriving 22 measures quantifying their textual properties, authors' reputations and product characteristics. Confirmatory factor analysis reveals five underlying quality dimensions representing reviewers' reputations in the community, the topical relevancy of the reviews, the ease of understanding them, their believability and objectivity. A correlation and regression analysis confirms that these dimensions are related to the helpfulness scores assigned by community participants. However, it also uncovers a strong relationship between the chronological ordering of reviews and helpfulness, which both community participants and designers should keep in mind when using this method of social navigation.

Author Keywords
Information quality, online community, social navigation, information overload, product reviews.

ACM Classification Keywords
H.5.3. Group and Organization Interfaces: Web-based interaction.

INTRODUCTION
It has long been known that tasks involving the interpretation of text are subject to information overload, a state in which someone becomes unable to fully exploit information available to him or her [8]. For this reason, online communities in which participants post unstructured text messages face a number of challenges. Characteristically, as such a community's popularity increases, users need a means to manage the large quantity of texts, identifying and attending to those that are interesting to them. Otherwise, they likely end up leaving the community frustrated [10]. The overwhelming amount of information available is not the only challenge. Another concern is the posting of low-quality, or even false, information [2]. In fact, the quality of information available at a community is often inversely related to the size of its membership [7].

To address these problems, community designers often use social navigation, in which judgments from participants are collected and used to prioritize the messages posted [5]. The idea is to guide other users to interesting content, without having to hire moderators to screen each posting. This is the approach adopted by the community currently studied, the product review forum at Amazon.com, which has long been considered an e-commerce leader [19]. This hands-off approach is an important feature of Amazon's community, since consumers view it as being a relatively unbiased source from which to learn about others' opinions [22].

What intrigues us about Amazon is its very simple approach to social navigation. In contract to other communities in which participants rate the "interestingness" of messages on an established scale (e.g. Slashdot.com [12]), Amazon's participants are simply asked whether or not reviews are "helpful." As can be seen in Figure 1, others may then sort the reviews for a given product by the number of respective "helpful votes." As also depicted in the figure, participants have access to the profile of the reviewer.

A serious challenge for this approach has been noted in previous research. In particular, soliciting enough participation in rating content is considered to be one of the most critical issues in designing online communities [18]. According to Ghose and Ipeirotis [4], the helpful vote mechanism is not very useful for ranking reviews, since it takes time to accumulate a reasonable number of ratings. Similarly, Zhang and Varadarajan [25] suggest that at least 10 votes per review are required for the

