2008

Managing Information in Online Product Review Communities: Two Approaches

Jahna Otterbacher
University of Cyprus, jahna@ucy.ac.cy

MANAGING INFORMATION IN ONLINE PRODUCT REVIEW COMMUNITIES: A COMPARISON OF TWO APPROACHES

Otterbacher, Jahna, Department of Public and Business Administration, University of Cyprus, P.O. Box 20537, CY-1678 Nicosia, CYPRUS, jahna@ucy.ac.cy

Abstract

Virtual communities often suffer from a number of problems, including questionable information quality and information overload, which threaten their utility and stability. To address this, social filtering techniques may be used, in which users rate the postings, guiding others to the important ones. This method is contrasted to information retrieval techniques, in which intrinsic properties of texts, such as length or keywords, are used to rank them by perceived relevance to a topic. Each approach has advantages and disadvantages. Social navigation assumes that users actively rate messages, however, soliciting sufficient participation is a known challenge. Additionally, what is interesting for one user may not be for others. Currently, we compare these approaches in the context of an e-commerce product review forum at Amazon.com. We find that while a significant proportion of reviews go unrated, these reviews are typically of low quality. Interestingly, we also find that the rankings produced using reader-assigned "helpful votes" are correlated to the rankings assigned by some simple information retrieval algorithms. The conclusion is that a number of approaches for filtering product reviews could effectively be used in such online communities in order to accommodate user preferences, and thus, in reinforcing the utility of the community.

Keywords: Online communities, Information overload, Product reviews, Information retrieval.

1 INTRODUCTION

A standard component of a B2C retail website is a forum where consumers post reviews of the products offered. These forums, which are typically only minimally monitored by the site owner, stand to benefit businesses and consumers alike. For the firm, the availability of product reviews can help establish a potential customer's trust in both the product and vendor (Ba & Pavlou 2002). In addition, hosting a site where consumers can express themselves projects a customer-centric image (Wagner & Majchrzak 2007) and fosters lines of communication, thus helping to solidify long-lasting relationships (McWilliams 2000).

Many firms have introduced new functionality to their customer forums, giving consumers multiple means to participate in discussions and create content for mutual benefit (e.g., wikis, product tagging, recommendation lists). Thus, the forums have evolved into virtual communities. Frequent participants maintain personal profiles, and are often ranked as to the quality or helpfulness of their reviews and their participation in the well-being of the community. Continual improvement and expansion of such communities can enhance a business since a "sticky" community helps draw potential customers to a firm's sales site (Preece & Maloney-Krichmar 2002).

The benefits to the customer are also well established. When contemplating a purchase, consumers often turn to online reviews as an unbiased source of product information, as compared to that obtained directly from a manufacturer (Dellarocas 2003, Schindler & Bickart 2005). More explicitly, a number of studies have shown that the prices of goods and services sold online are affected by the reputations of both the products and the vendors (Senecal & Nantel 2004, Chevalier & Mayzlin 2006).

However, despite the potential value of these communities, several factors threaten their utility. Ironically, their increasing popularity can create problems of information overload (Jones et al., 2004). In other words, a potential customer, who comes to the site searching for information, may quickly become overwhelmed by the amount of text, and might leave the site frustrated. Another problem faced in customer communities is the need to evaluate the quality of information available. The fact that participation in product review communities is typically open to any user, with minimal monitoring from the sponsoring firms, is a double-edged sword. On the one hand, it means that users can search to find opinions and experiences with products that are relatively unbiased (Schindler & Bickart 2005). On the other hand, the community is subject to bogus postings or even to reviews posted by parties promoting the products for financial gain (David & Pinch 2006).

1.1 Strategies: Social navigation and information retrieval

In order to combat such problems, many communities use social navigation to help guide users to useful information on the site. In contrast to having a moderator who must screen postings, such systems use the judgments of all participants to prioritize messages (Goldberg et al., 1992). This then helps others to distinguish between high and low quality postings (Lampe et al., 2007). However, community designers face some significant challenges in implementing such systems. One of the most critical ones is the elicitation of sufficient participation from users (Rashid et al., 2006). Obviously, if few users take the time to rate postings, the system will be unreliable.

