Star Quality: Aggregating Reviews to Rank Products and Merchants

Mary McGlohon*†, Natalie Glance†, Zach Reiter†
mmcgloho@cs.cmu.edu, nglance@google.com, zachreiter@google.com

†Google, Inc, Pittsburgh PA * Carnegie Mellon University, Pittsburgh PA

Abstract

Given a set of reviews of products or merchants from a wide range of authors and several reviews websites, how can we measure the *true quality* of the product or merchant? How do we remove the bias of individual authors or sources? How do we compare reviews obtained from different websites, where ratings may be on different scales (1-5 stars, A/B/C, etc.)? How do we filter out unreliable reviews to use only the ones with "star quality"? Taking into account these considerations, we analyze data sets from a variety of different reviews sites (the first paper, to our knowledge, to do this). These data sets include 8 million product reviews and 1.5 million merchant reviews. We explore statistic- and heuristic-based models for estimating the true quality of a product or merchant, and compare the performance of these estimators on the task of ranking pairs of objects. We also apply the same models to the task of using Netflix ratings data to rank pairs of movies, and discover that the performance of the different models is surprisingly similar on this data set.

1. Introduction

The perceived value of reviews on the Web is uncontested: consumer surveys show that people cite product reviews as a top influencer in purchase decisions. According to Nielsen, consumer recommendations are the most credible form of advertising among 78% of survey responders (Survey 2007); and a BIGresearch survey indicates that 43.7% of consumer electronics purchases are affected by word of mouth (BIGresearch 2009). Additionally, retailers see $15 - 100\%$ greater conversion rates and decreases in product returns for items with reviews (PowerReviews ; BazaarVoice). On the other hand, a recent article in the *Wall Street Journal* publicized that the average rating for top review sites is an astoundingly positive 4.3 out of 5 stars (Fowler and Avila 2009). Given the important influence of reviews, we might then ask, how accurate *are* user review ratings on the Web? More particularly, is it possible to extract an aggregate signal from a collection of reviews that accurately reflects the relative quality of the objects under review?

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

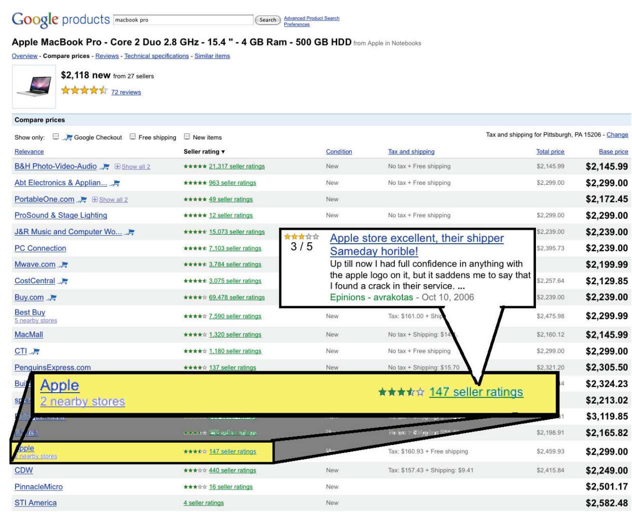


Figure 1: The list of merchants for a particular product in Google Product Search, ordered by average rating. Apple, a widely-used seller, appears toward the bottom of the list, weakened by the aggregates. (Note that the default sorting uses a different heuristic.)

The de facto aggregate review score used by almost all Web properties is the average rating per item. However, as shown by Hu et al. (Hu, Pavlou, and Zhang 2006), this is not always the best way of measuring the true quality of a product or merchant. For instance, we see in Figure 1 that for a Macbook computer, Apple is one of the lower-ranking merchants. Since Apple very often ships directly to the user, the lower rank would seem surprising. Does this suggest that average review is a poor reflection of an item's true quality? A few Web properties use "secret sauce" to calculate proprietary composite rating scores or item rankings from a collection of reviews, like alaTest and raveable.com. Do such approaches yield better results?

There are many other issues that arise in aggregating reviews data across sources and authors. Different sources have different rating scales (1-5 stars, 0-10 stars, etc.) or rating distributions (almost all positive, mostly "complaints", etc.) Authors may vary not only in their opinions of products, but in their biases, which may cloud the signal. Furthermore, reviews may be plagiarized, otherwise faked, or irrelevant (reviewing a brand instead of a product, or a product instead of a merchant) (Danescu-Niculescu-Mizil et al. 2009).

Our goal in this work is to address some of these issues and compare the performance of average rating against more sophisticated techniques for determining a composite quality score. Specifically, we detail several algorithms for generating composite scores. We then use the scores to rank reviewed objects and compare the performance of the different algorithms against a test set of ranked pairs.