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

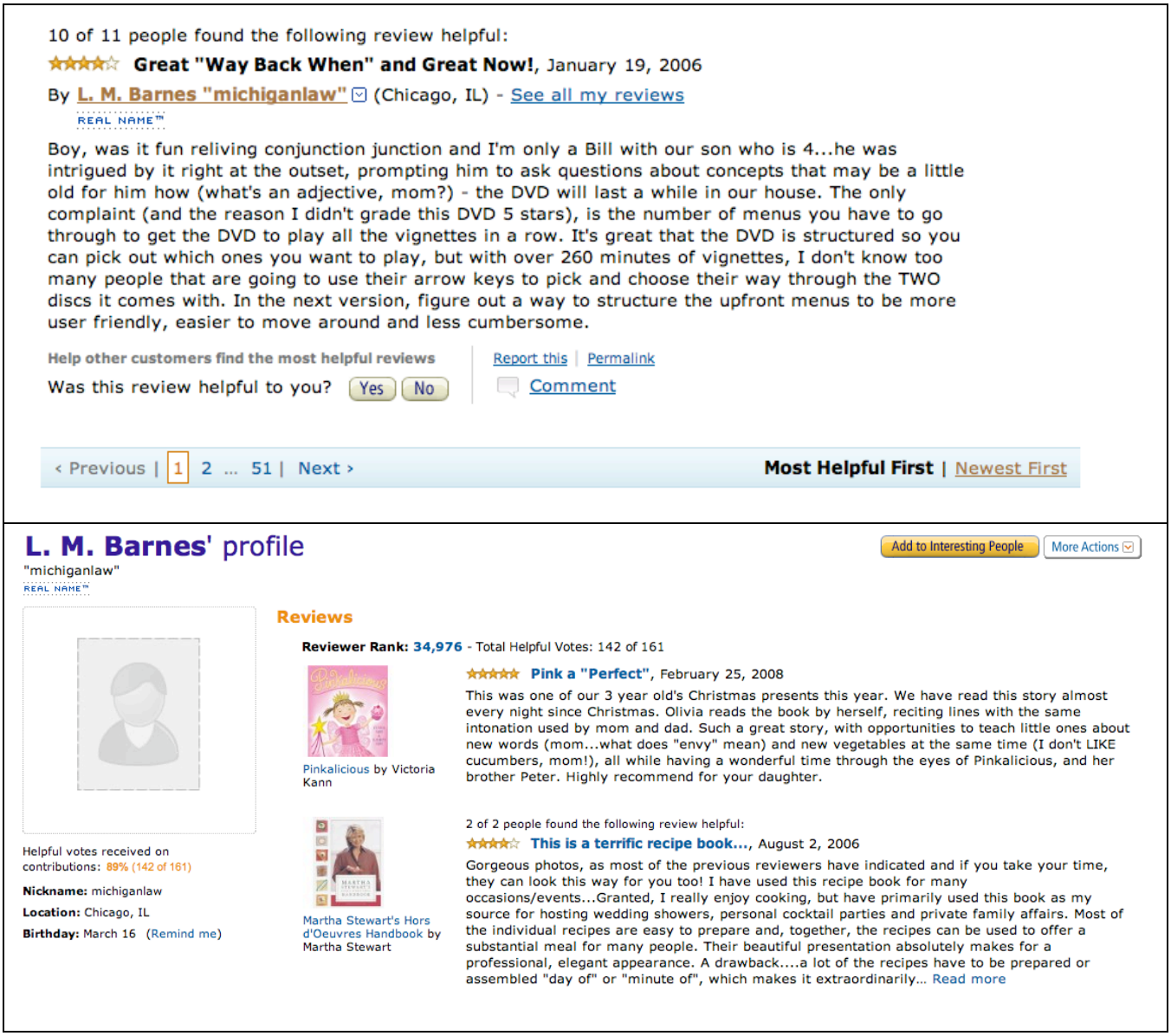

Figure 1: Example review for Schoolhouse Rock! Special 30th Anniversary Edition DVD (top) and its reviewer's profile (bottom).

ratings to be robust. Nonetheless, Amazon continues to use this scheme. In addition, several major online retailers, including JCPenney.com and BestBuy.com, host review forums that closely mimic Amazon's, also employing the "helpfulness" method of social navigation.

Given its simplicity yet apparent success, we wish to explore the nature of message "helpfulness" in the Amazon community. By providing insight as to what helpfulness reflects about textual reviews, we can help community participants better use these ratings in their search for information about a product of interest. In addition, we will discuss the implications of our findings for community designers in improving the effectiveness of this social navigation technique.

Specifically, we will address the following questions:

- What is "helpfulness" in the context of the Amazon community? Is it a measure of message quality?
- What are the important dimensions of "helpfulness"?

We conduct a study of a large set of product reviews, using a well-established framework for assessing data quality put forward by Wang and Strong [24]. In the next section, we explain this framework in detail along with the data set examined. We will also explain how we operationalize the various dimensions of quality, by quantifying 22 properties of the textual reviews and their authors as well as the products they describe. Following that, in the analysis section, we will explain the results of

a confirmatory factor analysis, from which we recovered five dimensions of message quality. We also show, using a correlation and regression analysis, that the quality dimensions are able to explain a great deal of variance in the extent of "helpfulness" among the reviews. Finally, we will conclude by examining the implications of our analysis for both information seekers and designers of online communities that use the "helpfulness" social navigation mechanism.

METHDOLOGY

Data Set

We selected 50 products at random from four Amazon categories: DVDs, Electronics, Music and Software. Our data includes all reviews for the 200 products along with their respective "helpfulness" ratings, posted on or before August 25, 2008. In addition, we captured the text of the profile pages of the respective reviewers, as well as the text from the main page of each product.

Table 1 shows the attributes of the reviews in the data set. First, we can observe the amount of textual information available to consumers. While the distribution of the number of reviews posted per product is skewed, typically, there are well over 200 available for a given product. Therefore, it is clear that techniques for managing information overload are quite essential.

	Mean	Median
# Reviews posted per product	340.3	235
Length (sentences)	8.7	6
Length (words)	146.0	88
# Total ratings per review	8.8	4
Helpfulness (# helpful votes / # total ratings)[1]	0.53	0.55

Table 1: Product review attributes.

Table 1 also displays information about the level of participation in rating posted content. The number of total ratings is skewed to the right, with the median being 4. In other words, while a few reviews receive many ratings, most receive a rather modest number. In fact, 14.6% of the reviews received no ratings at all. We will return to this issue in the analysis section. Finally, the distribution of "helpfulness" across the reviews follows an approximately normal distribution, with about half of the participants who rated a review finding it helpful.

Four characteristics of reviewers are summarized in Table 2. Amazon reviewers can earn badges, which "tell other customers something interesting" about themselves[2]. The badges might help reviewers attract attention as they are prominently displayed before the respective review, as shown in Figure 1. In our data, 58% of the reviews were displayed with a "real name" badge. However, only 3.6% of the reviews were written by a "top reviewer." From the reviewers' profile pages we gleaned additional information about their experience and reputation in the community. As can be seen, the average number of reviews contributed was highly skewed, the median being only 4. In addition, the total number of helpful votes received is skewed, with a median of 10.

