Discovering User Bias in Ordinal Voting Systems

Alyssa Lees
Google Research
alyssalees@google.com

Chris Welty
Google Research
cawelty@gmail.com

ABSTRACT

Crowdsourcing systems increasingly rely on users to provide more subjective ground truth for intelligent systems - e.g. ratings, aspect of quality and perspectives on how expensive or lively a place feels, etc. We focus on the ubiquitous implementation of online user ordinal voting (e.g 1-5, 1 star-4 stars) on some aspect of an entity, to extract a *relative* truth, measured by a selected metric such as vote plurality or mean. We argue that this methodology can aggregate results that yield little information to the end user. In particular, ordinal user rankings often converge to a indistinguishable rating. This is demonstrated by the trend in certain cities for the majority of restaurants to all have a 4 star rating. Similarly, the rating of an establishment can be significantly affected by a few users [10]. User bias in voting is not spam, but rather a preference that can be harnessed to provide more information to users. We explore notions of both global skew and user bias. Leveraging these bias and preference concepts, the paper suggests explicit models for better personalization and more informative ratings.

ACM Reference Format:
Alyssa Lees and Chris Welty. 2019. Discovering User Bias in Ordinal Voting Systems. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3308560.3317080

1 INTRODUCTION

Crowdsourcing systems increasingly rely on users to provide more subjective ground truth for intelligent systems - e.g. ratings, perspectives on how expensive or lively a place feels, etc. Not surprisingly, there are no questions for which a pure consensus can be reached through on-line voting. Disagreement is the normal case, and the important issue is how to most productively process that disagreement into a usable signal. In this paper we examine the problem of how to process *ordinal votes*, i.e. votes on a fixed scale (e.g. 1-4 or 1-5), with Likert properties[8], from multiple users on different subjective aspects of a rated entity. Likert[1] referred to the human responses to subjective questions as *attitudes*.

We begin with a few observations of ordinal systems in practice. There are many publicly available systems that provide results of ordinal voting, such as star ratings, of movies, restaurants, music, etc:

- Mean ratings tend to converge: many rated things (movies, restaurants, songs) have 4 out of 5 stars.
- In use-cases where fractional ratings are displayed based on ordinal votes, discrepancies between ratings in the same range yield little information in terms of differentiation of quality.
- There is user bias: some users consistently vote higher or lower than others and these votes can be affected by underlying preference
- A single entity ranking can be significantly affected by a few users [10].
- Votes are *subjective*, they reflect not only attitudes, but contextual information that may or may not be accessible when processing the votes
- *There is no ground truth*

Our research question is derived from these observations:

> **RQ1**: Can we identify *consistent* user voting bias and productively harness its signal for more reliable and/or informative ratings?

The standard way to process ordinal votes is to take the majority, or plurality (the choice that receives the most votes), or some other aggregate, most commonly the mean. A naive approach standard way to deal with user bias is to throw out votes that disagree with the majority, or to down-weight them as a function of their overall agreement. Any methodology expanding on this approach is performing a form of outlier detection and will devolve in to the all 4-star restaurants problem addressed above. It is worth noting, that many existing systems have expanded on these notions to learn more elaborate user *trust* scores [2]. User trust is an integral element of evaluating any online voting system, but only represents one dimension.

In this position paper we outline an approach for identifying different kinds of bias, from users as well as the rated entities themselves, and how it can be productively harnessed. If user voting patterns are consistent, we should be able to map them into a usable signal – if someone consistently votes lower than everyone else, in theory it should be possible to re-normalize their voting behavior and turn disagreement into agreement.

Finally, one of the largest sources of voter disagreement may be attributed not to explicit behavioral bias, trust or expertise, but rather user preference. Distinguishing preference, i.e. preferring one type of movie over another, is integral to personalization of systems as well as finding a system where all global ratings don't devolve to a standard value.

Aligning biased user votes can be important in a number of ways, not least of which is increasing the number of usable votes that are available at places for which there are few votes already. For example, users probably don't need to know what other users think of McDonald's or Star Wars, there is not much we could

[1] pronounced lick-ert [7]

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317080

learn, but some small out-of-the-way shop or Indie movie that has three recorded votes may turn out to be a gem, if only we could understand that seemingly divergent votes actually agree.

2 BACKGROUND

Identifying consistent voting behaviors can be considered a generalization of the Bayesian Truth Serum [12], which cleverly identifies a particular voting pattern among experts, in cases where expertise applies, such as the famous wine tasting game; experts tend to disagree with the crowd, but are also capable of predicting that disagreement.