We study three different data sets: product reviews and merchant reviews from Google Product Search, and Netflix movie ratings. The merchant and product review data sets are compiled from hundreds of third party review collections, including Epinions, Amazon and CNET for product reviews and Resellerratings and Shopzilla for merchant reviews. As a result of the issues associated with aggregating a wide range of sources, we initially hypothesized that average rating over aggregated reviews, even with re-scaling, would be a relatively poor predictor for ranking reviewed items. To our surprise, the average proved to be equally accurate as more sophisticated composite scores.

2. Related Work

There has been a wide range of work on consumer reviews, from that studying the motivations of consumer behavior in terms of both purchasing and reviewing, to mining product features and predicting scores.

Reviews of products can have a large impact on how well a product sells. One study (Chevalier and Mayzlin 2006) showed that there exists a *causal* relationship from book reviews to purchasing behavior. Another (Archak, Ghose, and Ipeirotis 2007) further examined the relationship between product features in reviews and sales. However, there are a number of biases reviewers tend to display. Wu and Huberman found that viewing existing reviews caused subsequent reviews written to become increasingly polarized (Wu and Huberman 2008). This may be explained, as (Talwar, Jurca, and Faltings 2007) suggested, by users having an "expectation" for a product based on prior reviews– and their rating is then impacted based on whether or not the product (in this case, a hotel room) met expectations. On the other hand, authors (Gilbert and Karahalios 2010) showed that in many cases reviewers simply echo previous reviews without adding anything new, and report interviews with reviewers on the motivations for doing so.

Our problem lies in developing a good relative measure of a product's "true quality" based on user ratings. (Hu, Pavlou, and Zhang 2006) suggest the average review score becomes an unreliable indicator of a product's "true quality". However, a better composite measure for a product's true quality yet to be determined. We therefore explore the possibility that some reviews are more "reliable" than others in terms of their ratings. Duplicate reviews (David and Pinch 2006) and opinion spam (Jindal and Liu 2008) are common, the latter study showing that it is difficult to identify untruthful reviews (plagiarism or deliberately misleading), but two other types can be detected using typical classification techniques: reviews that are irrelevant because they review the brand (not the product), and non-reviews. Authors found spam was more frequent among low-selling products, in reviews which deviate significantly from the average rating for a product, and from people who write multiple negative reviews on one brand.

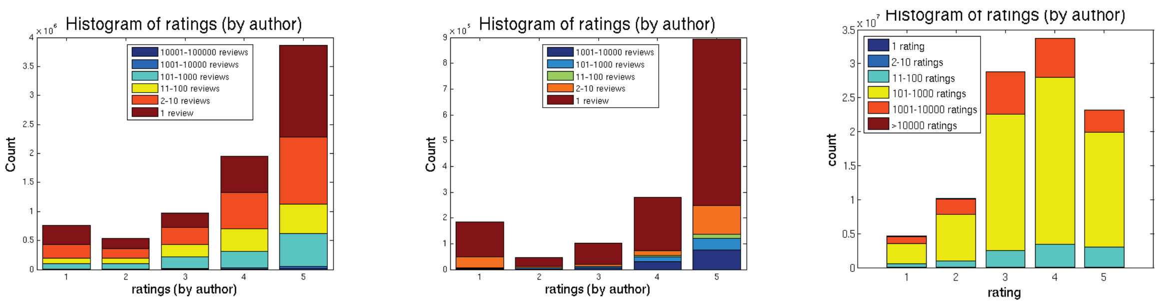
Many reviews sites also allow users to label reviews as "helpful", which has a disproportionate impact on purchase decisions (Chen, Dhanasobhon, and Smith 2008). While the problem of finding helpful reviews (finding individual reviews that help a user make a buying decision) is only tangentially related to our problem (deducing true quality of products), it lends insight in how one may want to assess a reviewer's reliability.

Even helpfulness votes, however, have user bias. Authors of one study (Otterbacher 2009) analyze a set of reviews and note a number of biases and preferential behavior in terms of helpfulness scores. Another study (Danescu-Niculescu-Mizil et al. 2009) analyzed a set of reviews from Amazon.com, along with their "helpfulness" scores, as rated by users. Some domains (amazon.us and amazon.de) were more "controversial" in reviews than others (amazon.uk, amazon.jp). Furthermore, helpfulness scores seemed to depend on the variance of the reviews for a product: for highly controversial products, reviews at the extremes were most helpful.