	%	Mean	Median
Reviews displaying a "real name" badge	58%		
Reviews displaying a "top reviewer" badge	3.6%		
#Reviews written		64.0	4
#Helpful votes received		523.2	10

Table 2: Reviewer Attributes.

Research Framework

To study message quality in the Amazon review forum, we look to the Management Information Systems literature, where the concept of data quality[3] has been studied extensively. Wang and Strong [24] developed a framework for data quality from the end user's perspective. Conducting a large-scale survey, they uncovered four major categories of data quality, each of which is made up of several dimensions:

- Intrinsic quality: emphasizes that data have quality in their own right. Important dimensions of this attribute include believability, accuracy, objectivity and reputation.
- Contextual quality: stresses the need to consider quality with respect to the user's specific task. Its dimensions include relevancy, timeliness, completeness and quantity.
- Representational quality: has to do with the format and meaning of the data. Its key dimensions are interpretability, ease of understanding, representational consistence and concise representation.

[1] Following [4] and [25], we use this definition of helpfulness throughout the paper.

[2] http://www.amazon.com/gp/help/customer/display.html?ie=UTF8&nodeId=14279681 (accessed January 2009).

[3] While we recognize that "information" is typically interpreted as being the product of processed "data", following [16], we use these terms interchangeably in the current study.

- Accessibility: concerns whether the user has access to an information system in order to meet her information needs. Its dimensions include accessibility and access security.

Pipino and colleagues [16] noted that the framework can be used in an objective assessment of quality in particular contexts. Metrics should be developed that operationalize the quality dimensions relevant to the data set and task at hand. For example, a previous study used the framework to predict quality in news articles [23]. Textual properties such as length and the presence of key vocabulary were found to correlate to aspects of quality.

We determined that to assess quality in Amazon reviews, only the first three categories in the framework are needed. Accessibility is not relevant since participants in the community are using the same information system (i.e. the virtual community environment). Table 3 shows the quality framework developed for the current study. As can be seen, we have incorporated 9 aspects of quality across the first three categories. The third column of Table 3 describes the metrics used to operationalize the dimensions of quality. We have incorporated information from four sources: the textual properties of the reviews (e.g. length, vocabulary), metadata of the reviews (e.g. age), information from the respective reviewer's Amazon profile, and properties of the products themselves

Category	Dimensions	Metrics	Explanation / Justification
Intrinsic quality	Accuracy Objectivity	Textual similarity between the review and description on product's page. In particular, the (1) cosine, (2) bigram overlap, and (3) normalized longest common subsequence between the two texts were calculated [14].	[6] proposed that there are two types of information in reviews: objective, which is textually similar to the product description, and subjective, that differs from the description.
	Believability Reputation	(4) Product rating (on a 5-point scale) assigned by reviewer (5) Reviewer uses real name (6) Reviewer has top reviewer badge (7) Reviewer's rank in the community (8) Total reviews contributed by reviewer (9) # Helpful votes received by reviewer (10) Perplexity of textual review (11) Entropy of textual review	(4): Consumers with extreme opinions of a product are more likely to write reviews and often want to vent their frustrations [1]. (5)-(9): These attributes might be used by community members to assess reviewer reputation. (10)-(11): If we consider the distribution of words used in all reviews of a product, perplexity and entropy quantify the deviation of a review from what is expected [14].
Contextual quality	Relevancy	(12) Centroid (textual centrality) score of product review, as described in [17].	A weighted vector of words used across all reviews of a product is created. A review's centroid score quantifies the extent to which it contains words that are statistically important across reviews.
	Appropriate amount	Length of review measured as: (13) # Sentences (14) # Words	Trivially, longer texts contain more information. However, some reviews could be too long for users to read.
	Timeliness	(15) Days lapsed since the earliest review was posted about the respective product	Older reviews tend to have fewer ratings [4, 15].
Representational quality	Ease of understanding Interpretability	"Readability" measures of review: (16) Characters-to-sentences ratio (17) Words-to-sentences ratio	Texts that score high on these measures are more complex and take more effort to understand [3].

Table 3: Wang and Strong's (1996) data quality categories, dimensions and the metrics used to quantify them.

