Targeted Solicitation of Product Reviews

Nhat X.T. Le*, Ryan Rivas†, James M. Flegal‡ and Vagelis Hristidis§
*†§*Department of Computer Science & Engineering,* ‡*Department of Statistics*
University of California, Riverside
{*nle020,†rriva002,‡jflegal}@ucr.edu, §vagelis@cs.ucr.edu

***Abstract*—Customer reviews have become an essential resource when people search for goods or services on the Internet. Previous work has shown that reducing a product's uncertainty is critical to its purchase decision. Thus, reviews are more effective when they reduce a product's uncertainty. Existing e-commerce platforms typically ask users to write free-form text reviews, which are sometimes augmented by a small set of predefined questions, e.g., "rate the product description's accuracy from 1 to 5."**

In this paper, we argue that this "passive" style of review solicitation is suboptimal in achieving low-uncertainty "review profiles" for products. Its key drawback is that some product aspects receive a very large number of reviews while other aspects do not have enough reviews to draw confident conclusions. Therefore, we hypothesize that we can achieve lower-uncertainty review profiles by carefully selecting which aspects users are asked to rate.

To test this hypothesis, we propose various techniques to dynamically select which aspects to ask users to rate given the current review profile of a product. We use Bayesian principles to define reasonable review profile uncertainty measures; specifically, we apply Bayesian inference to measure an aspect's rating variance. We compare our proposed aspect selection techniques to several baselines on several review profile uncertainty measures. Experimental results on two real-world datasets show that our methods lead to better review profile uncertainty compared to aspect selection baselines and traditional passive review solicitations.

***Index Terms*—review solicitation; customer reviews; review analysis; sentiment analysis;**

I. Introduction

It is well-known in marketing and management sciences that product uncertainty, the buyer's difficulty in evaluating product characteristics, plays a crucial role in customer shopping decisions [1]. Dimoka et al. [1] found that high product uncertainty has stronger negative effects than high seller uncertainty. Kim and Krishnan [2] noted that consumers are unlikely to buy expensive products (defined as higher than $50) online if there is a high degree of product uncertainty, even if they have a lot of online shopping experience. Therefore, e-commerce companies have sought to mitigate product uncertainty in several ways, such as providing detailed descriptions, including multimedia and virtual reality tools, and most notably soliciting customer reviews.

In this paper we focus on how to maximize the effect of reviews to the decrease in product uncertainty. Khare et al. [3] found that reviews' volume and the level of consensus have a fundamental impact on consumer judgment. Hence, an effective review solicitation strategy must account for both these factors.

TABLE I
A REVIEW PROFILE; NUMBERS ARE RATING COUNTS.

	⋆	⋆⋆	⋆⋆⋆	⋆⋆⋆⋆	⋆⋆⋆⋆⋆
Screen	0	0	0	5	31
Battery	26	10	3	0	0
Speed	0	1	2	0	3

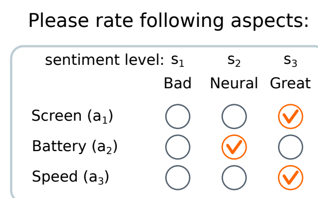


Fig. 1. Ask a customer about a smartphone

Existing e-commerce platforms typically ask users to write a free-text review. These reviews can then be analyzed by feature and sentiment extraction methods (Section II) to estimate the overall opinion of reviewers for each aspect (feature) of a product. Other websites provide a static (predefined) set of aspects for the user to rate, typically with a score from 1 to 5. For example, "How clean was your room?" or "How would you rate the reliability of the car?"

A key drawback of existing review solicitation methods is that some aspects receive too many ratings, which is especially wasteful if reviewers generally agree with each other. For example, consider product "smartphone" with aspects "screen," "battery," "design" and so on. Hundreds of reviewers may rate the "screen" as 5-stars. Conversely, a more controversial aspect, e.g. the "speed," may only receive a few ratings. This leads to a review profile with high uncertainty, as users typically try to compare various products across several aspects (features) by using past users' reviews.

An example of a smartphone's review profile is presented in Table I, which intuitively shows that screen has high rating with high confidence, battery has low rating with high confidence, and speed has high rating with low confidence. A key question that this paper studies is: *given the current reviews profile of a product, is it better to let users write a free-text review (and then extract the aspects and opinion using existing methods [4], [5], [6], [7], [8], [9]), or to ask the user to rate a small number of carefully selected aspects as in Figure 1?* A second question is: *how should this small set of aspects be selected, given the current review profile?*

In this paper, we study these two questions in a principled manner by first considering a Bayesian statistical model to estimate the probability distribution of each aspect's rating and

978-1-5386-6783-5/19/$31.00 ©2019 IEEE
DOI 10.1109/ICSC.2019.00061

then dynamically selecting aspects whose estimated ratings have the highest posterior variance. Intuitively, this method solicits reviews for the aspects that have few reviews or have diverse opinions. This means subsequent users may be asked to rate different aspects of the product.

We understand that reviews' uncertainty may also be affected by other factors like spam reviews [10], [11], or the helpfulness of the text of the reviews [12], [13], [14]. These are important factors, orthogonal to our focus, and outside the scope of this paper.

To design and compare various aspect selection methods, we must first come up with a reasonable definition for review profile uncertainty, as no such standard measure exists in the literature. In our method, we estimate a review profile uncertainty by the expected rating variance of each aspect, which we model based on a well-accepted Bayesian inference model. This model is consistent with the aforementioned points that reviews' volume and consensus are the key factors in consumer's evaluation of a product, as a high number of reviews or high review agreement reduce a rating's posterior variance. To avoid comparing various review solicitation methods based solely on the variance of the aspects, which may favor our proposed methods, we also consider other uncertainty measures based on the confidence interval (from a frequentist statistician's point of view, in contrast to our Bayesian measure), and the number of aspects whose confidence is above a threshold.