An alternative approach to guiding users to key information in a mass of textual postings is to use the intrinsic properties of the texts themselves (e.g., (Sack 2001)). For example, traditional information retrieval (IR) algorithms, which rank an input set of texts, might use the texts' length or the presence of key vocabulary, in determining their relative importance with respect to a given topic (Salton & McGill 1986). However, users might trust automatic rankings less than social navigation, and automatic mechanisms may not be able to identify "quality" information as a human reader would. In other words, it remains to be seen if a text that is relevant and important to a topic (i.e., according to an IR approach) is also one that is seen as "helpful" by the general readership of a community.

1.2 Goals of the current work

We focus on examining these two approaches to managing problems of information overload and quality control in the context of a particular community – Amazon.com. We chose Amazon because its community is well-established and popular with consumers, and because it has built a solid reputation as a model for online firms (Rindova et al., 2007). We assume that one reason consumers visit Amazon.com is to seek general information from third parties about products. This is in line with previous findings that users employ the Web for both formal (i.e., to answer a specific question about a product) and informal (i.e., to learn about a type of product) information searching (Choo et al., 2000, Kumar et al., 2005). For example, a consumer might use Amazon's reviews to answer the question "What are people saying about this product?" This problem might be described as a sense-making task in that the user wants to understand a large number of textual product reviews describing the experiences and opinions of others (Weick 1996). Therefore, we will be concerned with techniques that can help users find general information of a high quality rather than information that is relevant to a specific query (e.g., "Is the zoom on this digital camera effective?")

The paper proceeds as follows: Section 2 provides a brief overview of related research in the areas of information overload and information quality in the context of online communities; our research questions and methodology are explained in Section 3, while Section 4 presents the analysis. Finally, we follow up with conclusions as well as the implications for the management of e-commerce websites with virtual customer communities.

2 RELATED WORK

2.1 Information overload in online communities

Product reviews provide a wealth of information to consumers for which briefer summary statistics (e.g., numerical ratings) cannot substitute. From reviews, consumers can learn of the more subjective, "experience" attributes of the product (Schindler & Bickart 2005). However, tasks involving the interpretation of text are susceptible to problems of information overload, thus users require a means to filter such information (Hiltz & Turoff 1985). More concretely, if a potential customer visits a community in order to investigate others' experiences with a particular item and finds hundreds of reviews, she requires a means to identify those that are most useful.

Information overload has consequences not only for individual users but also for community sustainability. There is a fine line between reaching a critical mass, such that there is enough activity to attract interest (Hiltz & Turoff 1978), and reaching a state of overload, in which users cannot utilize or process the available information (Rogers & Agarwala-Rogers 1975). To this end, the judgments of other users might be used as indicators of message quality or usefulness as in a collaborative filtering system (Goldberg et al., 1992). This eliminates the need for a site moderator and allows the general readership to contribute to the community's well-being. For example, while reading product reviews at Amazon.com, users are encouraged to "help other customers find the most helpful reviews" by answering "yes" or "no" to the question "Was this review helpful to you?" Once this information is collected other users may then sort the reviews by their "helpful votes" in order to aid them in deciding which texts to read.

Alternatively, properties of the texts themselves might be used to predict their utility. For example, Zhang and Varadarajan (2006) suggested that highly-rated reviews stand out due to the manner in which they are written. Therefore, another means to sort postings would be to use intrinsic properties to predict texts' relative utility. If it were shown that such ranking mechanisms are correlated to those assigned by humans, this approach would offer some important advantages, in particular because IR techniques do not depend upon sufficient user participation.

2.2 Information quality in online communities

Consumers, who rely on product review communities as a source of unbiased information, are also faced with evaluating information quality. Known threats include bogus reviews, texts copied from product to product and even postings by manufacturers. For instance, David and Pinch (2006) investigated book reviews at Amazon.com and found that hundreds were copied from other reviews. They also noted that paid editors, whose ulterior motive is to convince users to purchase specific books, post reviews. Similarly, Mayzlin (2006) found that firms post online reviews in an effort to promote their own products. In analyzing eBay's reputation system, Resnick and Zeckhauser (2002) reported that users are not always honest and frequently exhibit "Pollyanna" behaviour, in which they write comments similar to those already posted. These problems illustrate the need for tools that can help users sort through the plethora of reviews about a given product.