Data Treatment and Control Variables

Measurements on the 17 attributes described in Table 3 were collected, with each of the 68,393 reviews in our data treated as an observation. In addition, measurements on 5 control variables were collected for all observations:

- Review #, where reviews are sorted in reverse chronological order and #1 is the most recent review contributed to the forum. Currently, users may sort reviews either by date or perceived helpfulness.
- Product sales rank within its category. For example, the Schoolhouse Rock DVD is ranked #1 in the "Movies - Kids & Family" category. In cases where a product has multiple ranks (because it falls into multiple categories), we use the rank displayed first.
- Retail price of the product.
- Average product rating over all reviewers. At the product's main page, this is displayed prominently from 1 to 5 stars under the product name.
- Total number of reviews posted about the given product. This may tell us something about the product's popularity and thus, how excited people are to read about it and participate in rating its reviews.

Variables that deviated from a normal distribution were transformed. In particular, we used the natural log of 11 variables: the review number, product sales rank, retail price, total number of reviews posted, total number of reviews written by the reviewer, number of helpful votes collected by the reviewer, the centroid score of the review as well as its perplexity score, the age of the review and the length of the review in words and sentences.

ANALYSIS

Confirmatory Factor Analysis

The data on the 17 metrics (i.e. the 68,393-by-17 matrix) were subjected to a factor analysis, in order to determine if we could recover the underlying dimensions of quality. We note that two of the 17 variables are categorical ("real name" and "top reviewer"). While interval data are typically assumed for factor analysis, it has been noted that categorical variables can be included so long as the researcher examines the factor loadings to confirm that such variables are not "difficulty factors" or overly correlated to one another [6]. As will be seen in the analysis, these two variables were not difficulty factors and in fact, only "top reviewer" ended up being among the important dimensions.

We compared candidate models using Bentler and Bonett's normed fit index (NFI), as described in [13]. Since it is clear that our 17 measurements are not uncorrelated, we use the one-factor model as an informed baseline. Table 4 shows the NFI for three models, with their smallest eigenvalues. Generally, models with an NFI greater than 0.90 are considered acceptable, while those with an NFI above 0.95 are considered good. However, one concern with the NFI is that the more parameters that one adds, the larger the NFI. Therefore, we also considered the Kaiser criterion [11], which calls for dropping any factor with an eigenvalue under 1.0. We chose the model with five factors, since the factor with the smallest eigenvalue still accounts for 11.5% of the variance in the data. To contrast, in the six-factor model, the smallest factor only accounts for 3.8% of the variance. We note that the unrotated and the varimax solutions are very similar. Here, we present and discuss the varimax solution.

# Factors	NFI	Smallest Eigenvalue
4	0.928	2.08
5	0.966	1.23
6	0.984	0.43

Table 4: Comparison of candidate models.

The loadings of the 17 metrics onto the five factors are displayed in Table 5. To aid in interpretation, those with an absolute value of 0.5 or greater are in bold font. In addition, the proportion of the total variance in the data that is accounted for by each factor is shown. Together, the 5 factors account for 100% of the explained variance in the data. Below, each factor will be interpreted.

F1: Relevancy

The first factor concerns the topical relevancy of the reviews. As shown in Table 5, three of the 17 variables contribute significantly to this dimension of quality, namely, the length of the review (measured both in terms of the number of words and sentences) as well as the centroid score. As mentioned, in a trivial way, one expects the length of a text to be positively correlated to its information content. However, reviews that are atypically long or short can indicate that the review is of lower quality (e.g. someone "ranting" about a bad experience or something accidentally posted).

To contrast, the centroid score quantifies the extent to which a review contains a large number of words that are statistically important across all reviews about that product. For example, important words in the centroid for the Apple 30GB iPod product include "music," "battery," "player" and "iTunes." Reviews that include a relatively large number of such words are considered to be more central to the main topic expressed in the set of reviews, as compared to those containing fewer of these words.

F2: Reputation

The second factor recovered is the "reputation" dimension of intrinsic quality. The four variables that load onto this factor concern the reputation of the reviewer in the

	F1 Relevancy	F2 Reputation	F3 Representation	F4 Believability	F5 Objectivity
Proportion of variance	0.2556	0.2425	0.1972	0.1895	0.1152
	Factor Loadings				
Bigram overlap between review and textual product description	0.2071	0.0862	0.0720	-0.0773	**0.5333**
Cosine between review and textual product description	0.4089	0.1429	-0.1226	-0.0317	**0.5115**
Normalized LCS between review and textual product description	0.1262	0.0266	-0.0502	-0.0323	**0.7887**
Reviewer's rating of product	-0.1218	-0.0408	-0.0341	-0.0521	0.0614
Real name	-0.1142	-0.0831	-0.0201	0.0684	-0.0587
Top reviewer	0.1159	**0.5108**	0.0861	-0.0381	-0.0029
# Reviews written by reviewer	0.1582	**0.8932**	0.0900	-0.0430	0.0198
Reviewer's rank in community	-0.2462	**-0.6534**	-0.0725	0.0460	-0.0746
# Helpful votes reviewer received	0.2528	**0.9598**	0.1091	-0.0420	0.0359
Centroid of review	**0.7355**	0.1631	0.1685	-0.0358	-0.0591
Perplexity of review	-0.0741	-0.0394	-0.0437	**0.9953**	-0.0198
Entropy of review	-0.0707	-0.0402	-0.0434	**0.9948**	-0.0188
Age of review	-0.1053	-0.1203	-0.0283	-0.0841	-0.1408
Review length (sentences)	**0.9515**	0.2069	-0.0765	-0.0802	0.1330
Review length (words)	**0.9149**	0.2222	0.2801	-0.0941	0.0955
Characters / sentence	0.0948	0.1048	**0.9767**	-0.0520	-0.0137
Words / sentence	0.0838	0.0780	**0.9925**	-0.0373	-0.0228