In the case of star-ratings, one could argue that some amount of expertise and background knowledge applies, but for more quantitative tasks like rating the price level of a restaurant, expertise is less obvious. We could consider the range of places a person has visited as expertise. In general, we simply search for consistent, patterns of voting behavior by comparing users to some relative truth, such as the plurality vote, the mean or as suggested in this paper, an informed prior.

Much recent work on bias in machine learning looks for different kinds of bias, bias that favors one population of users in ways we may consider unfair, such as race, gender, etc [4]. However, in this work we focus on explicit differences in user voting behavior only. These differences can be roughly divided into subcategories of voting behavioral distribution trends, preference and reliability/trust.

Existing work on user reliability in the context of online social media usage and online question answer systems [2, 3], and user voting behavior bias [18] suggest a re-ranking methodology to remove the existence of these behavioral data biases. Explicit user behavioral preference in online queries can be exploited for better click-through performance [19].

There is a wealth of work in personalization, e.g. via Bayesian probabilistic models extracted from learned explicit user preferences for recommendation systems [15]. Online aspect ratings can be predicted from user reviews [16] as well as user data in ordinal voting systems [11]. The focus on *aspects* of the places being voted on, such as ambience, food quality, etc. in [11] explicitly analyzed the bias of user populations via an Ordinal Aspect Bias Model using Bayesian inference.

While this paper will focus on a high level model of different user biases in ordinal voting systems, we are inspired by the rich background of collaborative filtering techniques that assist in clustering and identifying behavioral patterns and rating prediction [5, 9]. It is important to note that despite the nature of the ordinal votes used in these techniques, there has been an underlying assumption that the rankings follow a continuous valued distribution. We will continue that assumption in our bias discussion.

Many statistical techniques utilizing Bayesian inference require cutoff points in a Gaussian distribution, other work uses a logitic regression model for category data [6]. Another alternative is the logistic stick-breaking process for separating clusters and mapping the categorical likelihood [14].

Finally, there has been working expanding existing Bayesian coclustering techniques [17] for identifying user behavior preferences. This work has exploited Gibbs sampling [1] to generate examples from the posterior for improved parameter estimation.

3 VOTING BIAS IN ORDINAL SYSTEMS

In order to process votes to identify user bias, we make a few assumptions. Given a set of rated entities, $e \in \mathcal{E}$, users $u \in \mathcal{U}$, and ordinal user ratings $r_{u,e} \in \mathbb{N}^{[0,M]} = \{x \in \mathbb{N} | 0 \leq x \leq M\}$, where M denotes the maximum ordinal voting value. We denote the *true rating*, independent of bias, of a rated entity in a continuous domain as $v_e \in \mathbb{R}$. Although we have no expectation to ever know any particular v_e, we do make assumptions on the prior distribution from which any particular v is drawn. For simplicity, we model the prior with a Gaussian normal distribution, $P(\mathcal{V} = v) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ that is truncated at the domain value cutoffs.

Next we allow for an aggregation function on the set of votes for an entity, to produce a set of raw ratings $r_e \in \mathbb{R}$ from the set of user ratings, $r_e = \mathfrak{F}_e([r_{u,e}])$. Typical choices for the function $\mathfrak{F}$ are the mean or plurality (most popular ordinal vote). Since individual votes are subject to bias, the distribution of raw ratings will be skewed by this bias, and will have a different, possibly non-Gaussian, distribution.

In the simplest case, and the standard approach to disagreement, users vary consistently in their reliability, so that there is a user trust vector $\bar{t}$, such that $v_e = \mathfrak{F}'_e([r_{u,e}], [t_u])$. Typical choices for $\mathfrak{F}'$ are weighted mean, or weighted plurality.

Mitigating the ineffectual information relayed by such votes can begin simply by filtering outliers (both in voting and users), changing how distributions of votes are displayed, re-weighting and catering to specific user preferences.

We explore several avenues towards imparting more informative user votes:

- Explore explicit aspect voting bias to identify common voting behaviors and preferences via clustering.
- Identify user zeitgeist behavior (bias) to inform global distributions of votes
- Utilize location bias (say for restaurants and hotels) to change conditional distributions of votes.
- Identify individual user preference (bias) and leverage for personalization.
- Identify individual voting behaviors and normalize bias-independent display
- Leverage user trust scores for increased accuracy in ranking.