2.3 Current strategies at Amazon.com

The strategy used by Amazon for guiding users to quality reviews, is a system in which reviews are assigned binary "helpful votes." A list of reviews may be sorted in a number of ways: by helpful votes (i.e., "most helpful first"), by date (i.e., "newest first"), or by users' ratings of the product. Therefore, aspects of social navigation are used at Amazon, however, intrinsic textual properties of the reviews are not currently used. Users may also search the reviews for keywords. However, since we are concerned with users searching general information, we have not examined this facility. In the following section, we describe our dataset of product reviews as well as our research questions, which examine this system from the point of view of controlling information quality and overload on the Amazon site.

3 METHODOLOGY AND RESEARCH QUESTIONS

3.1 Data collection

We collected 34,343 reviews and their respective user-assigned ratings, by considering the top 50 selling products (on the 1st of March, 2007) across four product categories for a total of 200 products. Table 1 shows the properties of the dataset. It should be noted that the data collection from the Amazon website and the calculation of the properties of each textual review were done automatically using the Perl programming language, as to avoid the introduction of human biases and errors.

Category	# Reviews	Reviews per Product	Length (words)	Age (days)	Time Span
DVDs	7,894	157.9 / 106	213 / 131	164	154
Electronics	8,966	179.3 / 103	134 / 90	274	341.5
Music	14,738	294.8 / 180.5	157 / 102	133	167.5
Software	2,745	54.9 / 45	151 / 105	109	206.5
All	34,343	171.7 / 85.5	163 / 104	175	200

Table 1 . Dataset: mean/median number of reviews per product and review length, median age (days from first posting) and time span of reviews (oldest to newest review).

3.2 Text analysis metrics

We will examine properties of product reviews that might be correlated to several points of interest (i.e., whether a review has been rated by users). Therefore, we need to introduce some metrics for text analysis, which will be used and discussed in the remainder of the paper.

3.2.1 Centroid score

The centroid score quantifies the salience of a text with respect to a set of related texts. More concretely, given a collection of reviews about a particular product, it quantifies how well a review represents the "gist" of the information conveyed in the overall set. A review with a relatively high centroid score is one that should represent "what people are saying about the product." The implementation of the centroid used is adapted from that introduced by Radev and colleagues (2004).

First, the centroid for the entire set of reviews is created. This is done by representing the set of texts as a weighted vector of all the words used over all reviews. Each word (element) in the vector is weighted corresponding to its TF*IDF value. TF refers to the term frequency of the word in the reviews while IDF, or inverse document frequency, measures the proportion of all documents in a large collection of documents that contain the word (Salton & Buckley 1988). This procedure assigns a high weight to content words and a lower weight to more commonly used words that do not express as much topical information. Next, to score a given product review, we find the TF*IDF value for each word used in the review, and sum them up. The result is the review's "centroid score." A review will have a high centroid score if it contains a relatively large number of content words that are expressed throughout the collection of reviews about the product. To illustrate, Figure 1 lists some of the key content words for an electronics product from the dataset, the Apple 30GB iPod.

iPod, music, battery, player, great, love, apple, iTunes, like, don't, bought, screen, movies, thing, songs, buy, really, very, videos, sound, nice, mp3, easy, accessories, PC, awesome

Figure 1. Top content words for the Apple 30GB iPod.

3.2.2 Entropy and perplexity

Perplexity and entropy can be used to measure how unique or "surprising" a review is. The creation of a review is viewed as a random event. In particular, the text is seen as a sequence of randomly selected words. The random variable X can take on values (words) in a discrete set of symbols (vocabulary across the texts), and its probability distribution function, $p(x) = P(X = x)$, $x \in X$, is estimated based on the set of all other reviews of the product. The entropy of the review in question is then defined as the average uncertainty of the random variable X (Manning & Schutze 2000):

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Perplexity is related to entropy in a very simple way:

$$Perplexity(X) = 2^{H(X)}$$

but it is often preferred to entropy since it has a more intuitive interpretation. A perplexity of *k* refers to the extent of "surprise" involved in guessing between *k* choices of equal probability. Therefore, textual product reviews with lower perplexity are less surprising as compared to the remaining reviews while those with higher perplexity are more unexpected, given what the other reviews say.