We next extend our methods to account for dependencies among a product's aspects. For example, if "screen" and "contrast" are two correlated aspects and there are many and in-agreement reviews for "screen," it may be wasteful to solicit reviews for "contrast." For this, we consider a dependency-aware Bayesian inference model to estimate the correlation of two aspects. Then, we generalize the previous definition of expected rating variance to infer an aspect's variance from others if they are highly correlated.

We compared our methods on two real datasets: Amazon reviews with annotated aspect ratings introduced by Bing Liu, et al. [4], [15] and crawled automobile reviews from edmunds.com. We first compare our method to the passive text-based solicitation method, which is simulated by picking top aspects based on the order in which they appear in reviews. Users' answers are reproduced using the actual aspects' sentiments extracted from their free-text reviews. In another group of experiments, we compare our method to various baselines that also select set of aspects to solicit users. In these cases, we utilize random generators to generate sentiments as the answers. Specifically, we consider three uncertainty measures: rating variance (as introduced in our model), rating confidence interval length, and ratio of highly confident aspects (independent of our model). Our contributions are summarized as follows:

- We define the problem of dynamically selecting aspects to solicit targeted reviews to reduce uncertainty (Section III-A).
- We propose a principled method to select the best aspects to query based on canonical Bayesian inference (Sections III-B and III-C).
- We propose a preliminary extension of our aspect selection method that considers aspects' correlation (Section IV).
- We conducted detailed experiments (Section V) on two real-world datasets, which show that our methods lead to superior review profiles compared to passive text-based solicitation and other aspect selection methods. We published our code and used datasets on our supporting web page [16].

The remainder of the paper is organized as follows: we discuss the related work in Section II, and the conclusions and future work in Section VI.

II. Related Work

Commercial Reviewing Web Sites: Most sites solicit free-text reviews, along with an "overall rating" typically expressed with 1 to 5 stars. Other web sites have a small, predefined set of questions that they ask reviewers; for instance, vitals.com, which is a doctor reviewing site, asks users to assign a score to "bedside manner" and "courteous staff." The only web site that we found that has a dynamic set of questions is tripadvisor.com, which asks users to rate different hotel aspects (e.g., "service," "location" and "sleep quality") for different hotels. However, we have no knowledge of how these aspects are selected as this is a proprietary system.

Dynamic Questionnaires: USHER [17] is a system for form-based survey design that aims to improve the quality of collected data. USHER uses a probabilistic model on the form questions, learned from previous form submissions, to adapt the form layout (question ordering) dynamically to emphasize the most important questions, or re-ask questions that may have been answered incorrectly. A key difference is that in USHER the goal is to collect information about all the questions from each user, whereas our goal is to collect enough (and reliable) information for each product aspect. For this, we analyze our aspect ratings' certainty, which is not the case in USHER.

Multi-armed Bandit Problem [18] is one of the fundamental problems in Artificial Intelligence. In its simple form, a gambler presented with a row of slot machine must decide her playing strategy, i.e. which machine to play next given the sequence of past plays, to maximize her reward. The key property of this problem is that rewards of successive plays on a machine i are independent and identically follow a distribution of an unknown expected value R_i. In our problem, the reward is the decrease in the uncertainty of each aspect, where these uncertainties may be dependent to each other (Section IV). Another difference is that in the multi-armed Bandit problem, the gambler is guaranteed the highest reward in the long run if she found the machine with the highest expected reward value, then played on that machine only. In our case, there is no aspect that will forever produce highest expected uncertainty drop when we keep getting more rating of this aspect.

Reviews Analysis There has been much work on analyzing text reviews. These works generally have two phases. First, they extract aspects (features) like "zoom," and second, they estimate the sentiment associated with each aspect using its surrounding context. *These works are complementary to our work, as they facilitate converting text reviews to structured review profiles, which can then be processed by our algorithm to select which aspects to elicit in future reviews.*

Aspect Extraction: The most common approaches to extract aspects from product reviews are based on keyword statistics and syntactic rules. Existing works [4], [5] use association rule mining to find frequent aspects, and then filter out meaningless or redundant ones using predefined syntactic dependency-based rules. After that, these frequent aspects and opinion words are utilized to discover more infrequent aspects using another set of rules. Another technique [6] decides if an aspect candidate is actually an aspect by checking the *Point-wise Mutual Information* (PMI) score between it and its product class using their Web search engine hit counts. Another approach, adopted by Jakob and Gurevych [7], models this task as an information extraction problem and applies Conditional Random Field techniques to extract aspects. Topic modelling has also been used for this problem, as in Titov and McDonald [8], who discover global and local aspects; and Mukherjee and Liu [9], who extract and categorize aspects given some seeds.

Sentiment Analysis: This problem has been investigated extensively, and has been comprehensively surveyed by Liu and Zhang [19]. Traditional methods [20] focus on creating a comprehensive, good dictionary of opinion words that are looked up when analyzing text reviews. Other authors such as Turney [21] exploit syntactic patterns to detect opinion phrases containing adjectives or adverbs. A supervised learning algorithm was first introduced to classify movie reviews as positive or negative based on vectors of reviews using the Bag-of-Words model [22]. In this model, authors experimented with Naive Bayesian and SVM classifiers that offer accurate results. Recently, the use of deep neural networks and representation learning have improved the performance of this task significantly [23], [24], [25], [26], [27]. For instance, Le, et al. [27] use an unsupervised neural network model to learn reviews' representational vectors that are later fed to a standard supervised classifier for sentiment analysis.