There have been several studies to automatically assess helpfulness or usefulness of individual reviews. RevRank (Tsur and Rappoport 2009) uses feature selection techniques to construct a "virtual core review" to represent the review space to find a set of the most helpful reviews. Other models that have been used to classify review helpfulness or to identify product features in reviews include (Ghose and Ipeirotis 2007; Hu and Liu 2004; Kim et al. 2006; Liu, Hu, and Cheng 2005; Liu et al. 2007; Zhang and Tran 2008; Zhang and Varadarajan 2006).

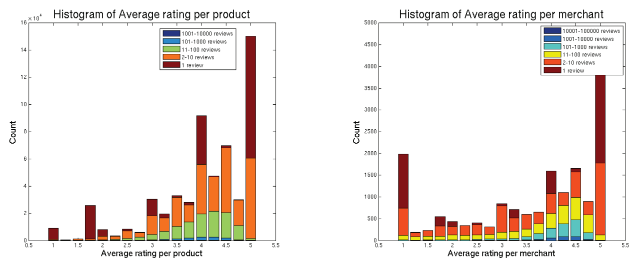
Before proceeding, it is important to distinguish between our problem of finding an overall composite ranking, and the similar problem of finding a ranking for an individual user (personalization). While the problem of aggregating ratings to obtain a composite ranking may seem like it would be an easier problem than that of personalization, in some ways it presents its own challenges. Personalization (e.g. recommender systems, such as that in the Netflix Prize (Bennett, Lanning, and Netflix 2007)) relies heavily on user data and comparisons between users, while finding an aggregate "one size fits all" ranking should still be reliable for users that are outliers or have no data available. Moreover, in the case of personalization, performance is directly observable: each user can agree or disagree with their recommendations; while in ranking such a subjective domain there is no ground truth.

Given that personalization may be a more well-defined problem, one might ask why one would want to obtain an overall ranking. There are several reasons: First, rich user data is not available for many users. Second, even if the data is available, individual preference may not be relevant under some circumstances. In a way, finding the aggregate



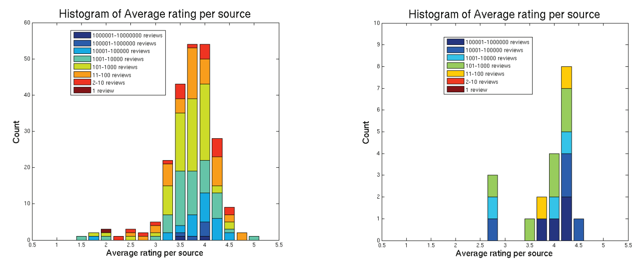
(a) Product reviews (b) Merchant reviews (c) Netflix ratings

Figure 2: Distribution of ratings in the different data sets, segmented by prolificity of authors.



(a) Product reviews (b) Merchant Reviews

Figure 3: Histogram of the average review score for different objects (segmented based on the number of reviews an object receives), for (a) product reviews, and (b) merchant reviews. In both data sets, while highly-reviewed products/merchants have an average of around 4.5, those with very few reviews tend to have average scores of 1 or 5.



(a) Product reviews (b) Merchant Reviews

Figure 4: Histogram of the average review score from different sources, for (a) product reviews, and (b) merchant reviews. Notice in (b) that while most sites have an average of a little over 4, there are a few sites with a much lower average of around 2.75.

ranking lends more insight into the *reviewed object* itself, while personalization also factors in the *individual user*.

3. Data Description

We have three data sets we will use: product reviews, merchant reviews, and Netflix movie ratings. Each review has an *author* (an Epinions.com user, for example), an *object* (product, merchant, movie), and a *rating* (1-5 stars). Product and merchant reviews, aggregated from a crawl of several reviews sites, also include the *source*. Product reviews consist of books, consumer electronics, and other retail items; and merchant reviews consist of users rating experiences with different merchants (similar to rating an Ebay transaction). Both types of reviews are between 1 and 5 stars (normalized according to different scales). Netflix is a narrower space of reviews: it consists of users rating movies from 1-5 stars. Though this data does not have multiple sources, many of the same methods used to rank products and merchants will still hold for Netflix movies.

There are many subtleties that arise in aggregating data from multiple sources. Each reviews site has its own bent: there may be sites focused largely on complaints (such as those meant to call out scams in merchants), which may translate into a high prior for a review being negative. On a finer scale, the sites may have different foci on each review– some merchant sites may ask for an overall merchant experience while others elicit per-transaction reviews. Some product sites may focus on certain product verticals, such as video games or movies. Furthermore, many review sites are retail sites, and can enforce some power over reviews, whether by following up with users or by removing inaccurate reviews at the request of a merchant. (In these cases, merchants who do not keep a close eye on their reviews may have relatively lower ratings.)