3.3 Research questions

In order to appreciate the user's task, we will first examine the quantity of information that is typically available for a product of interest and how much this varies across products. We will also see how many of the reviews for a given product are actually unique and how helpful users find them to be. After that, we will focus on two key research questions, as described below.

- *Participation in rating reviews*: As previously explained, the system for rating the reviews breaks down when an insufficient number of community members participates. Therefore, an important question is how many reviews go unrated. It would also be useful to know which reviews are likely to go unrated. It may be the case that some low-quality reviews go unattended to by readers, in which case an unrated review should be treated as an unhelpful one.

- *Ranking reviews by perceived helpfulness:* Finally, we wish to characterize which reviews are perceived as being helpful by Amazon community participants. Having collected a large set of reviews and their respective helpful votes, we wish to examine the properties of highly rated, as well as low rated, reviews. We will consider the following hypotheses:

  H1: The helpfulness of a review is correlated to its *age*.

  H2: The helpfulness of a review is correlated to its *length*.

  H3: The helpfulness of a review is correlated to its *textual centrality*.

  H4: The helpfulness of a review is correlated to its *novelty or uniqueness*.

Finally, once we establish correlations between characteristics of the reviews and their perceived helpfulness, we will propose algorithms that rank them automatically using these characteristics. We will then compare the rankings of these mechanisms to the rankings of the reviews based on the user-assigned helpful votes, in investigating the final hypothesis:

H5: There are ranking mechanisms based on the intrinsic properties of textual reviews that are correlated to the "helpfulness" judgments of users.

4 ANALYSIS

4.1 The Amazon environment: How much information?

As shown in Table 1, when researching a popular product, a consumer typically encounters around 85 textual reviews, however, this number varies by product. For example, the median number of reviews for music products is 180.5, while that of software products is only 45. In addition, within a given category there is variance as in Table 1, one observes that the distribution of reviews posted is skewed to the right. This is particularly true of music products. For instance, some classic albums have an unusually large number of reviews (e.g., Pink Floyd's Dark Side of the Moon has over 1,000 reviews). Finally, reviews are around 100 words in length, with products typically having some very long reviews, as within all product categories the distribution is again skewed to the right.

From the information retrieval literature, it is known that when presented with a ranked list of items (e.g., documents found by a search engine), users are more likely to select those displayed on the first page of results as opposed to those on later pages and that the higher a text appears on the list, the more likely it is to be viewed, irrespective of how helpful it actually is (Joachims et al., 2005, Jansen et al., 2000). While the number of reviews that are displayed per page at Amazon varies, it is obvious that for a typical product there will be multiple pages of reviews. In addition, research on online communities has found that there are indeed negative externalities associated with large numbers of postings, which can burden users with excess information processing (Butler 2001, Gu et al., 2007). Therefore, users clearly need a means to identify the most helpful reviews.

4.2 Participation in rating reviews

Over all reviews, the distribution of "total votes" received is skewed (median of 5, mean of 10.8). In addition, 4,131 reviews (12.0%) were completely unrated. Given that on average, a product has 85 reviews available, this means that at least 10 of them have no "total votes" and the user will have no way to gauge their importance. Table 2 compares the properties of reviews broken out by number of total votes. A general trend is that the ignored reviews are older, shorter, and less central to the main topic. Since the median number of total votes is 5, we tested the differences between the group of reviews with less than 5 votes, and those with at least 5 votes. They differ with respect to age, centroid score and length[1] , all with a p-value of 0. In addition, the group with less than 5 votes has significantly higher mean perplexity (p-value of 0.04).

	Total Votes = 0	**Total Votes < 5**	**Total Votes >= 5**	**Total Votes > 10**	**Total Votes >5 Helpfulness = 0**
Age	294	251	124	99	192
Centroid	0.0721	0.0896	0.135	0.1452	0.0601
Length	62	79	135	145	58
Rating	5	5	5	4	3
Perplexity	58.2	29.0	18.9	19.1	20.4
Entropy	4.2	4.2	4.2	4.2	4.3

Table 2: Properties of reviews by total number of votes assigned by users: median age (days), centroid, length and consumer product rating (from 1 to 5); mean perplexity and entropy.