III. Modeling a Product's Review Profile and Aspect Selection Algorithm

A. Problem Definitions

An online product (or service) has a set of aspects (also referred as attributes or features in other papers) denoted as $a_1, a_2, \ldots, a_m$. Each aspect receives ratings from l sentiment (star) levels $s_1, s_2, \ldots, s_l$.

The review profile of a product is a summary of the aspect ratings, as exemplified in Table I. To model the quality of the review profile, we define the *review profile's statistical summary (RPSS)* as a set of tuples:

$$< a_h, r^{a_h}, cert^{a_h} > \qquad \text{with } h = 1, \ldots, m \qquad (1)$$

where r^{a_h} is the expected rating of a_h and $cert^{a_h}$ is the certainty level of r^{a_h} estimation, which are discussed in Section III-B. We also call $uncert^{a_h}$ as the uncertainty level inversely proportional to $cert^{a_h}$ (i.e. $uncert^{a_h} = 1/cert^{a_h}$). A particular aspect a_h gets $n_i^{a_h}$ votes for sentiment s_i ($i = 1, 2, \ldots, l$), and in total n^{a_h} ratings ($n^{a_h} = \sum_i n_i$).

This paper studies the problem of selecting the *top-k Uncertain Aspects (k-UA)*: *Given current users rating history:* $< a_h, n_i^{a_h} >$ $(h = 1, \ldots, m; i = 1, \ldots, l)$*, which are the* k *aspects to ask the next reviewer to rate in order to optimize the review profile?* In Section IV we extend this problem definition to consider aspect rating correlations, by accounting for the co-occurrences of aspect ratings within reviews.

Note that the top-k aspects are recomputed for each new reviewer. The computational cost is negligible, so even for high throughput of reviews, the algorithm can dynamically update the top-k aspects. The product's aspects can either be explicitly listed at the reviewing web site, or may be extracted automatically from a text review. In the former case, the reviewer selects a number of stars for each aspect, and in the latter case the sentiment is estimated automatically. These methods are discussed in detail in Section II.

In the following sections, we will present our Bayesian approach to model an aspect's certainty level in a RPSS, and then propose our algorithm for the *k-UA* problem.

B. Bayesian Inference Model

Our model focuses on measuring aspect a_h's uncertainty level $uncert^{a_h}$ of its expected rating r^{a_h}. In this section, to simplify the notation we ignore the superscript a_h in $r^{a_h}, uncert^{a_h}$ and $n_i^{a_h}$. Let $p = (p_1, p_2, \ldots, p_l)$ be a random vector representing the probabilities (degree of belief) that users rate the aspect with $s_1, s_2, \ldots, s_l$ stars, respectively. We follow a typical Bayesian inference for categorical data [28] to account for this probability vector. In particular, each sentiment level s_i is a category that users' ratings fall in.

Suppose that the prior distribution of $p = (p_1, p_2, \ldots, p_l)$ is a Dirichlet distribution of order $l \geq 2$ with parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_l)$, $\alpha_i > 0, \forall i$: $g(p) = \frac{1}{B(\alpha)} \prod_{i=1}^{l} p_i^{\alpha_i - 1}$, where $B(\alpha)$ is the Beta function. It is common to consider the uniform case as the prior: $\alpha_i = 1$ $(\forall i)$ since the likelihood will dominate the prior over time.

Also assume that the likelihood $f(n|p)$ of observed data $n = (n_1, \ldots, n_l)$ (sentiment counts) is a multinomial distribution: $f(n|p) = \frac{N!}{n_1! \ldots n_l!} \prod_{i=1}^{l} p_i^{n_i}$, where $N = \sum_{i=1}^{l} n_i$ is the total number of sentiment counts. Hence we have the posterior $h(p|n)$:

$$h(p|n) \propto f(n|p)g(p) = \frac{N!}{n_1! \ldots n_l!} \times \frac{1}{B(\alpha)} \times \prod_{i=1}^{l} p_i^{n_i + \alpha_i - 1}$$

Let $\beta_i = n_i + \alpha_i$, $\beta_0 = \sum_i \beta_i = N + \sum_i \alpha_i$. Then the posterior $h(p|n)$ is also a Dirichlet distribution with parameter $(n_1 + \alpha_1, \ldots, n_l + \alpha_l)$, or $(\beta_1, \ldots, \beta_l)$ with mean, variance, and covariance, respectively:

$$E[p_i|n] = \frac{n_i + \alpha_i}{\Sigma_{i=1}^{l}(n_i + \alpha_i)} = \frac{\beta_i}{\beta_0}$$

$$Var[p_i|n] = \frac{\beta_i(\beta_0 - \beta_i)}{\beta_0^2(\beta_0 + 1)} \quad (2)$$

$$Cov[p_i, p_j|n] = \frac{-\beta_i \beta_j}{\beta_0^2(\beta_0 + 1)} \qquad \text{for } i \neq j \quad (3)$$

The aspect's expected rating is $r = \sum_i s_i p_i$, and hence

$$E[r|n] = E[\sum_i s_i p_i|n] = \sum_i s_i E[p_i|n] = \sum_i s_i \frac{\beta_i}{\beta_0}$$

$$\begin{aligned} Var[r|n] &= Var[\sum_i s_i p_i|n] \\ &= \sum_i s_i^2 Var(p_i|n) + \sum_{i \neq j} s_i s_j Cov(p_i, p_j|n) \\ &= \frac{1}{\beta_0^2(\beta_0+1)}[\sum_i s_i^2 \beta_i \beta_0 - \sum_i \sum_j s_i s_j \beta_i \beta_j] \quad (4) \end{aligned}$$