Aside from source biases, the data points themselves may be noisy. Different sites have different rating scales– some are 1-10, some are 0-5, and some even have only positive/negative labels. Furthermore, some sites simply have a larger number of reviews, which may mean more consistent data, and/or more opportunity for less reliable reviews.

To illustrate some of the inherent biases to confront in aggregating data, we next describe some empirical behavior of the data.

3.1 Product reviews

The product reviews data set, gathered from 230 sources, consists of over 8 million ratings, of 560,000 products, reviewed by 3.8 million authors. The distribution of all ratings given is shown in Figure 2(a). A overwhelming majority of ratings are positive. Bars are segmented based on the number of reviews an author has given. For instance, a majority of the 5s awarded are by authors who have only given one review.

Examining authors giving a single review, we see that approximately 37% of 5s are given by these authors, but only 25% of 4s are. Likewise, this set of single-review authors appears to give disproportionately more 1's. This makes intuitive sense– one explanation may be that fake reviews are likely to be anonymous to escape blocking (although testing this hypothesis is outside the scope of this paper).

However, there does not seem to be a significant correlation between the number of reviews received for a product and its rating– a product with a single review is equally likely to receive a five-star rating as a product with 100 reviews. This is surprising, considering that in Jindal and Liu (Jindal and Liu 2008) single-review products were often found to be spam (where spam also tends to have extreme ratings). As shown in Figure 3(a), most products have an average review of around 4.25.

Further calculations suggest there is less variance within an author rating different products than within a single product's reviews. This is unsurprising: we would expect a product's reviews to cover a variety of opinions, while an author is likely to have the same "bias" around products they rate. More prolific authors tend to have wider variance.

What is more surprising is that within a single *source* there is also little variance. As indicated in Figure 4, however, there is a great deal of variance *between* sources, with some having an average as low as 1.5 and others having an average as high as 5! This vastly different rating behavior between sites suggests we should pay attention to the source of a review when determining its quality or bias.

3.2 Merchant reviews

The merchant reviews data set is smaller, with 1.5 million ratings for 17,000 merchants. There are 1.1 million authors, from 19 sources.

The distribution of ratings is shown in Figure 2(b). The dominating number of 5's is even more prevalent here. As in product ratings, most ratings come from less active authors. As we saw with product reviews data, if an author has written a single review, then the author is disproportionately more likely to have given a 5. Also, as with product reviews, the sources vary widely in terms of rating behavior (see Figure 4(b))– for example, ReviewCentre.com has an average rating of about 2.9 while Pricegrabber.com has an average rating of around 4.5.

Looking more closely we see some different effects than what we saw in product reviews. Reviews given to a merchant may have a high variance, more so than for products. While a disproportionate number of merchants with a single review receive either a 1 or a 5 (as in product data), the average review for any given merchant hovers at around 4.5. Authors of merchant reviews had, overall, a higher average *and* a higher variance than authors of product reviews (that is, an author is more likely to use the full scale of 1-5 stars when rating merchants than s/he is when rating products), as suggested by Figure 3(b).

3.3 Netflix ratings

The Netflix data does not contain multiple sources as the previous two data sets did, but had a larger number of ratings. It consists of around 100 million user ratings (on a scale of 1-5), for 17,770 movies, by 480,189 authors. The data provided for the Netflix Prize was primarily sampled from users with many ratings. As shown in Figure 2(c), the mode of ratings is at 4, not 5 as in our other data sets. (However, among authors the few authors with a single rating, 5 is the mode.) We observe more correlation between the number of reviews for a movie and its average rating: the movies that had more ratings tended to have more positive ratings. For example, movies with over 100,000 ratings had 63% positive (4 or 5), while movies with 101-1,000 ratings had only 42% positive. This effect was not as strongly observed in the other reviews data.

Now that we have provided a brief overview of the data, we will describe our objective in using the data and how we intend to evaluate solutions.

4. Problem Statement

Formally, given a set of ratings R where each rating $r(o_i, a_j)$ is a numeric value representing author a_j's opinion of object o_i, our overall goal is to correctly rank each object o_i's "true quality" q_i, relative the other objects. A challenge to ranking is that this "true quality" is unobservable (if such a thing is even definable in such a subjective space), so we will later propose a framework for evaluation.

Each of our models will provide an estimate $\hat{q}_i$ of the quality of each object o_i, and then rank the objects in order according to that estimated "score". For example, for our baseline method estimates the quality score $\hat{q}_i$ as the average rating given by all reviewers, objects with the highest average rating will appear at the top of a ranked list. We will next detail some proposed models for estimating $\hat{q}_i$ in order to perform the task of ranking objects.