Table 5: Loadings of the 17 quality metrics on the five factors.

Amazon community. Three of these variables (helpful votes received, total reviews written and "top reviewer") are positively correlated to factor 2. To contrast, the reviewer's rank is negatively correlated to this dimension, since the reviewer with rank of 1 is considered the best.

F3: Representation / Ease of Understanding

The third factor has to do with representational quality and in particular, with the ease of understanding the reviews. As seen in Table 5, only two variables are correlated to this factor: the words-to-sentences ratio of a review as well as its characters-to-sentences ratio. As explained, these metrics, which were originally proposed as means to analyze the sophistication of student essays [3], quantify how complex a text is. It can be noted that these characteristics were also used in previous studies where the goal was to predict the quality [7] and helpfulness [4] of postings in online communities.

F4: Believability

Factors 4 and 5 represent aspects of data accuracy in the quality framework. Factor 4 has to do with believability, and is correlated to two variables: the perplexity and entropy of the review. These metrics quantify how "surprising" a text is and are derived in the following way. First, the creation of a review is viewed as a sequence of randomly selected words. The random variable, X, can take on values (words) in a discrete set of symbols, which is the vocabulary used across all reviews of a particular product. In other words, the distribution of the variable X is estimated based on the entire set of reviews of the product. The entropy of a review is literally the average uncertainty of the variable X. To contrast, the perplexity quantifies the extent of "surprise" in the review, given the distribution of X [14].

In [15], we found that perplexity is useful in detecting reviews that are unusual, either because they represent unpopular opinions or because the postings are actually junk. To clarify, examples from reviews about a product in our data, Pink Floyd's "Dark Side of the Moon" album, are shown in Figure 2. On the left, we observe a posting that is junk as well as a review that is likely a minority opinion about the product. To contrast, on the right, we observe a review with low perplexity that is likely representative of the majority opinion about this product.

• High perplexity (47.9) "Mike Rotch here…just making sure, you know." • High perplexity (33.1) "This CD is the clearly the best…of the WORST! Never have I heard such filth in my life!"	• Low perplexity (5.6) "Dark Side of the Moon - quite possibly the best album of all time!"

Figure 2: Reviews with relatively high and low perplexity.

F5: Objectivity

Finally, the fifth factor represents the objectivity dimension of intrinsic data quality. The three contributing variables quantify the extent to which a review is similar to the textual description of the product, which is provided on its main page. The first metric is the longest common subsequence (LCS). It first finds the longest phrase that the two texts have in common. The length of this phrase is then normalized by the length of the review.

The next variable is the bigram overlap. Here, we are looking for the proportion of bigrams (i.e. sequences of two words) in the review, which also appear in the product description. To calculate the third metric, the cosine between the review and the description, the two texts are represented in vector space, with each element representing a unique word and its weight, the number of times the word is used in the text. The cosine between the two vectors represents the similarity between the texts.

To summarize, we recovered 5 of the 6 dimensions of quality outlined in Table 3. Reviewers' reputations, topical relevancy, the ease of understanding the reviews, and their believability and objectivity were recovered as salient factors explaining significant proportions of variance in the data. One of the dimensions, "timeliness," was not recovered. From the point of view of a user seeking information to inform a purchasing decision, "timeliness" may be too subjective to quantify. For example, it may relate to when the user reads a review, in relation to when she needs to make a decision. In any case, review age, which we used to operationalize "timeliness," is not an important variable in the analysis.