When users sort by "helpful votes," the unrated reviews appear with the low quality reviews. Therefore, the next logical question is whether or not this is appropriate. The last column of Table 2 shows the properties of reviews that have been rated by at least 5 users and are known to be unhelpful. Indeed, the unrated and the unhelpful reviews share several properties – in particular, relatively low centroid scores, and short lengths. Therefore, unrated reviews are also likely of a low quality.

4.3 Helpful reviews

Figure 2 depicts the distribution of "helpfulness" of the reviews[2] among those with at least 5 total votes (17,818 reviews). The mean and median helpfulness are 0.6 (i.e., 60% of readers find the review helpful). Obviously, since the reviews vary significantly with respect to their perceived helpfulness, it would be useful to know which reviews tend to be helpful. A comparison of the features of the reviews, by their level of helpfulness (broken out by quartiles), is presented in Table 3. The median review age, centroid score, length and customer product rating (from 1 to 5) are shown as well as the mean perplexity and entropy. Statistically significant differences between each group of reviews and

[1] According to the Mann-Whitney non-parametric rank-sum test (Siegel, 1956).
[2] Proportion of "helpful votes" to "total votes" assigned.

the next group are indicated with asterisks[3]. Note that while some of the significant differences (particularly for entropy) appear small, this is due to the large sample size and small standard deviations.

	Helpfulness < Q1 (0.239)	**Q1 (0.239) ≤ Helpfulness < Q2 (0.6)**	**Q2 (0.6) ≤ Helpfulness < Q3 (0.973)**	**Helpfulness ≥ Q3 (0.973)**
N	2,509	5,196	7,442	2,671
Age (days)	165	83	123	223
Centroid	0.744***	0.1141***	0.1660***	0.1755
Length	81***	115***	162	161
Rating	3***	4***	5	5
Perplexity	19.2*	19.1***	18.9***	18.6
Entropy	4.19**	4.22***	4.20**	4.18

Table 3: Comparison of reviews with at least 5 total votes.

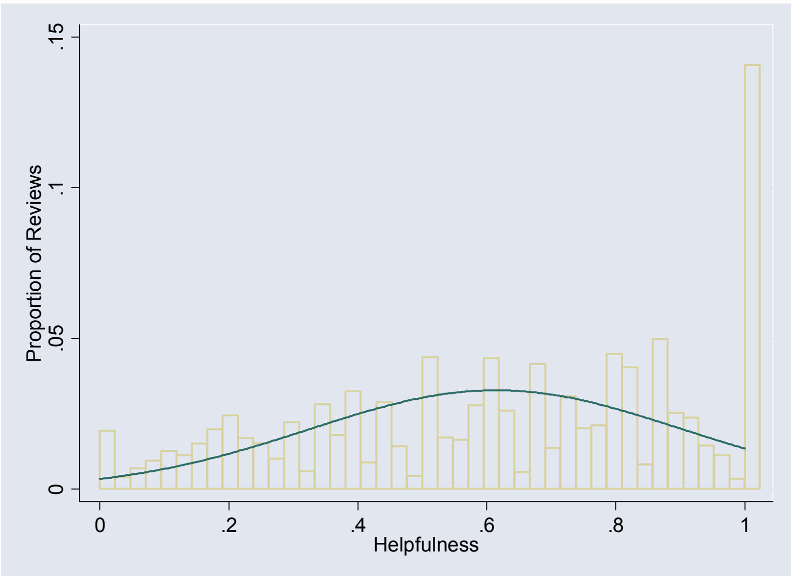

Figure 2: Proportion of votes claiming review was helpful for reviews with at least 5 total votes.

Given that a review has received at least 5 votes, there does not appear to be a strong correlation between its helpfulness and its age. However, reviews need to be of a certain length in order to be helpful. In addition, the centrality of a review is positively correlated to its helpfulness. In other words, if a review uses words that are statistically important to the set of all product reviews (i.e., has a high centroid score) and is not lexically surprising (i.e., low perplexity), it is likely to be more helpful. Finally, the valance of a review, given by the consumer's rating of the product on a scale of 1 to 5, is also positively correlated to the quality of the review. This may tell us something about how users employ the reviews. For instance, if they tend to search for reviews that reinforce their desire to

[3] Difference with next group is significant at α of 0.01 (***). Difference is significant at α of 0.05 (**). Difference is significant at α of 0.10 (*).

purchase the product, this would lead them to find the more positive reviews to be helpful, as compared to more negative reviews.