4 BIAS/PREFERENCE MODEL

We propose modelling user ordinal voting systems with an assumption of a theoretical *true* underlying prior probability distribution function for the ranking of the entities E. In this idealized model, individual unbiased *true* entity rankings are drawn from the PDF. There are two not necessarily intuitive points with this assumption: 1. the underlying entity rankings are drawn from continuous distribution 2. that de-biased entity model may reflect a different distribution than the ordinal voting samples.

For a subjective voting system, such as movie ratings, a suitable prior may be a normal distribution. While unbiased selections should naturally follow a bell-curve, with the majority ranking average and a few achieving excellence, a more objective voting task, such as one requesting users to rate the price level of a restaurant, may be informed by the real-life skew of establishments. In sample data sets across many regions, the authors found that restaurant

data demonstrates a skew towards inexpensive and moderate establishments. In this scenario, a suitable prior could be modelled by a Gamma distribution.

We extend our definition of the *true* rating of an entity to cover this expected distribution:

$$v_e : e \in \mathcal{E}, P(\mathcal{V} = v) = F_V(v)$$

where V is a continuous random variable with a prior modelled by the function $F_V(v)$. Again, this *true* value is not necessarily observed in sampled user votes and likewise is not necessarily representative of what should ideally be displayed in any online ranking/rating system for the entities $\mathcal{E}$. Rather we argue that these continuous values are skewed by several kinds of bias. The first kind of bias, which we will deal with later, is user bias. The second kind of bias, which we discuss in this section, is a natural bias caused by extrinsic and global factors, that we consider properties of the voting question. Examples of such natural bias include culture, geographic location and genre, and there are many more.

Location bias includes the notion of rural versus urban; a rural area may apply different standards of expectations for price and menu items than a more populated area. Similarly, different countries and regions have demonstrated clear differences in dominant preferences and cultural tastes. The concept of location applies to search query rankings, movie rankings, and almost any conceivable online ranked result.

Genre bias includes particular styles for a given entity class. In the space of movies, genre may encompass *action* or *romantic comedy*. In the space of restaurants, genre could be a type of cuisine such as *Italian* as well as experiential features of the establishment such as *trendy, casual,* or *upscale*.

Genre and location are often intertwined. For example the genre of Bollywood movies are more popular in specific regions. Likewise the skew of ranking may be similarly affected.

4.1 Natural Bias

These notions of **Natural Bias** where for particular combinations of e.g. location and genre, the zeitgeist opinion of users or the actual entities themselves demonstrate a modified distribution from the underlying F_V prior. The point here is that natural bias may distort and skew the distribution of votes according to trends that are separable from individual user bias. Action movies tend to rate higher, they have their own distribution, rural areas have different price level standards, and its hard to find a genuinely expensive restaurant.

To account for these natural biases we define sets to partition the entities, and distributions within those sets. For example, we define genre $g_e \in \mathcal{G}$ as the genre of entity $e \in \mathcal{E}$, and $l_e \in \mathcal{L}$ as the location of entity e. Note that in the context of user votes, the location of the user voting l_u would be a separate consideration, but is outside the scope of this paper.

We call the distribution of the rated entities with their natural bias score $Z_{g,l}$ and we define $z_e \in Z$ to be the natural bias adjusted rating of an entity. Again, we don't actually have these values, but we expect the overall distribution to be closer to the distribution of entities by their raw votes, $r_e \in \mathcal{R}$. We define a function $h : \mathbb{R} \mapsto \mathbb{R}, (v, g, l)$ that reflects the mapping between the ideal distribution and the natural bias, such that:

$$P(z_e|V, G, L) \sim h(v_e, g_e, l_e)$$

Similarly, we define the inverse, mapping naturally biased entity ratings back to the ideal distribution V as

$$h^{-1}(Z, G, L) \sim V$$

In other words, if we can characterize the bias according to genre and location as distributions, given an ideal distribution we can collect the bias-adjusted rating of individual entities.