Correlation to Helpfulness

We now examine the extent to which the quality factors are related to the helpfulness of product reviews, as judged by Amazon participants. We begin by considering the correlation between the five factors and helpfulness. In addition, we examine the correlations between the control variables and helpfulness. The correlation coefficients are shown in Table 6. As can be seen, one of the factors, believability, has a negative correlation to helpfulness. This is expected, since the main variables contributing to this factor, perplexity and entropy, quantify how surprising a review is. In other words, less surprising (or more believable) reviews tend to be more helpful. The other four factors are positively correlated to helpfulness, indicating that reviews that are topically relevant, are written by reviewers with established reputations in the community, are relatively easy to read and are objective tend to be more helpful.

	r
F1: Relevancy	0.2279
F2: Reputation	0.0934
F3: Ease of understanding	0.0590
F4: Believability	-0.0302
F5: Objectivity	0.0376
ln(Review number)	-0.3354
ln(Sales rank)	-0.0275
ln(Price)	0.0914
Average rating	0.0079
ln(Total reviews)	-0.0414

Table 6: Correlations between all variables and helpfulness.

The correlations between the control variables and helpfulness are also as expected, with the exception of total reviews. We see that the correlation between review number and helpfulness is relatively strong and is negative. This means that reviews that are posted earlier to a product's forum tend to be less helpful than those posted more recently. We also see that sales rank is negatively correlated to helpfulness, such that top selling products' reviews are more helpful than those written about less popular products. In addition, a product's price and its average numerical rating are positively correlated to helpfulness. Finally, the number of total reviews posted about a product is negatively correlated to helpfulness. This variable might indicate a product's popularity with community members. Therefore, we expected it to be positively correlated to helpfulness. However, this does not appear to be the case.

We also inspect the correlations between the explanatory and control variables. We find a significant correlation between review number and total number of reviews (r = 0.6473). In addition, we find a strong negative correlation between F2 (reputation) and review number (r= -0.4219). To avoid problems with collinearity in the regression analysis, we leave out F2 and number of total reviews. We include review number as a control because of its strong correlation to helpfulness.

	β	t	Sig
F1: Relevancy	0.0690	49.7	0.00
F3: Ease of understanding	0.0165	12.1	0.00
F4: Believability	-0.0128	-9.4	0.00
F5: Objectivity	0.0226	13.9	0.00
ln(Review number)	-0.0969	-89.4	0.00
ln(Sales rank)	-0.0108	-17.2	0.00
ln(Price)	0.0321	27.7	0.00
Average rating	0.0438	15.5	0.00
Constant	0.8245	56.9	0.00

Table 7: Regression analysis using full data set.

We regressed helpfulness onto four explanatory and four control variables. The model overall is highly significant (p-value = 0.00), with an R^2 of 0.17. As seen in Table 7, all four factors are significant. We conclude that, even when controlling for the review number (or rank in chronological order), the sales rank of the product, its retail price and average rating, the four quality dimensions are significantly related to perceived helpfulness.

	β	t	Sig
F1: Relevancy	0.0575	30.9	0.00
F3: Ease of understanding	0.0125	7.6	0.00
F4: Believability	-0.0131	-7.1	0.00
F5: Objectivity	0.0219	9.6	0.00
ln(Review number)	-0.0894	-75.2	0.00
ln(Sales rank)	-0.0189	-19.1	0.00
ln(Price)	0.0309	18.8	0.00
Constant	0.962	81.8	0.00

Table 8: Regression analysis restricted to reviews with at least 10 ratings.

As mentioned, problems occur with the "helpful vote" scheme when not enough ratings are collected. In their work on predicting helpfulness scores of Amazon reviews, Zhang and Varadarajan [25] restricted their analysis to reviews with at least 10 ratings. Therefore, we investigate how our model changes if we eliminate the observations with less than 10 ratings (14,714 reviews remaining). The result is shown in Table 8. The model is again highly significant (p-value = 0.00) but with a much better R^2 of 0.40. Note that one of the controls, average product rating, was dropped since it was not significant.

Returning to the issue of how many ratings are needed before the scores are reliable, we examine how the R^2 of the model changes as a function of the number of ratings. Figure 3 plots the R^2 of the model from Table 8, restricting the data more and more. As illustrated, there is a basis to requiring 10 ratings in order to consider the scores stable, as we observe the steepest increase in R^2 at this point. Also, we can see that at 40 ratings, where the model accounts for 67% in the variance of helpfulness, we stop achieving significant increases in R^2.

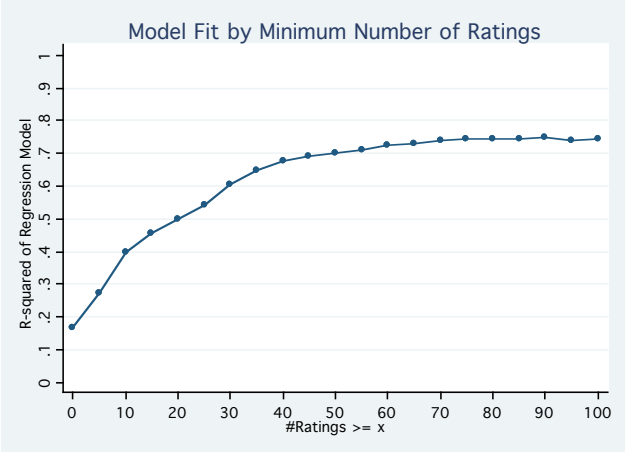

Figure 3: Model fit improves as the reviews receive additional ratings from community participants.