5. Proposed Models

Our proposed models fall into two main categories: statistical and reweighting. Our baseline model, average rating, falls into the first category. Reweighting models involve filtering reviews out or downgrading their influence in the composite score, considering some reviews to be more important or reliable than others.

We will use the following notation: the estimated quality by a model, for an object o_i, is $\hat{q}_i$. The set r_{i*} represents all ratings *to* a given object o_i, and the set r_{*j} represents all ratings *from* a given author a_j.

5.1 Statistical models

Average rating: This is the model we use as a baseline measure. The estimated quality of an object is the average rating it has received from all authors in the training data. Formally, $\hat{q}_i = \bar{r}_i = \frac{1}{|r_{i*}|} \sum_{j \in r_{i*}} r_{ij}$

Median rating: This is set up identically to average rating, except for the statistic used. Here the estimated quality of an object is the *median* rating the object has received from all authors in the training data.

Lower bound on normal confidence interval: Some products have more consistent ratings than others. For example, we would like to give a higher score to a product that has received 100 5-star reviews than to a product that has received a single 5-star review, even though the average rating model would give these the same score. We may also trust a product with a solid string of 4s more than one with a noisier signal. We approximate $r_i \sim N(q_i, \sigma_i^2)$– that is, a rating for a product falls in a distribution around its true quality, with some variance. We then use the *lower bound* for the quality score. More precisely, $\hat{q}_i = \bar{r}_i - z_{\alpha/2}\frac{\sigma_i}{\sqrt{|r_{i*}|}}$, where the constant $z_{\alpha/2} = 1.96$, for a 95% confidence.

Lower bound on binomial confidence interval: Such a normal approximation may not be accurate. However, we could instead simplify the star ratings into positive/negative– for instance, every rating of 4 stars or above is positive– and then take the lower bound of the confidence interval of the *percentage of positive reviews*. Also known as the Wilson Score, it is calculated in the following manner: First, obtain $\hat{p}$, the proportion of "positive" ratings for a given object o_i. We also define $n = |r_{i*}|$, the number of reviews for an object. Next, the statistic is:

$$\hat{q}_i = \hat{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2}\frac{\sqrt{[\hat{p} * (1-\hat{p}) + z_{\alpha/2}^2/4n]/n}}{(1 + z_{\alpha/2}^2/n)}$$

This measure was suggested in (Miller 2006) as a better way to aggregate ratings for products with few ratings.

Average percentile of order statistic: One issue with aggregating across several review sites is the different scales used in ratings. For example, while one site uses 1-5 stars, another may use a binary "positive/negative" label, two scales that may not easily translate. How does one maintain a faithful comparison of products, particularly when objects are not reviewed by all sources?

We would like to make a statement such as, "most of the time, object o_1 was ranked above object o_2". We can devise a method for this. We rank products *according to each site*, and calculate a score for an object based on where it occurs on each list. Specifically, we take the *average percentile* for this object over all sites and use that as $\hat{q}_i$.

This score is calculated in a few steps:

1. Aggregate reviews by source, and for each object, calculate the average rating $\bar{r_{ij}}$ from all authors j that are reviewers *from that source* (for example, the average of all ratings received for a product on Epinions).
2. Sort all objects o_i for each source, based on that average.
3. From each sorted list, assign a percentile score for a source-object pair.
4. For each object, take $\hat{q}_i$ to be the average *percentile* it receives for all its sources.

For example, suppose that an object o_i was reviewed on two different sites s_1 and s_2. We would then sort *all products* on each site according to average rating. Suppose that after doing this, o_i was in the 80th percentile on site s_1, and in the 50th percentile on site s_2. The "score" for o_i would then be 0.65.

We can use the same process with authors instead of sources to counteract author bias, but this data is more sparse in the product and merchant reviews where many authors rated few objects. However, it may be useful for the Netflix ratings where authors are comparatively prolific.

5.2 Re-weighting models

If we have some idea of which reviews are more reliable– having more "star quality"– we can decide to give more importance to these when training a model. An instance of re-weighting is *filtering*, which decides that certain reviews are likely to be misleading, and assigns these a weight 0. In essence, we are pruning the training set. We next detail models we used in this class.

Filter out anonymous reviews: Using regular expressions on the reviewer name ("A Shopper", "Anony*", etc.), remove from the training data any reviews from an apparently anonymous reviewer. Then, calculate the average rating given to an object in the remaining data. Anonymous reviews comprised between 5-10 percent of the training data in products and merchants data. This model is irrelevant for Netflix ratings, where all users were anonymized.

Filter out non-prolific authors: Sort the data by authors, and remove from the training set any review from an author with fewer than m total reviews (we used $m = 10$ in experiments).