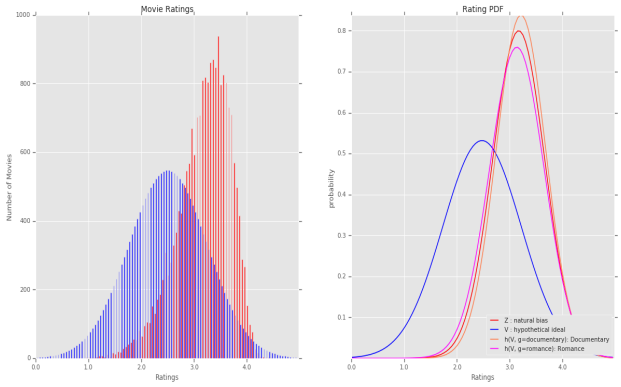

Figure 1: Idealized Movie Rating Distributions with Natural Bias

All we really have, however, is raw user votes on each entity $(R_e = r_{1,e} \ldots . r_{n,e})$, where each vote is on the ordinal scale, leading to some raw distribution of entities on a continuous, real-valued scale. Our next problem is converting the scale of those distributions back into the ordinal scale. If, for example, the entities follow a normal distribution centered on 1.4, at what point do we "cut off" and say an entity has a 2, or a 1, rating?

Rating Likelihood: Stick Breaking Dirichlet Process. We represent the probability of any given user-entity rating with a specific ordinal vote value as $P(r_e = m)$ where m is an ordinal value $m \in \{1, .., M\}$. There are many ways to model discrete values to a continuous distribution. Here we will follow established precedent to employ the Dirichlet Stick-Breaking process. In this representation, the probabilities $P(r_{u,e} = m)$ are defined by a procedure of cutting up a unit length stick at cut points $C = \{c_1, ..., c_{M-1}\}$ where $c_1 < c_2 < \ldots < c_{M-1}$. The purpose of this cutting is essentially to find the best real-valued points in $[1, M]$ to use as cut points to divide the distribution into the original ordinal scale.

The stick-breaking Dirichlet process itself contains two variables, locations $\{\theta_m\}$ that are independent and identically distributed over the process and the corresponding probabilities $\{\beta_m\}$. The Dirichlet process probability mass distribution for a discrete set of values has a probability mass function:

$$f(\theta) = \sum_{m=1}^{m=\infty} \beta_m * \delta_{\theta_m}(\theta)$$

and an indicator function $\delta_{\theta_m}(\theta_m) = 1$ and zero for every other ordinal voting value. We will need to relate θ to our cut points in C.

The stick breaking process begins by viewing β_m as a length of a piece of a unit stick broken at m and the remaining stick as $\beta_m^{'}$, such that

$$\beta_m = \beta_m^{'} * \prod_{i=1}^{m-1}(1-\beta_i^{'})$$

In our scenario, $\zeta_e^m = |c_m - z_e|$ is the distance from the idealized rating to the cut point, so that the actual probabilities align, $P(r_e = m|z_e, C) = p(r_e = m|\zeta_e^m)$. To translate to the stick breaking formula, we use the sigmoid function to ensure that the ζ distance from cutoffs is reduced to a unit length stick. As such,

$$\beta_m = s(\zeta_e^m) = \frac{e^{\zeta_e^m}}{1+e^{\zeta_e^m}}$$

and

$$P(r_e = m) = \prod_{m^{'}<m}(1 - s(\zeta_e^{m^{'}})) * s(\zeta_e^m)$$

Finally, we approximate z_e as the expected value of the individual user ratings:

$$P(Z = z_e) \sim E(r_e) \sim \frac{1}{|U_e|}\sum_{u \in U_e} r_{u,e} \sim h(v_e, g_e, l_e)$$

4.2 User Reliability

As highlighted in the introduction, real life scenarios are riddled with instances where a few user voters can significantly alter the outcome of online voting. Examples include spam, where a user votes randomly or for one value on numerous entities[13]. Alternatively, there are examples where friends of an establishment repeatedly vote online to mitigate negative online rankings. Notable instances have been recorded where a single disgruntled customer can offer a poor rating that throws the displayed ranking of a product shopping sight off for months. This is especially problematic, when ratings translate effectively become binary - i.e. everything above say a 4 start level is about the same and regarded as decent and anything below is viewed with skepticism.

To address, our model of user bias must incorporate an individual user trust score $t_u \in \mathbb{R}^{[0,1]}$ and T_U is the set of trust scores across users in U.

$$P(Z = z_e) \sim \frac{\sum_{u \in U_e} t_u * r_{u,e}}{\sum_{u \in U_e} t_u}$$

Modelling user reliability/trust can be learned via historical data on individual basis. Alternatively suspected outlier detection can be learned more broadly in the user and entity domain spaces.

4.3 User Voting Behavior and Preference Biases

Additionally we assume that users have *behavioral* and *preference* bias that should be treated separately from *trustworthiness*. These types of bias can be used for personalization, recommendation and individualized tailoring of ranking results.