Having established these correlations, we move on to comparing an automatic review-ranking (IR) approach versus the ranking of reviews by helpfulness (social navigation approach). To do this, we performed the following experiment:

- For the set of 17,818 reviews with at least 5 total votes, we produced five automatic rankings. We also produced the ranking based on the proportion of helpful votes (descending order).
- The five automatic ranking mechanisms were: by centroid score (descending order), length (descending), perplexity (ascending), the sum of centroid score and length (descending), and the product of the centroid score and length (descending).
- We compared each of the automatically-produced rankings with the manual ranking using Kendall's rank correlation coefficient (Kendall's Tau) as described in (Abdi 2007). A Tau of 1 indicates perfect agreement whereas 0 indicates that the rankings are no better than chance.

Figure 3 shows the distribution of Kendall's Tau across the 200 products when their reviews are ranked by centroid score. This simple procedure results in a Tau that is better than chance 94% of the time. More detailed results can be seen in Table 4. The rankings produced using the centroid score and length have better agreement with the user-assigned rankings as compared to those produced from the perplexity of reviews. For example, when we rank the reviews by their centroid scores, we obtain a statistically significant correlation (at $\alpha=0.05$) to the user-assigned rankings for 136 out of 200 products in the dataset, compared to only 47 out of 200 products when we use perplexity.

While our goal was not to search for the optimal review-ranking algorithm, we tried combining the information gleaned from the centroid scores and lengths in two simple ways: by using their sums and their products. As can be seen in Table 4, neither of these procedures helped us to improve the correlation of the automatic rankings to the user-assigned rankings. Therefore, it remains to be seen if a more sophisticated scoring function could improve the automatic rankings.

	Centroid	**Length**	**Perplexity**	**Centroid+Length**	**Centroid*Length**
Better than chance (Tau >0)	188 (94%)	183 (92%)	156 (78%)	184 (92%)	185 (93%)
Significant at $\alpha=0.10$	149 (75%)	151 (76%)	64 (32%)	150 (75%)	151 (76%)
Significant at $\alpha=0.05$	136 (68%)	136 (68%)	47 (24%)	137 (69%)	136 (68%)

Table 4: Significance of Kendall's Tau rank correlations between rankings based on intrinsic features versus those assigned by users (proportion helpful).

5 DISCUSSION AND CONCLUSIONS

In order to be truly useful for consumers, online product communities need to provide users with a means to navigate through the available textual information. Currently, we examined two approaches for addressing the problems of information overload and quality control in the context of the Amazon.com community: social navigation (i.e., ranking reviews by "helpful" votes) and automatic mechanisms based on the intrinsic properties of the reviews. We found that although a substantial proportion of reviews go unrated by readers, these are most likely of a low quality and thus, it is appropriate that they are displayed with "unhelpful" reviews.

We found correlations between the rankings of textual reviews produced by simple IR techniques and those produced by the user-assigned helpful votes. In particular, the lexical centrality of a given review and its length are correlated to the proportion of helpful votes received. This finding is interesting for a number of reasons. First, as mentioned in the introduction, the "helpfulness" of a review is a rather subjective property. What one reader at Amazon finds helpful will not be helpful for all. Nonetheless, this subjective characteristic of reviews is correlated to their textual properties.