Weighting authors according to maximum likelihood A more sophisticated re-weighting model involves deriving some "bias" for different authors and re-weighting their ratings accordingly. We propose a model based on an assumption of the distribution of ratings around "true quality". We model $r_{ij} = r(a_j, o_i)$ as a stochastic function of the "true quality" q_i of an object o_i and the "noise" of an author– some authors are more precise than others, so we say each author has a variance σ_j^2. That is, we make the following assumption:

$$r_{ij} \sim N(q_i, \sigma_j^2)$$

Based on this assumption, the maximum likelihood estimate for the parameters q and σ is:

$$argmax_{\sigma,q}\Pi_{r_{ij}}\frac{1}{\theta_j\sqrt{2\pi}}exp(\frac{(r_{ij}-q_i)^2}{2\theta_j^2})$$

To find each q_i and σ_j, we use the EM algorithm (Dempster, Laird, and Rubin 1977) to maximize the value above, iteratively updating each q and σ until convergence. The update equations are as follows: $\hat{q}_i$ is a weighted average of ratings by all authors, where each author is weighted according to the noise among their ratings.

$$\hat{q}_i = \frac{\sum_{r_{i*}}\frac{1}{\sigma_j^2} * r_{ij}}{\sum_{r_{i*}}\frac{1}{\sigma_j^2}}$$

The noise for an author is then simply the sample variance around the quality scores.

$$\hat{\sigma^2}_j = \frac{1}{|r_{*j}| - 1} \sum_{r_{*j}} (q_i - r_{ij})^2$$

One can add more parameters to the assumed distribution of ratings, assigning a bias to an author in addition to a variance term, or assigning variance to a product's quality. However, we found that in practice the more complex models did not outperform the two-parameter one, at least on the Netflix data set.

6. Evaluation

6.1 Methodology

Since there is no ground truth for the true quality of a product, merchant, or movie, deciding how to evaluate the proposed models is another interesting problem. We cannot exactly answer the question "Can we rank objects correctly, according to quality?"

Therefore, we propose an approximation– to answer a related question, which is "Given no knowledge of a user, how often can we replicate their ranking of a pair of objects?" To accomplish this, we can first rank objects based on our estimated quality scores $\hat{q}_i$. Then, we can sample from user ratings and see how reliably our ranking matches the preferences of users. Thus, for the purposes of evaluating different estimates of $\hat{q}_i$ to rank objects, we propose evaluating using a holdout method to obtain training and test sets of reviews. The steps for this are as follows:

1. For each author a_j with greater than n reviews ($n = 100$ in our experiments), pick k pairs of objects rated by the author. Each of these pairs $\{r(o_{k,1}, a_j), r(o_{k,2}, a_j)\}$ will become one data point in the test set.
2. For each pair, label it with whichever object has received a higher rating from the author.[1] *The goal of any model will be to reproduce this ranking for each pair.*
3. Reviews not used in the test set are placed in training set.

Any given model will use the training set to come up with an overall estimated quality, $\hat{q}_i$ for each object o_i and thereby an ordered ranking of objects. Then, for each pair of objects in the test set, the ranking between the pairs is compared to how those two objects are ranked in the model's list. [2]. If the model correctly reproduces the relative ranking for the pair, it counts as a hit; otherwise it is a miss. Accuracy is then determined based on the number of pairs it ranks correctly. For example, a random model would arbitrarily choose one of the two objects as the better one for each pair, and receive an accuracy score of about 50%. (It is possible to have conflicting data points in the test set, if the same pair of objects is selected from two authors with differing opinions, however that occurrence is unlikely given the sparse sample in the test set.)

[1]We require that the difference between the pair to be greater than some threshold to use it in the test set– it is more important a model distinguish between a 4-star object and a 1-star object than between a 5-star and a 4-star object.

[2]Since the goal is not personalization, the identification of the author of the pair of reviews in the test set is not used.

Method/Accuracy	Products	Merchants	Netflix
Random[4]	50%	50%	50%
Average rating	70.4%	69.3%	69.0%
Median rating	48.2%	50.2%	40.7%
Lower bound: normal	69.0%	70.2%	68.9%
Lower bound: binomial	65.1%	68.3%	69.1%
Order statistic (by sources)	69.9%	66.3%	N/A
Order statistic (by authors)	62.1%	58.5%	69.1%
Filtering anonymous	67.1%	68.7%	N/A
Filtering non-prolific authors (minimum of 10 reviews)	68.6%	38.6%	69.0%
Reweighting authors by reliability			69.4%

Figure 5: Results from running various aggregate quality metrics.