CONCLUSION

We now return to our original questions: What does helpfulness mean? Does it reflect the quality of information in the reviews? The answers can help community participants better understand and employ the ratings when searching for information. Also, the findings have implications for the designers of virtual communities in which the postings are textual messages.

Implications for Community Participants

When presented with many reviews, a user would like to employ helpfulness ratings to determine which postings to read. While there is indeed a relationship between review quality and helpfulness, users need to know that there are other factors that impact helpfulness. Most notably, the chronological ordering of reviews is strongly correlated to helpfulness. Early reviews systematically have lower scores as compared to more recent reviews, other factors controlled. When presented with an ordered list of

documents, as when using a search engine, users often do not look past the first page of results (e.g. [9]). Therefore, it is likely that as reviews move down the list and onto the next page, they stop collecting ratings. Because of this bias, it may be a good idea to browse older reviews, especially if one doesn't immediately find what she is looking for among the most "helpful" reviews.

Implications for Community Designers

For designers, a key challenge remains how to solicit more participation in rating content. In our sample of reviews, only 20.5% had received 10 or more ratings. In addition, given the impact of review number on helpfulness, designers need to take care in how ratings are collected. One solution might be to initially randomize the order in which reviews are presented. This might allow all reviews a chance to appear at the top of the list and to collect more ratings. Users could then have the option to sort them by helpfulness or chronologically.

Secondly, the mechanisms by which reviewers achieve reputations in the community might be reexamined. Currently, the most important contributions to reputation are the number of helpful votes received and the number of reviews written. As discussed, reputation is negatively correlated ($r = -0.42$) to review number. Thus, a reviewer will collect more helpful votes if she manages to post at a time when the forum is popular with users rather than posting a review early on. In that sense, reputation reflects popularity rather than good citizenship. Designers might want to consider if there are other characteristics valued by the community that could be incorporated into reputation. For example, participants might get recognized for being early posters or for participating in rating others' reviews.

Limitations

We considered a snapshot of 200 products at a given point in time. Like any online community, the content, participants and ratings at Amazon are continually changing. We have no reason to believe that the trends should change, given that the current method of social navigation remains the same. Of course, it is unlikely to remain the same as Amazon adds more features that participants use to judge review quality.

We also note that the metrics we used are rather simple in that they represent surface properties of the texts. Deeper methods, such as semantic or syntactic analyses of textual reviews, could certainly be added to the framework. This might allow us to capture further aspects of quality.

Unique Contributions

We set out to examine the nature of message helpfulness. Since our focus was on understanding rather than predicting helpfulness, we adopted a framework for quality assessment that is well established in the data quality community. This gives us a theoretical foundation that helps us interpret the factors influencing helpfulness.

While we used simple linear regression (SLR), our models account for a relatively high amount of variance in the independent variable. When we restricted the analysis to the reviews with at least 10 ratings, the R^2 was 0.40. For comparison, the highest R^2 reported in previous work where the goal was to predict helpfulness using SLR models was 0.16 in [25] and 0.10 in [4]. It is important to point out that [25] relied exclusively on linguistic properties to predict helpfulness while [4] used both textual information and review and product metadata. To contrast, our framework included not only properties of the reviews, their metadata and product information but also cues about the reviewers' reputations.

Directions for Future Work

Here, we briefly summarize three directions for future research. First, by going deeper into the social network at Amazon, we could incorporate additional aspects of reputation to further examine its relation to helpfulness. For example, members can add reviewers to their trusted "friends" network so the number of friends that a reviewer has, and who those friends are, could be examined. In addition, some reviewers share personal information about themselves, such as their professions and areas of expertise, which could be exploited.

Secondly, our study concerns group behavior since we examined what users collectively judge to be helpful. Also, we studied the information artifacts left at Amazon, rather than directly observing how users assess helpfulness. Therefore, it would be beneficial to conduct a related user study, in order to see if individuals would confirm that the quality dimensions are important. For example, we could conduct interviews with Amazon users, asking them to talk through a task in which they identify reviews that are helpful to them.

Finally, we are interested in comparing two approaches to organizing postings. Social navigation and automatic methods that rely on textual properties (as in [20]) have both been used to combat information overload in communities where textual messages are exchanged. We are curious as to how correlated the two approaches are (i.e. if they produce similar rankings). In addition, we would like to examine how the user experience differs.