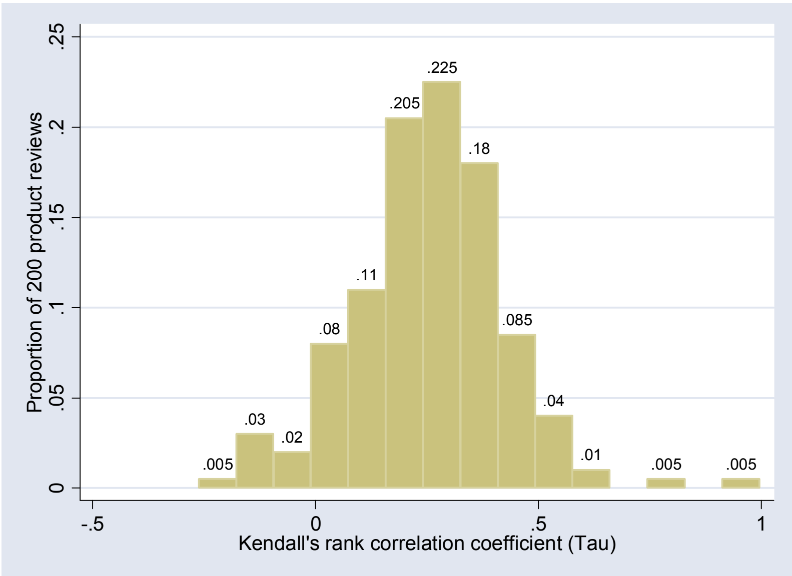

Figure 3: Distribution of Kendall's Tau across the 200 products for centroid-based rankings.

This finding suggests that a wider range of review rankings might be offered to readers at an online review community. For example, users could opt to sort by "representative reviews," "reviews that others found helpful," date posted or "uniqueness of reviews." Such an approach would provide users with tools for searching through a large number of texts, and at the same time allow them to sort texts by their own personal preferences and ideas about what "helpfulness" is (Hiltz & Turoff 1985).

While previous work considered how to predict the usefulness of product reviews based on their textual properties (e.g., (Zhang & Varadarajan 2006)), our work takes a more user-oriented perspective. We have assumed that the ultimate goal is to display the reviews as a ranked list, so that the user can quickly find useful information. Therefore, we have proposed text analysis metrics that compare a given review to others, rather than describe its properties in isolation (e.g., the centroid and perplexity relative to the set of all reviews). In addition, we have proposed an evaluation framework using Kendall's Tau, which allows us to directly measure the effectiveness of the various metrics at ranking the reviews.

5.1 Limitations

To our knowledge, this is the first work that attempts to compare the social navigation approach and the information retrieval approach to managing information overload in an online community. At the same time, its limitations should be mentioned. Firstly, as explained earlier in the paper, we studied top-selling products at Amazon, since they are more likely to have large numbers of reviews as compared to less popular items. Therefore, it is important to note that the salient properties of the reviews of less popular products may differ from those currently examined.

Also, as explained in the analysis, the text analysis metrics we examined are quite basic. Future work should consider the use of more sophisticated metrics as well as scoring mechanisms of the review ranking procedures. Additionally, our analysis and findings are obviously restricted to information management in communities involving asynchronous communication between participants. The techniques examined cannot help the members of communities that must read and respond to messages quickly. Finally, as already explained, we have assumed throughout the current work that one way that consumers use Amazon.com is to find information relevant to a given product. Therefore, the goal was to present the user with a ranked set of textual reviews about the product. Such an approach requires that the texts are already clustered by topic (i.e., product), therefore, in order to be applied to other types of online communities, clustering of messages by topic would also have to be performed.

5.2 Questions for future research

In conclusion, several new research questions have arisen from our work and should be examined in future endeavours:

- What are the preferences of users of online product communities with respect to displaying textual product reviews? Can we profile users and then offer them a customized version of review rankings? Do they tend to sort on similar attributes or is it entirely idiosyncratic (Lampe et al., 2007)?
- What is the nature of "helpfulness"? How does helpfulness differ from the notion of "relevance," which is typically used in information retrieval for document ranking (Mizzaro 1997)?

References

Abdi, H. (2007). The Kendall rank correlation coefficient. In: Salkind, N. (ed.) Encyclopedia of Measurement and Statistics. Thousand Oaks, California: Sage Publishers.

Ba, S. and Pavlou, P. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behaviour. MIS Quarterly, 26(3).

Butler, B. S. (2001). Membership size, communication activity, and sustainability: a resource-based model of online social structures. Information Systems Research, 12(4), 346-362.

Chevalier, J.A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345-354.

Choo, C.W., Detlor, B., and Turnbull, D. (2000). Information seeking on the Web: an integrated model of browsing and searching. First Monday, 5(2).

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

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. Management Science, 49(10), 1407-1424.