In practice, to build the test sets for each data set, we took one pair of reviews from each author with more than 100 ratings[3]. Each test set was further pruned by the threshold mentioned earlier: if the difference in ratings that the author gave the two products was less than 2 stars, the test data point was not used. This resulted in a test set size of 1423 pairs in product reviews, and 205 pairs in merchant reviews, and 13,141 pairs in Netflix– small test sets due to the selectivity of test points.

6.2 Results

We test the various models on the different data sets. Results are summarized in Figure 5. Each model was trained on data from all authors (save data points in the test set), and results are computed on the test sets from a select few authors. Accuracy is calculated as the number of *correctly ranked* pairs. Ties and unclassified pairs were considered to be misclassifications. Overall, we found that average rating performed significantly better than random, around 70% in all data sets. Surprisingly, in spite of the different domain, the different measures performed similarly on Netflix data as on the product and merchant reviews.

Some statistic-based models performed up to, but not exceeding average rating. The confidence-interval based models performed promisingly, although a common source for error were objects with a single rating: a variance of 0 made the lower bound of the confidence interval equal to that one rating. (There may be ways to combat this issue, which are worth exploring.) Order statistic-based measures tended to perform as well as average, suggesting that there were few pitfalls in this approach, but it is perhaps more complex than would be useful. Median performed poorly (worse than random) due to a vast number of ties; most ratings were one of the whole numbers.

[3]We recognize that this is a potential source of bias; however, we believe that sampling in this manner was likely to give better quality test data. One interesting future direction is to automatically choose which two rated objects, from which authors, are most meaningful to compare against each other.

Performance of re-weighting models suggested that removing data entirely is not always useful. Filtering out anonymous reviews did not have a significant effect either positively or negatively, but filtering out non-prolific authors removed a large amount of data in the merchant reviews, making the model unable to score many pairs in the test set as an object was not rated at all in the filtered training data[5]. We explored the more sophisticated measure of re-weighting "more reliable" authors as described earlier on the Netflix data. Also surprisingly, the results were nearly identical, at 69.1% accuracy.

7. Conclusion

7.1 Potential future directions

While average rating's performance matched that of a large variety of more sophisticated approaches, we have not disproven the existence of a more accurate ranking system. We next explore some potential future possibilities for searching for such a system.

In this work we chose to focus on ratings themselves, to gain insight into what the ratings mean independent of the text of reviews. However, for the purposes of building a reliable ranking system, there may be additional features one would like to explore, to leverage rich data available through online reviews.

Leveraging additional features of reviews In favor of simplicity, most of our filtering methods were based on a single heuristic such as anonymity or prolificity of authors. However, careful selection of sources, authors, and reviews; and accounting for bias, may be promising. It may be useful to do more longitudinal analyses, with an eye to the motivation of users and the content of their reviews, as the existing literature makes some interesting observations in that vein. Some features in the data that may be telling include helpfulness ratings, timestamps (including which reviews were already live), and typical vs. outlier patterns in a user's ratings.

The source of reviews may be further investigated: we may choose to re-weight the sources, as we re-weighted authors for the Netflix data. As we saw in our data description, some sources appear to have bias associated with them– a few sources tended to give merchants lower average reviews than others. Deciding how to weight reviews coming from particular sources therefore may be of use. However, re-weighting authors in the Netflix data set did not outperform average, so it is not clear that there is a simple way to do this.

Cleaning dataset for plagiarism or spam By using heuristics found in the literature, it may be useful to remove potentially spam reviews from both training and test sets. Some heuristics include reviews which are outliers, singleton reviews (per product), and people who write multiple reviews on a single brand. Also looking for repeated sentences across several reviews is a way to filter suspect reviews.

Leveraging other data sources In some cases, there may be a more authoritative source for reviews. For example, Better Business Bureau ratings may give some insight into which merchants are most reliable. Leveraging this data either directly for ranking, or indirectly for evaluating ranking measures, may be useful.

7.2 Concluding remarks

We have explored in depth three reviews data sets, including a data set of aggregated product reviews, one of aggregated merchant reviews, and one made up only of movie ratings. Our objective has been to compare different metrics for ranking objects by "true quality", given an aggregated set of ratings for that object. We have tested several statistic-based models and various forms of data-cleaning on this task, and while none thus far have been able to outperform the average rating model (which performs well, but not as well as would be desired), our analysis provides several new observations and promising directions.

Our major contributions are as follows:

- This is the first work, to our knowledge, over aggregated reviews from different sources. We observe that there are often biases of different sources and authors– different authors and review communities will often have very different behavior. We compare reviews coming from these different review sites and investigate how this may help deduce the true quality of an object rated.
- We propose several diverse models for ranking the true quality of reviewed objects.
- We build a framework for evaluating the true quality of reviewed objects and test different approaches.
- We compare performance of different models on multiple datasets, and find surprisingly similar results in terms of performance of different measures.