Since $Var[r|n]$ reflects the fluctuation of an aspect's rating around its expected value, $Var[r|n]$ can be interpreted as the uncertainty measurement of our estimation of the aspect's rating, i.e. $uncert = Var[r|n]$. $Var[r|n]$ also has an intuitive property that it is roughly inversely proportional to the number of votes N (via β_0 in the denominator of Equation (4)). If an aspect has a very high uncertainty value, i.e. $Var[r|n]$, it means that we are not ready to make a conclusive estimation of its rating. Also note that, asking a controversial aspect still alleviates its variance slowly even if its new ratings are truly polarized. In the common practice, a uniform prior is used in this Bayesian inference, thus $\alpha_i = 1$. As a result, $\beta_i = n_i + 1$ and $\beta_0 = N + l$. Note that in our experiments we also consider alternative measures of uncertainty when comparing the proposed algorithms.

C. Aspect Selection Algorithm

Algorithm 1 Highest variance pick

Input: previous vote counts $n_1, \ldots, n_l$ of aspects, number k
Output: k aspects

```
1: procedure PICK_HIGHEST_VARIANCE
2:     for all i in 1 ... l do
3:         α_i = 1                      ▷ uniform prior for every aspect
4:     for all aspect a do
5:         for all i in 1 ... l do
6:             β_i^a = n_i^a + α_i       ▷ posterior parameters
7:         β_0^a = Σ_{i=1}^{l} β_i^a
8:         Calculate Var[r^a|n^a] using Equation (4)
       return top k aspects with highest Var[r^a|n^a]
```

We present our solution to the *k-UA* problem in Algorithm 1. In particular, Lines 2-3 set up common uniform prior parameters, while lines 5-7 compute posterior parameters for every aspect. We finally calculate rating variance of all aspects in line 8, then output the top k aspects with highest variances (i.e., degree of uncertainty).

Note that $Var[r|n]$ can be computed faster using vectorized version of Equation (4). Specifically, $Var[r|n]$ is the variance of a linear combination of column vector s and random vector p, so $Var[r|n] = s^T \Sigma s$, where Σ is the covariance matrix built up using Equations (2) and (3) that can be vectorized as well.

TABLE II
TOY EXAMPLE OF 4 ASPECTS WITH COUNTS OF 1, 2 OR 3 STARS RESPECTIVELY.

	Weight	Cost	Battery	Design
Star count	0, 5, 28	4, 9, 20	11, 11, 11	1, 3, 7

TABLE III
RPSS OF TABLE II, WHERE $uncert$=VARIANCE.

	Weight	Cost	Battery	Design
Expected Rating	2.78	2.44	2	2.43
Variance	0.006	0.014	0.018	0.035

D. Toy Example

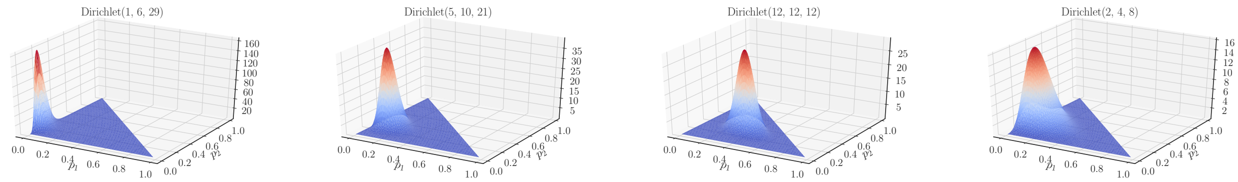
To explain the intuition of our model, consider a toy example where we are looking at a smartphone with four aspects: weight, cost, battery and design. Each aspect can be rated with 1, 2 or 3 stars (i.e., bad, neutral or good). The previous ratings of these aspects are presented in Table II. The question is which aspects we should ask users about to improve this smartphone's RPSS? Following the previous model, we can model aspects' expected rating as Dirichlet posteriors that are demonstrated in Figure 2 and the RPSS in Table III. We then use Algorithm 1 to calculate each aspect's rating variance and order them to select the k most uncertain aspects. In this case, the algorithm will pick aspect "Design" first, then "Battery," "Cost" and finally "Weight." Design is a clear choice since it has far fewer ratings to estimate its rating with high confidence. The other three aspects have the same number of ratings but Battery has more diverse opinions, so it is selected next. Weight is picked last because of its high rating count and very skewed rating distribution.

IV. EXTEND TO ACCOUNT FOR CORRELATION BETWEEN ASPECTS

Section III provides a framework to model the uncertainty level of aspect ratings, where aspect ratings are assumed to be independent of each other. However, in practice aspects are often correlated. For example, screen and brightness, or design and easy-to-use are similar to each other, and often receive similar rating. Intuitively, if one of two highly correlated aspects (e.g., "screen") has high rating certainty, then it is less important to solicit more ratings for the other aspect (e.g., "brightness"). Next, we first show how to estimate the correlation between the ratings of two aspects, and then show how this can be used to define a correlation-aware version of the uncertainty score of each aspect (recall that the aspect selection algorithm selects the k aspects with the highest uncertainty).

TABLE IV
COUNTING WHEN TWO ASPECTS WERE RATED TOGETHER BY AN USER.