(1) **Voting Behavioral Bias** : We will define voting behavioral bias as $b(u, e)$. Voting behavioral bias is NOT due to user trust and can take many forms.
- Users may vote different if given a series of ordinal questions online. Studies have demonstrated that more attention may be given to the first question then later items.
- Some users have more optimistic or pessimistic trends. In other words, some users may have a bias towards voting one level higher than the norm while others may vote lower. If identified, these voting distributions can be re-normalized for finer grained results.
- Expert versus Novice voting. A novice may conform to more popular voting conventions while an expert may display a wider range and finer grained discrepancy between votes.

(2) **User Preference** : Individual users display different individualized preferences dictated by genre and location: $a(u, g_e, l_e)$. These individuals preferences are separate from the Global Bias trends due to these factors. For example, a user may prefer action movies or romantic comedies, and hence are prone to giving higher rankings to *Action* movies independent of underlying movie quality. The same is applicable to restaurants, shopping etc. Even in a less subjective domain, such as price level, a user's particular preference of upscale versus affordable, may affect the expectation of what is expensive versus not.

Preferences in particular can be inferred by user vote rankings, stated preferences, and history of visiting or ranking particular genre/location pairs.

Capturing these differences in preference and behavior is integral for informed universal ranks as well as personalized recommendations.

We loosely incorporate the individualized behavior preferences and biases into our model, arguing that weighting w/ trust scores alone will not ensure that user rating scores will approximate the globally biased distributions Z_l for a given location l. Two factors are to be considered

- If user voting preferences for genres in a given location, demonstrate different marginal probabilities for members of G than the globally biased distributions Z_l, then re-weighting user votes by preference is necessary.
- If general user voting behavior for a given genre and location $U_{g,l}$ does not reflect the underlying prior for the *true values* V, then the distribution of individual votes need to be transformed.

We consider $\hat{R}_{U_{L,G}} s \in \mathbb{R}$ as the theoretical continuous distribution of ratings provided by users defined by discrete values with individual bias applied for a given for locations and genre. In an ideal scenario with a set of trusted users $U_{G,L}$ for a specific genre and location, we subtract global bias with the inverse of h, and apply individual de-biasing function

$$b(h^{-1}(\hat{R}_{U_{L,G}}, L, G), L, G) \sim V_{G,L}$$

to approximate a distribution of votes that emulates the prior for the *true value* V. In practice, we consider b to be a scaling function, so that if users in a given location and genre demonstrate very tight distribution around a value, the function will map to a wider range of values.

Similarly, $a(u, g, l) \in \mathbb{R}^{[0,1]}$ maps an individual user preference for a particular genre in a given location, where 1 is viewed as complete preference and 0 is absolute dislike. The preferences across users for a particular location can be weighted to match the prior of a given $p(G = g)$.

Informed Rankings. To add finer granularity to the displayed ranking$\hat{r}_e$ is to weight users votes by the individual users preference for the given entity e genre/location pair. The intuition behind this suggestion is that a user with a known preference for a given genre $g \in G$ and $l \in L$ will have informed expertise on the given entity e. Applying preference scores can ensure that users with low preference for a particular genre are down-weighted and hence possible inconsistency (or down-weighted) scores are removed.

$$z_e \sim \frac{\sum_{u \in U_e} b(r_{u,e}, g_e, l_e) * a(u, g_e, l_e) * t_u}{\sum_{u \in U_e} a(u, g_e, l_e) * t_u}$$

In the same manner, personalization can conceivably be incorporated into the system. In this scenario, the displayed rankings would be customized to a *target* user's specified preferences.

In order to accomplish this task, votes can be weighted by a *similarity* score between a target user of the system and the user vote participants. In this context the function $a(u, g, l)$ can be replaced with $S(u, u')$, where S is a similarity metric and u' is the target user fro display.

$$\hat{z}_{u',e} \sim \frac{\sum_{u \in U_e} b(r_{u,e}, g_e, l_e) * S(u, u') * t_u}{\sum_{u \in U_e} S(u, u') * t_u}$$

A simple similarity metric could be the cosine-distance between tensors of users voting history or utilizing another collaborative filtering technique. Depending on the system, similarity can be defined by voting history, outside information, user specified preferences or user visit information. It is outside the scope of this work, but to ease in computation, users can be associated with predefined preference clusters, as such the choice of rating displayed is cached to a small number of pre-defined entities.