Goldberg, D., Nichols, D., Oki, B.M. and Terry, D. (1992). Using collaborative filtering to weave an information tapestry. Communications of the Association for Computing Machinery, 35(12), 61-70.

Gu, B., Konana, P., Rajagopalan, B., and Chen H.M. (2007). Competition among virtual communities and user valuation: the case of investing-related communities. Information Systems Research, 18(1), 68-85.

Hiltz, S.R. and Turoff, M. (1978). The Network Nation: Human Communication via Computer. Addison-Wesley Publishing Company, London.

Hiltz, S.R. and Turoff, M. (1985). Structuing computer-mediated communication systems to avoid information overload. Communications of the Association for Computing Machinery, 28(7), 680-689.

Jansen, B., Spink, A. and Saracevic, T. (2000). Real life, real users, and read needs: a study and analysis of user queries on the web. Information Processing and Management, 36(2), 207-227.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (SIGIR 2005), Salvador, Brazil.

Jones, Q., Ravid, G. and Rafaeli, S. (2004). Information overload and the message dynamics of online interaction spaces: a theoretical model and empirical exploration. Information Systems Research, 15(2), 194-210.

Kumar, N., Lang, K.R., and Peng, Q. (2005). Consumer search behavior in online shopping environments. In Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS-38), January.

Lampe, C., Johnston, E. and Resnick, P. (2007). Follow the reader: Filtering comments on Slashdot. Proceedings of the Association for Computing Machinery Conference on Human Factors in Computing Systems (ACM CHI '07), San Jose, California.

Manning, C.D. and Schutze, H. (2000). Foundations of Statistical Natural Language Processing. Massachusetts Institute of Technology Press.

Mayzlin, D. (2006). Promotional chat on the Internet. Marketing Science, 25(2), 155-163.

McWilliams, G. (2000). Building stronger brands through online communities. Sloan Management Review, 41(3), 43-54.

Mizzaro, S. (1997). Relevance: the whole history. Journal of the American Society for Information Science, 48(9), 810-832.

Preece, J. and Maloney-Krichmar, D. (2003). Online communities. In: Jacko, J. and Sears, A. (eds.) Handbook of Human-Computer Interaction, Lawrence Erlbaum Associates Inc. Publishers, New Jersey, 596-620.

Radev, D.R., Jing, H., Stys, H., and Tam, D., (2004). Centroid-based summarization of multiple documents. Information Processing and Management, 40, 919-938.

Rashid, A.M., Ling, K., Tassone, R.D., Resnick, P., Kraut, R., and Riedl, J. (2006). Motivating participation by displaying the value of contribution. Proceedings of the Association for Computing Machinery Conference on Computer Human Interaction (CHI 2006), Montreal, Canada.

Resnick., P. and Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In: Baye, M. R. (ed.) The Economics of the Internet and e-Commerce. Volume 11 of Advances in Applied Econometrics, Elsevier Science, Amsterdam.

Rindova, V., Petkova, A.P. and Kotha, S. (2007). Standing out: How new firms in emerging markets build reputation and knowledge creation. Strategic Organization, 5(31), 31-70.

Rogers, E.M. and Agarwala-Rogers, R., (1975). Organizational communication. In: Hanneman, G.L., and McEwen, W.J. (eds.) Communication Behavior, Addision Wesley, Reading, Massachusetts, 218-236.

Sack, W. (2001). Conversation map: An interface for very large-scale conversations. Journal of Management Information Systems, 17(3), 73-92.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513-523.

Salton, G. and McGill, M.J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York.

Schindler, R.M. and Bickart, B. (2005). 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, 35-61.

Senecal, S. and Nantel, J. (2004). The influence of online product recommendations on consumers' online choices, Journal of Retailing, 80, 159-169.

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw Hill.

Wagner, C. and Majchrzak, A., (2007). Enabling customer-centricity using wikis and the wiki way. Journal of Management Information Systems, 23(3), 17-43.

Weick, C. (1996). Sensemaking in organizations. Sage Publications, Newbury Park, California.

Zhang, Z. and Varadarajan, B. (2006). Utility scoring of product reviews. In Proceedings of the Conference on Information and Knowledge Management (CIKM'06), Arlington, Virginia.