	Design–1	Design–2	Design–3	
Cost–1	3 (n_{11}, p_{11})	2 (n_{12}, p_{12})	0 (n_{13}, p_{13})	5 (n_1^c)
Cost–2	1 (n_{21}, p_{21})	5 (n_{22}, p_{22})	2 (n_{23}, p_{23})	8 (n_2^c)
Cost–3	1 (n_{31}, p_{31})	4 (n_{32}, p_{32})	7 (n_{33}, p_{33})	12 (n_3^c)
	5 (n_1^d)	11 (n_2^d)	9 (n_3^d)	



(a) Posterior of Weight (b) Posterior of Cost (c) Posterior of Battery (d) Posterior of Design

Fig. 2. Toy example: posteriors of aspects' rating

To estimate the correlation of two aspects, we propose to look at their ratings at the same time. In particular, we count the number of times that two aspects were rated together in the same review. For instance, in Table IV we consider two aspects (cost and design) in a three-star system. Using similar notation as before, n_{ij} and p_{ij} are, respectively, the number of reviews and the probability users rate aspect "cost" i stars and "design" j stars at the same time. Also, $n_i^c = \sum_j n_{ij}$ and $n_i^d = \sum_j n_{ji}$ (cost and design are shortened as "c" and "d" in this clear context). We focus our interest on these two aspects' rating correlation $Cor(r^c, r^d|n)$ before generalizing to any aspect pairs. First note

$$p_i^c = \sum_j p_{ij}, \qquad p_t^s = \sum_q p_{qt} \tag{5}$$

There are l sentiment levels $s_1, \ldots, s_l$, so

$$r^c = \sum_j s_i p_i^c = \sum_i \sum_j s_i p_{ij} = \sum_i \sum_j s_i p_{ij} \tag{6}$$

$$r^d = \sum_t s_t p_t^d = \sum_t \sum_q s_t p_{qt} = \sum_t \sum_q s_t p_{qt} \tag{7}$$

Following our Bayesian approach as in Section III, we model probabilities $p_{11}, \ldots, p_{ll}$ by a Dirichlet posterior of parameters $(n_{11} + \alpha_{11}, \ldots, n_{ll} + \alpha_{ll})$. Denote $\gamma_{ij} = n_{ij} + \alpha_{ij}(i, j = 1, \ldots, l)$ and $\gamma_0 = \sum_{i,j} \gamma_{ij}$. We get variance $Var[p_{ij}]$ and co-variance $Cov(p_{ij}, p_{qt})$ in similar forms as Equation (2), (3).

$$Var[p_{ij}|n] = \frac{\gamma_{ij}(\gamma_0 - \gamma_{ij})}{\gamma_0^2(\gamma_0 + 1)} \tag{8}$$

$$Cov[p_{ij}, p_{qt}|n] = \frac{-\gamma_{ij}\gamma_{qt}}{\gamma_0^2(\gamma_0 + 1)} \qquad (ij \neq qt) \tag{9}$$

These are the building blocks to model $Cor(r^c, r^d|n)$.

$$\begin{aligned}
Var(p_i^c|n) &= Var(\sum_j p_{ij}|n) \\
&= \sum_j Var(p_{ij}) + \sum_{j \neq t} Cov(p_{ij}, p_{it}) \\
Var(p_t^d|n) &= \sum_q Var(p_{qt}) + 2 \sum_{j \neq q} Cov(p_{jt}, p_{qt}) \\
Cov(p_i^c, p_q^c|n) &= Cov(\sum_j p_{ij}, \sum_t p_{qt}|n) = \sum_{j,t} Cov(p_{ij}, p_{qt}) \\
Cov(p_j^d, p_t^d|n) &= Cov(\sum_i p_{ij}, \sum_q p_{qt}|n) = \sum_{i,q} Cov(p_{ij}, p_{qt}) \\
Cov(p_i^c, p_t^d|n) &= Cov(\sum_j p_{ij}, \sum_q p_{qt}|n) = \sum_{j,q} Cov(p_{ij}, p_{qt})
\end{aligned}$$

Now we compute

$$\begin{aligned}
Var(r^c|n) &= Var(\sum_i s_i p_i^c|n) \\
&= \sum_i s_i^2 Var(p_i^c) + \sum_{i \neq j} s_i s_j Cov(p_i^c, p_j^c) \\
Var(r^d|n) &= \sum_t s_t^2 Var(p_t^d) + \sum_{q \neq t} s_q s_t Cov(p_q^d, p_t^d) \\
Cov(r^c, r^d|n) &= Cov(\sum_j s_i p_i^c, \sum_t s_t p_t^d|n) \\
&= \sum_{i,t} s_i s_t Cov(p_i^c, p_t^d|n) = \sum_{i,t} \sum_{j,t} s_i s_t Cov(p_{ij}^c, p_{tk}^d)
\end{aligned}$$

Finally, $Cor(r^c, r^d|n)$ can be estimated by Pearson correlation

$$Cor(p^c, p^d|n) = \frac{Cov(r^c, r^d|n)}{\sqrt{Var(r^c|n) \times Var(r^d|n)}} \tag{10}$$

This formula provides the correlation of any two aspects. We can then generalize an aspect's uncertainty level provided in Equation (4) as

$$uncert_{a_i} = \min_{j=1,\ldots,m} \frac{Var(r^{a_j}|n)}{|Cor(r^{a_i}, r^{a_j}|n)|} \tag{11}$$

where a_i is an aspect. Note that, on the right hand side of above equation (11), when $j = i$, we have $Cor(r^{a_i}, r^{a_j}|n) = 1$. Hence, we get $Var(r^{a_i}|n)$ as a factor constituting $uncert_{a_i}$. The intuition behind Equation (11) is that we can take advantage of one aspect's rating to infer about the other's rating. Specifically, when two aspects are highly correlated, $|Cor(r^{a_i}, r^{a_j}|n)|$ is close to 1, thus the two aspects share the variance of the one with smaller variance.

We do not present the experimental results of this extended model as it does not show substantial improvement on key measurements so far. We doubt that it is due to the lack of a large dataset, though the model is in need of further study.