In conclusion, we found that the "helpfulness" of reviews at Amazon is correlated to several dimensions of message quality. Despite its simple nature, the construct of "helpfulness" is able to pick up on some underlying attributes of quality, such as the topical relevancy, objectivity and readability of reviews. This finding is encouraging in that even simple means of rating online content, that do not require a lot of participants' time, can be used in a meaningful way.

ACKNOWLEDGMENTS

The author is indebted to the reviewers and the associate chairs for their very thorough comments on this work, which significantly improved both the analysis and writing. In addition, she thanks Drago Radev and Zhu Zhang for their advice and support.

REFERENCES

1. Chevalier, J.A. and Mayzlin, D. The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research 43, 3 (2006), 345-354.
2. David, S. and Pinch, T. Six degrees of reputation: the use and abuse of online review and recommendation systems. First Monday 11, 3 (2006).
3. Foltz, P.W., Laham, D., Landauer, T.K. Automated essay scoring: Applications to educational technology. In Proc. EdMedia, (1999).
4. Ghose, A. and Ipeirotis, P.G. Designing novel review ranking systems: Predicting usefulness and impact of reviews. In Proc. International Conference on Electronic Commerce, ACM Press (2007).
5. Goldberg, D., Nichols, D., Oki, B.M. and Terry, D. Using collaborative filtering to weave an information tapestry. Communications of the Association for Computing Machinery 35, 12 (1992), 61-70.
6. Gorsuch, R.L. Factor Analysis. Lawrence Erlbaum, Hillsdale, New Jersey, 1983.
7. Gu, B., Konana, P., Rajagopalan, B., and Chen, H.M. Competition among virtual communities and user valuation: the case of investing-related communities. Information Systems Research 18, 1 (2007), 68-85.
8. Hiltz, S.R. and Turoff, M. Structuring computer-mediated communication systems to avoid information overload. Communications of the Association for Computing Machinery 28, 7 (1985), 680-689.
9. Jansen, B.J., Spink, A. and Saracevic, T. Real life, real users, and real needs: a study and analysis of user queries on the web. Information Processing and Management 36 (2000), 207-227.
10. Jones, Q., Ravid, G. and Rafaeli, S. Information overload and the message dynamics of online interaction spaces: a theoretical model and empirical exploration. Information Systems Research 15, 2 (2004), 194-210.
11. Kaiser, H.F. The application of electronic computers to factor analysis. Educational and Psychological Measurement 20, (1960), 141-151.
12. Lampe, D., Johnston, E. and Resnick. P. Follow the reader: Filtering comments on Slashdot. In Proc. CHI 2007, ACM Press (2007).
13. Loehlin, J. C. Latent Variable Models. Lawrence Erlbaum Associates, 1992.
14. Manning C. D. and Schutze, H. Foundations of Statistical Natural Language Processing. Massachusetts Institute of Technology Press, 2000.
15. Otterbacher, J. Managing information in online product review communities: a comparison of two approaches. In Proc. 16th European Conference on Information Systems (ECIS 2008), Association for Information Systems (2008).
16. Pipino, L.L., Lee, Y.W., and Wang, R.Y. Data quality assessment. Communications of the Association for Computing Machinery 45, 4 (2002), 211-218.
17. Radev, D.R., Jing, H., Stys, H., and Tam, D. Centroid-based summarization of multiple documents. Information Processing and Management 40, (2004), 919-938.
18. Rashid, A.M., Ling, K., Tassone, R.D., Resnick, P., Kraut, R., and Riedl, J. Motivating participation by displaying the value of contribution. In Proc. CHI 2006, ACM Press (2006).
19. Rindova, V., Petkova, A.P. and Kotha, S. Standing out: How new firms in emerging marketing build reputation and knowledge creation. Strategic Organization 5, 31 (2007), 31-70.
20. Sack, W. Conversation map: An interface for very large-scale conversations. Journal of Management Information Systems 17, 3, (2001), 73-92.
21. Salton, G. and McGill, M.J. Introduction to Modern Information Retrieval. McGraw-Hill Inc., New York, 1986.
22. Schindler, R.M. and Bickart, B. Published word of mouth: Referable, consumer-generated information on the Internet. In: Hauvgedt, C., Machleit, K. and Yalch, R. (eds.) Online Consumer Psychology: Understanding and Influencing Behavior in the Virtual World. Lawrence Erlbaum Associates, 2005, 35-61.
23. Tang, R., Ng, K.B., Strzalkowski, T., and Kantor, P.B. Automatically predicting information quality in news documents. In Proc. Human Language Technology - North American Association for Computational Linguistics, Association for Computational Linguistics (2003).
24. Wang, R.Y. and Strong, D.M. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems 12, 4 (1996), 5-34.
25. Zhang, Z. and Varadarajan, B. Utility scoring of product reviews. In Proc. of the Conference on Information and Knowledge Management, ACM Press (2006).