As we have shown, finding a consistently accurate ranking of objects based on a diverse aggregate set of reviews is not straightforward, but is rather a complex problem with many potential pitfalls in review quality, user and community bias, and in the open-ended nature of reviewing. Learning to properly navigate these challenges will help form a more complete perspective of not only online reviews themselves, but also of the consumer experience and online user behavior.

Acknowledgments

The authors would like to thank Brady Hunsaker, Charles Schafer, Kamal Nigam, Andrew Moore, and Christos Faloutsos for their insightful discussions. Mary McGlohon was partially supported by a Google Research Grant.

References

Archak, N.; Ghose, A.; and Ipeirotis, P. G. 2007. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *KDD '07: Proceedings of*

[5] Using average rating as "backup" in these cases seemed to still not produce an overall improvement over average rating alone (in cases where there was improvement, it was not statistically significant).

the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 56–65. New York, NY, USA: ACM.

BazaarVoice. http://www.bazaarvoice.com/products/interaction-suite/ratings-and-revie%ws.

Bennett, J.; Lanning, S.; and Netflix, N. 2007. The netflix prize. In *In KDD Cup and Workshop in conjunction with KDD*.

BIGresearch. 2009. Word of mouth influences most electronics, apparel purchases, according to rama survey. http://www.bigresearch.com/news/bignrf120709.htm.

Chen, P.-Y.; Dhanasobhon, S.; and Smith, M. D. 2008. All Reviews are Not Created Equal: The Disaggregate Impact of Reviews and Reviewers at Amazon.Com. *SSRN eLibrary*.

Chevalier, J., and Mayzlin, D. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* 43:345–354.

Danescu-Niculescu-Mizil, C.; Kossinets, G.; Kleinberg, J.; and Lee, L. 2009. How opinions are received by online communities: A case study on Amazon.com helpfulness votes. In *Proceedings of WWW*, 141–150.

David, S., and Pinch, T. 2006. Six degrees of reputation: The use and abuse of online review and recommendation systems. *First Monday* 11(3).

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

Fowler, G. A., and Avila, J. D. 2009. On the internet, everyone's a critic but they're not very critical. *Wall Street Journal* A1.

Ghose, A., and Ipeirotis, P. G. 2007. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*, 303–310. New York, NY, USA: ACM.

Gilbert, E., and Karahalios, K. 2010. Understanding deja reviewers. In *CSCW '10: The 2010 ACM Conference on Computer Supported Cooperative Work*. New York, NY, USA: ACM.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. New York, NY, USA: ACM.

Hu, N.; Pavlou, P. A.; and Zhang, J. 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, 324–330. New York, NY, USA: ACM.

Jindal, N., and Liu, B. 2008. Opinion spam and analysis. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, 219–230. New York, NY, USA: ACM.

Kim, S.-M.; Pantel, P.; Chklovski, T.; and Pennacchiotti, M. 2006. Automatically assessing review helpfulness. In *EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 423–430. Morristown, NJ, USA: Association for Computational Linguistics.

Liu, J.; Cao, Y.; Lin, C.-Y.; Huang, Y.; and Zhou, M. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 334–342. Poster paper.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, 342–351. New York, NY, USA: ACM.

Miller, E. 2006. How not to sort by average rating. http://www.evanmiller.org/how-not-to-sort-by-average-rating.html.

Otterbacher, J. 2009. 'helpfulness' in online communities: a measure of message quality. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, 955–964. New York, NY, USA: ACM.

PowerReviews. http://www.powerreviews.com/reviews.php.

Survey, N. G. 2007. Word of mouth the most powerful selling tool. http://nz.nielsen.com/news/Advertising_Oct07.shtml.

Talwar, A.; Jurca, R.; and Faltings, B. 2007. Understanding user behavior in online feedback reporting. In *EC '07: Proceedings of the 8th ACM conference on Electronic commerce*, 134–142. New York, NY, USA: ACM.

Tsur, O., and Rappoport, A. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews.

Wu, F., and Huberman, B. A. 2008. How public opinion forms. Technical report, Social Computing Lab, HP Labs, Palo Alto, CA 94304, USA.

Zhang, R., and Tran, T. 2008. An entropy-based model for discovering the usefulness of online product reviews. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 759–762. Washington, DC, USA: IEEE Computer Society.

Zhang, Z., and Varadarajan, B. 2006. Utility scoring of product reviews. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, 51–57. New York, NY, USA: ACM.