V. Experimental Evaluation

Our experiments were carried out on two real-world datasets: Amazon reviews provided by Bing Liu, et al. [4], [15], and Edmunds' car reviews that we crawled. We published our code, additional experiments and all used datasets on our supporting web page [16], for reproducibility purposes. The datasets were used to generate realistic sequences of reviews as described below.

TABLE V
DATASET STATISTICS.

	Amazon reviews [4], [15]	Car reviews
#Products	8	501
#Sentiments (l)	6	5
#Reviews/product	51	106.67
#Aspects/product	4–21	7
#Ratings avg/aspect	27.31	77.76

The Amazon review dataset has been widely studied in the sentiment analysis community since it provides the ground-truth aspects and sentiments annotated manually by the authors. Moreover, different product types have different numbers of aspect. We omit products that have less than 4 aspects with at least 10 ratings, so we have enough aspects for the algorithms to pick from and enough data to build a realistic rating distribution.

We crawled the second dataset using Edmunds' free open API on two car makes (Toyota and Honda) from 1990 to 2017, which resulted in 501 vehicles with 53,440 reviews in total. Our experiments were conducted on products that have at least 100 reviews (149 products). Furthermore, all vehicles share a fixed set of seven aspects: comfort, reliability, technology, value, performance, interior and safety. The datasets' characteristics are presented in Table V.

In our evaluation, we start each experiment with no previous ratings information, and for each new simulated reviewer, we ask them to rate k aspects of a product. We conducted experiments with various k but only present the case of $k = 3$ due to space limitation; the results for other values of k followed similar trends.

Measures: Throughout all experiments, our first two measures are based on individual aspect rating's uncertainty level $uncert^{a_j}$. The first measure utilize the uncertainty value $Var[r^{a_j}|n]$ in Equation (4), which we explained why it is a reasonable measure in Section III-B. To avoid biasing the results towards our selection algorithm that uses the aspects' variance, we introduced a second measure, which is the length of Confidence Interval (CI) of an aspect's ratings. The idea is that a smaller CI length means a higher degree of confidence we know about an aspect's rating. In our experiment the CI is $\overline{X} \pm t(S/\sqrt{N})$, where $\overline{X}$ and S are the sample mean and variance of an aspect's ratings, respectively, N is the total number of ratings, t is the critical value specified by Student's t-distribution with $N - 1$ degrees of freedom and confidence level $1 - \alpha$. We choose confidence level 95% for all experiments.

Based on above measures, the key overall uncertainty measure we consider for a product is the maximum uncertainty among its aspects. Maximum is more appropriate than average, given our problem's motivation where we want to make sure that no aspect is left behind, that is, no aspect has too uncertain rating. Specifically, a product has multiple aspects $a_1, \ldots, a_m$, with uncertainty values $uncert^{a_1}, \ldots, uncert^{a_m}$, will have uncertainty level $\max_{j=1}^{m} uncert^{a_j}$.

As a third measure, we report the ratio of the number of aspects that we are confident about its rating statistics, thus we name this measure *"High Confidence Ratio"*. The idea is that when the confidence interval length of an aspect's ratings is smaller than our desired threshold δ, then we can be confident about the aspect rating. We choose confidence level 95% and CI length threshold $\delta = 1$ for all experiments. High confidence ratio of 1 means that we are certain about all aspects' average rating. This measure reflects the degree of rating certainty instead of uncertainty as in the first two measures.

Since a dataset has multiple products, we report in the plots the uncertainty amount calculated by averaging uncertainty values over all products. In summary, we have three measures: "max variance", "max confidence interval length" and "high confidence ratio".

Baseline Aspect Selection Methods: Besides our proposed algorithm from Section III-C, we consider two intuitive baseline methods used to pick k aspects to consult a new user: "*pick random*," which picks k random aspects from the set of an interested product's aspects, and "*pick least count*," which selects the k aspects with the least number of ratings so far. Given our toy example in Section III-D, Table II, *pick random* randomly selects four aspects with equal probability, whereas *pick least count* chooses aspect "design" first, then gives the three remaining aspects equal chances (because they have the same number of ratings: 33).

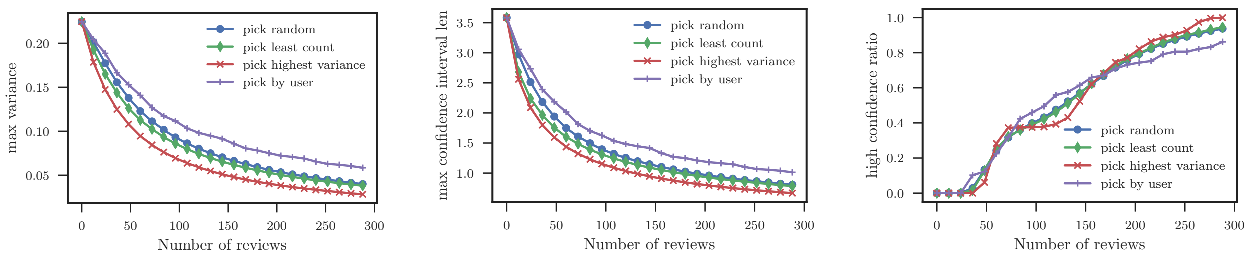
A. *"Active" Versus "Passive" Solicitation*

In the first experiment, we compare two approaches: letting the reviewer pick aspects to rate (passive, as in most existing Web sites) and actively asking them to rate specific aspects. In our datasets, the reviews of each product are fed to the various algorithm ordered by their generation timestamp. The result is presented in Figure 3, where a method asks a simulated user to rate k aspects. We refer to the user behavior in the traditional, passive solicitation as "*pick by user*" in the graphs. This method picks the first k aspects that appear in the review under consideration. If a review has less than k aspects, we decrease the same number of solicited aspects for this position in all active methods for fairness.