5 FUTURE DIRECTIONS

The concepts of trust, voting behavior and preference all have overlaps in interpretation and effective modelling. We separate explicitly in this paper on the user level and the global level, to dissect mechanisms in which informed ordinal rankings can be extracted.

For future work, we note that computing user preferences, trust, and voting bias on an individual user basis may not be computationally feasible or desirable. Instead of calculating individual voting bias, preference and trust, such values can be approximated via distance to pre-defined voting behavior and preference clusters.

An alternative consideration is personalizing rankings in a voting ordinal display. Instead of displaying rankings formed by weighting preference for a given entity genre/location, the individual votes can be weighted by the end users similarity to the voting users preference cluster.

REFERENCES

[1] Nicola Barbieri. 2011. Regularized Gibbs Sampling for User Profiling with Soft Constraints. *2011 International Conference on Advances in Social Networks Analysis and Mining* (2011), 129–136.

[2] Jiang Bian, Yandong Liu, Ding Zhou, Eugene Agichtein, and Hongyuan Zha. 2009. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *WWW*.

[3] Bee-Chung Chen, Anirban Dasgupta, Xuanhui Wang, and Jie Yang. 2012. Vote Calibration in Community Question-answering Systems. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 781–790. https://doi.org/10.1145/2348283.2348388

[4] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? *arXiv e-prints*, Article arXiv:1805.12002 (May 2018), arXiv:1805.12002 pages. arXiv:stat.ML/1805.12002

[5] Thomas Hofmann. 2003. Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*. ACM, New York, NY, USA, 259–266. https://doi.org/10.1145/860435.860483

[6] Yehuda Koren and Joseph Sill. 2013. Collaborative Filtering on Ordinal User Feedback. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*. AAAI Press, 3022–3026. http://dl.acm.org/citation.cfm?id=2540128.2540570

[7] Gary P. Latham. 2012. *Work Motivation: History, Theory, Research, and Practice*. Sage Publisher.

[8] R. Likert. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology* 140 (1932), 1âĂŞ55.

[9] Benjamin M. Marlin. 2003. Modeling User Rating Profiles For Collaborative Filtering. In *NIPS*.

[10] David Owen. 2018. Customer Satisfaction at the Push of a Button: HappyOrNot terminals look simple, but the information they gather is revelatory. *The New Yorker* (Feb. 2018). https://www.newyorker.com/magazine/2018/02/05/customer-satisfaction-at-the-push-of-a-button

[11] Lahari Poddar, Wynne Hsu, and Mong-Li Lee. 2017. Quantifying Aspect Bias in Ordinal Ratings using a Bayesian Approach. In *IJCAI 2017*.

[12] Drazen Prelec, Hyunjune Seung, and John McCoy. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541 (01 2017), 532–535. https://doi.org/10.1038/nature21054

[13] Vikas C. Raykar and Shipeng Yu. 2012. Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *J. Mach. Learn. Res.* 13 (March 2012), 491–518.

[14] Lu Ren, Lan Du, Lawrence Carin, and David Dunson. 2011. Logistic Stick-Breaking Process. *J. Mach. Learn. Res.* 12 (Feb. 2011), 203–239. http://dl.acm.org/citation.cfm?id=1953048.1953055

[15] David H. Stern, Ralf Herbrich, and Thore Graepel. 2009. Matchbox: large scale online bayesian recommendations.. In *WWW*, Juan Quemada, Gonzalo LeÃşn, YoÃńlle S. Maarek, and Wolfgang Nejdl (Eds.). ACM, 111–120. http://dblp.uni-trier.de/db/conf/www/www2009.html#SternHG09

[16] Hao Wang and Martin Ester. 2014. A Sentiment-aligned Topic Model for Product Aspect Rating Prediction. In *EMNLP*.

[17] Pu Wang, Carlotta Domeniconi, and Kathryn Laskey. 2009. Latent Dirichlet Bayesian Co-Clustering. 522–537. https://doi.org/10.1007/978-3-642-04174-7_34

[18] Xiaochi Wei, Heyan Huang, Chin-Yew Lin, Xin Xin, Xianling Mao, and Shangguang Wang. 2015. Re-Ranking Voting-Based Answers by Discarding User Behavior Biases. In *IJCAI*.

[19] Qianli Xing, Yiqun Liu, Jian-Yun Nie, Min Zhang, Shaoping Ma, and Kuo Zhang. 2013. Incorporating user preferences into click models. In *CIKM*.