We use the real reviews to realistically simulate the answers of the simulated user to the k selected aspects, as follows: we look up the sentiment of the asked aspect in the review currently under consideration if available. If the aspect is missing in the review, a simulated sentiment is computed from the rating distribution (which considers all reviews, not only the ones processed so far) of this aspect of this product. We refer to this rating scheme as "*answer almost real*" since it utilizes real user reviews in most cases.

We ran this experiment 200 times on all products independently, then take the average over all products. In each run, we solicit 300 reviews, up to $k = 3$ questions per review. If a product has less than 300 real reviews, we re-use its all available reviews to simulate answers. Since this experiment requires free-text review that is unavailable on our automobile dataset, we conducted it on Amazon review dataset only.

For all measures, we notice substantial improvements of the active methods over the passive solicitation method ("*pick by user*"). Illustrated by Figures 3(a), 3(b), and 3(c) respectively,



(a) Max Variance

(b) Max Confidence Interval Length

(c) High Confidence Ratio (higher is better).

Fig. 3. Comparing "passive" and "active" review solicitation (on Amazon reviews). Smaller is better, except for High Confidence Ratio measure.

the improvement is up to 52.6% for the "max variance", 34.7% for "max confidence interval length" and 14.4% for "high confidence ratio" measure with our "*pick highest variance*" method in the end of experiment. It is also worth noting that our method reaches the desired confidence on all aspects after about 270 reviews (when "high confidence ratio" is 1), while the passive method does not even reach this level by the end of the experiment (300 reviews).

The poor performance of "*pick by user*" is expected because users are normally biased toward common aspects with many ratings, while some aspects never get enough ratings to gain a reliable rating estimation. For example, for product "Nokia 6610," aspect "size" has around 210 ratings whereas "battery life" has only about 50 ratings, even though they have similar rating distribution shapes. Other methods distribute questions over aspects in a more balanced manner, thus get better performance. This result confirms our hypothesis that carefully selecting which aspects to ask users to rate can lead to higher review profile quality.

B. Comparison of Various "Active" Solicitation Methods

In this section, we compare our method "pick highest variance" to the two baselines, "*pick random*" and "*pick least count*," on both datasets. To scale to larger number of reviews and avoid the problem of the limited number of ratings for some aspects (e.g., "technology" and "safety" in the Edmunds dataset usually have less than 10 ratings per car), we consider a different answer generation scheme, where instead of using the real reviews one by one, we compute a ratings distribution for each aspect, and sample answers (ratings) from these distributions for each review. The experimental results are presented in Figure 4 and 5.

In both experiments, we solicit 300 reviews per product, 3 questions per review. We perform this simulation 200 times, then take the average for stable results. Our proposed method outperforms the two baselines consistently on both datasets and all measures. All methods start at the same point, then gradually diverge until the end of the experiments. By the end of the automobile reviews experiment, our method yields an uncertainty value that is 36.6% and 35.6% smaller than the value of "*pick random*" and "*pick least count*" accordingly in "max variance" measure (Figure 4(a) and 5(a)). The corresponding improvements in "max confidence interval length" measure are 21.5% and 20.9% (Figure 4(b) and 5(b)). In terms of confidence ratio, when our method reaches the full "high confidence ratio" (1) after about 60 reviews, the two baselines have the confidence ratio of 0.82 roughly and only reach full ratio after 90 reviews (Figure 4(c) and 5(c)).

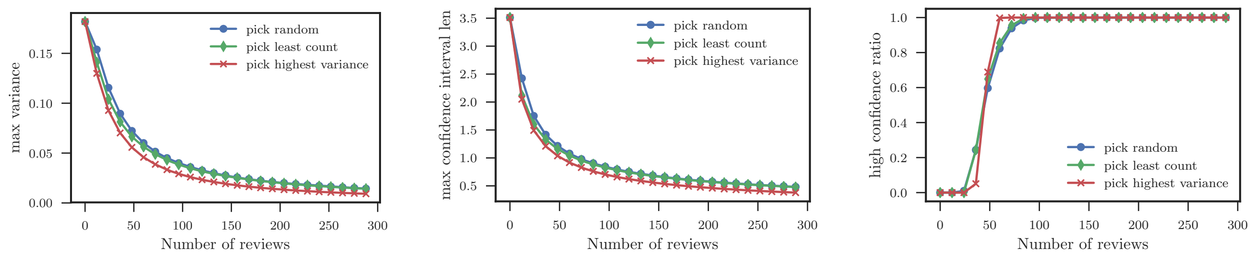
The corresponding results in Amazon reviews present similar trends. Comparing to the automobile review dataset under high confidence ratio measure, Amazon review dataset only has two differences. First, all methods reach the full ratio more slowly since Amazon products have larger aspect set, thus require more reviews. Second, our method's curve is smoother because Amazon products have a varying number of aspects instead of a fixed size (7 for Automobile products). Specifically, different number of product aspects result in different curves that are averaged to yield a pretty smooth curve as we observe.

It is worth mentioning that the two baseline methods behave slightly differently when the number of reviews performed is small; however, in the long run the number of times that aspects get selected evens out for both methods.

This is also a key difference between our method and baselines. Our method does not just ask about aspects equally as the baselines do. Instead, our method distributes more questions to aspects with contrasting ratings because these aspects need more information to solidify our belief of its rating. For instance, in toy example II, the two baselines treat "weight," "cost" and "battery" equally (same rating counts), while our method "*pick highest variance*" would ask about "battery" first due to its polarized ratings.

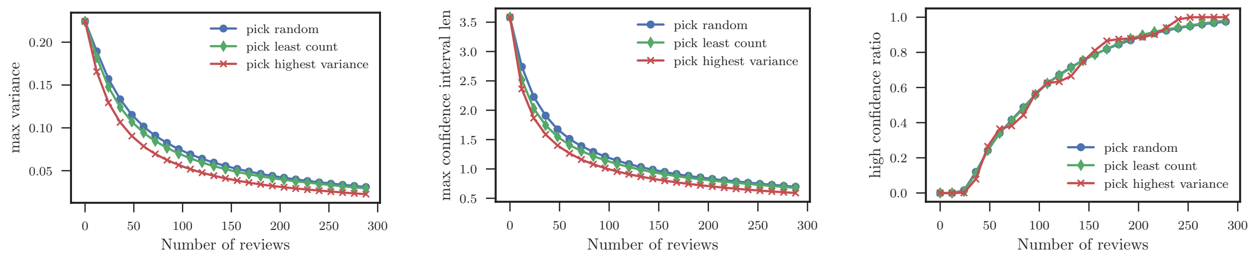
VI. Conclusions and Future Work

We have studied the problem of targeted review solicitation, which aims to achieve high-quality product review profiles, by actively soliciting aspects to rate. We adopted Bayesian inference statistics to model a review profile's key factors: product aspect rating estimation and its (un)certainty degree. We then introduced our algorithm to select k aspects to ask a new reviewer to optimize the review profile certainty. Using three different review profile quality measures, (variance, confidence interval length and high confidence ratio), we showed that our proposed "active" solicitation style clearly outperforms traditional "passive" solicitation methods on two real-world datasets. Moreover, in another set of experiments our method beats two "active" solicitation baselines under all measures. To assist others reproducing our results, all our code and



(a) Max Variance (b) Max Confidence Interval Length (c) High Confidence Ratio (higher is better)

Fig. 4. Automobile reviews. Smaller is better, except for High Confidence Ratio measure.



(a) Max Variance (b) Max Confidence Interval Length (c) High Confidence Ratio (higher is better)

Fig. 5. Amazon reviews. Smaller is better, except for High Confidence Ratio measure.

datasets are available online [16]. We also extended our model to account for correlated aspects.

In our future work, we plan to account for the uncertainty of a user answering a rating question, since our current model assumes that users always answer the questions we ask. Another research direction is to combine both "active" and "passive" solicitations simultaneously, that is, allow user to write free text in addition to rating a small set of aspects.

ACKNOWLEDGMENT

This work was partially supported by NSF grants IIS-1619463, IIS-1746031, IIS-1447826 and IIS-1838222

REFERENCES

[1] A. Dimoka, Y. Hong, and P. A. Pavlou, "On product uncertainty in online markets: Theory and evidence," *MIS quarterly*, pp. 395–426, 2012.
[2] Y. Kim and R. Krishnan, "On product-level uncertainty and online purchase behavior: An empirical analysis," *Management Science*, 2015.
[3] A. Khare, L. I. Labrecque, and A. K. Asare, "The assimilative and contrastive effects of word-of-mouth volume: An experimental examination of online consumer ratings," *Journal of Retailing*, 2011.
[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *SIGKDD*. ACM, 2004.
[5] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational linguistics*, 2011.
[6] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural language processing and text mining*, 2007.
[7] N. Jakob and I. Gurevych, "Extracting opinion targets in a single-and cross-domain setting with conditional random fields," in *EMNLP*, 2010.
[8] I. Titov and R. T. McDonald, "A joint model of text and aspect ratings for sentiment summarization." in *ACL*, 2008.
[9] A. Mukherjee and B. Liu, "Aspect extraction through semi-supervised modeling," in *ACL*, 2012.
[10] J. Weng, C. Miao, and A. Goh, "Protecting online rating systems from unfair ratings," in *International Conference on Trust, Privacy and Security in Digital Business*. Springer, 2005.
[11] N. Jindal and B. Liu, "Opinion spam and analysis," in *WSDM*, 2008.
[12] S. M. Mudambi and D. Schuff, "Research note: What makes a helpful online review? a study of customer reviews on amazon. com," *MIS quarterly*, 2010.
[13] N. Korfiatis, E. GarcíA-Bariocanal, and S. SáNchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," *Electronic Commerce Research and Applications*, 2012.
[14] H. Baek, J. Ahn, and Y. Choi, "Helpfulness of online consumer reviews: Readers' objectives and review cues," *International Journal of Electronic Commerce*, 2012.
[15] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *WSDM*, 2008.
[16] L. Nhat, R. Ryan, F. James, and H. Vagelis, "Reviews solicitation: Supporting material," www.cs.ucr.edu/~nle020/review_solicitation/, 2018.
[17] K. Chen, H. Chen, N. Conway, J. M. Hellerstein, and T. S. Parikh, "Usher: Improving data quality with dynamic forms," *TKDE*, 2011.
[18] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.
[19] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, 2012.
[20] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, 2011.
[21] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *ACL*, 2002.
[22] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *EMNLP*, 2002.
[23] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *ACL-HLT*, 2011.
[24] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," *arXiv preprint arXiv:1605.07725*, 2016.
[25] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *NIPS*, 2015.
[26] A. B. Dieng, C. Wang, J. Gao, and J. Paisley, "Topicrnn: A recurrent neural network with long-range semantic dependency," *arXiv preprint arXiv:1611.01702*, 2016.
[27] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *ICML*, 2014.
[28] A. Agresti and D. B. Hitchcock, "Bayesian inference for categorical data analysis," *Statistical Methods & Applications*, 2005.