Predicting the quality of new contributors to the Facebook crowdsourcing system

Julian M. Eisenschlos
Facebook
Menlo Park, CA 94025
julianeisen@fb.com

Abstract

We are interested in improving the quality and coverage of a knowledge graph through crowdsourcing features built into a social networking service. In this setting, most participants are casual users, making only a few contributions, and do so incidentally in the course of using the service. Techniques that make assumptions about the matching of users to questions, or the number of answers per user or per question do not work well under such circumstances. We present an approach to model user trust when prior history is lacking, so that we can incorporate more new users' contributions into crowdsourced decisions, and provide quicker feedback to new participants. Specifically, we present a logistic regression classifier for first-time contributions, and study the effect of prior knowledge about user demographics on this classifier using Facebook crowdsourcing datasets.

1 Introduction

Although automated systems are getting better at extracting knowledge from the web by parsing natural language and HTML [1], there are scenarios where human input is necessary [8] - for example, when information about the real world isn't catalogued in a programmatically accessible form, or is changing regularly or unpredictably, or when a machine learned predictor has insufficient confidence in its predictions. Crowdsourcing can help address these scenarios at scale.

We study the problem of gathering subject-predicate-object tuples for a database of international locations. The predicates under consideration include name, photo, phone number, website, address, open hours. For instance, we might know that *Facebook HQ* is located at *1601 Willow Road, Menlo Park, CA*, but we may be missing a phone number for this location. Our crowdsourcing interfaces enable users with differing levels of experience and reliability to suggest possible phone numbers, and our objective is to choose one (or more) phone numbers if we can determine that it is correct with high confidence.

Crowd contributors may be motivated intrinsically or extrinsically. Intrinsic motivations include the urge to prove oneself knowledgeable, to help others, or fix wrong information; while extrinsic motivations usually involve some material or monetary gain. To achieve worldwide coverage for places, we prefer to focus on the former because it is difficult to obtain access to enough paid workers with local knowledge. Further, [1] suggests that better results are achieved when one appeals to intrinsic motivations. In such a setting, most contributions are voluntary, and a majority of users provide a single or just a few answers.

Prior work on voting aggregation algorithms - using expectation maximization [4], message passing [2] or other techniques - shows that assessing the quality of the contributor by their prior history and level of agreement with the crowd leads to improved performance when compared to simple majority voting. These approaches make assumptions about regularity conditions on the distribution of users

and questions, which are hard to enforce in our volunteer-biased system, significantly limiting the utility of these approaches.

We therefore want to understand how to predict the quality of answers when prior history about user contributions is lacking. Being able to correctly perform this classification allows more contributors to have impact on the system, and more opportunities to provide feedback and re-engage users.

The summary of our work is as follows:

- We study the distribution of demographic properties across the population and among highly engaged and/or highly accurate contributors.
- We leverage that information to build a classifier of first-time contributions.

2 Crowdsourcing at Facebook

The crowdsourcing system at Facebook includes several product interfaces where a user can suggest new values for an attribute or vote on other users' suggestions. We use the term *claim* to refer to either a suggestion or a vote. Clearly, voting on suggestions requires less effort than making new suggestions.

One such crowdsourcing product, the Graph Editor [1], is customized for highly engaged contributors, and consists of an endless queue of locations requiring attention because they are missing attributes or suspected to have incorrect attributes. Experienced contributors can customize the queue to show them places they are familiar with, and make claims about these places. We refer to users that contribute lots of claims regularly as *power* contributors.

Graph Editor users, however only form a small fraction of the total crowdsourcing contributors on Facebook ($< 1\%$), while accounting for a disproportionate fraction of the claims (over 10%). The majority of contributions come from *casual* contributors who encounter missing or erroneous values in the course of using other features on Facebook, and make a crowdsourcing claim while reporting this. Most casual users will only interact very occasionally with the crowdsourcing system, making only one or two claims ever.

Our crowdsourcing algorithm has two main tasks:

- Measure the quality of a particular user's contributions
- Determine which S-P-O triple claims to accept or reject

Our algorithm is related to Karger et. al. [2] with an added time dimension. We run a process where we iteratively update the beliefs of triples associated with a subject (a Facebook Place) based on our current estimates of contributor quality, or update the estimates of a user's quality based on the triples that user has contributed to. When triples reach certain confidence thresholds, we can mark those claims as accepted or rejected.

Strictly speaking, we are modeling a claim's probability of correctness q, based on features including the contributing user's history, the Place in question, the interface that the claim was made in, the relationship between the contributor and the Place etc. Historically, we have required that contributors gain a sufficient track record before they are permitted to have impact on the system. This is conservative because new contributors are the most likely to be confused about the crowdsourcing feature, or to misinterpret guidelines, or even to be malicious; however in our setting this ignores a significant fraction of incoming claims. Incorporating other features about these new contributions could help us be smarter about valuing these claims, and give a voice to new contributors.

3 Demographics

In this section we present results about how key demographics are distributed among the average Facebook user, compared to the crowdsourcing population. For instance, crowd contributors are more likely to be male, slightly older, have been on Facebook longer, are generally more connected

[1] Publicly available at `www.facebook.com/editor`

Table 1: Comparing demographic aspects of contributors

Variable	Average increase	Kolmogorov-Smirnov statistic
Age	7%	0.12
Gender (% male)	11%	0.061
Time since joining Facebook	36%	0.21
Friends	88%	0.25
Following	183%	0.22
Followers	556%	0.30

Table 2: Comparing contributor density in different countries

Country	CD	Number of claims per contributor
Thailand	3.2	$1-10$
Laos	3.0	$1-10$
Taiwan	2.3	$20-30$
USA	0.9	>30

(i.e. friends, followers and followees) and are more engaged with Facebook. See Table 1 for detailed results using the *Kolmogorov-Smirnov* test [5,6]. A P-value below 1×10^{-15} proves that the test results are statistically significant.

We can also look at the geographic distribution of crowdsourcing. Some countries have a disproportional number of contributors (e.g. Thailand) relative to proportion of population on Facebook.

To quantify this, we can look at the contributor density (CD)

$$\mathrm{CD}(c) = \frac{\mathbb{P}(C = c, R = 1)}{\mathbb{P}(C = c)\mathbb{P}(R = 1)} \propto \mathbb{P}(R = 1|C = c)$$

where c is a country and C is the random variable that models the home country of a Facebook user and R the random variable that models whether a user is participating in crowdsourcing. These are simply the terms in the mutual information gain of C and R, a larger CD indicates more contributors than expected in that country. The observed density for some sample countries can be seen in Table 2 and we can look at the overall distribution in figure 1

As we can see, the geographical distribution of crowd contributors is not necessarily correlated to the propensity to contribute. Some countries have lots of causal contributors and others lots of power contributors. This is an important consideration while building up crowdsourcing communities, and balancing volume and spread generated by casual contributors against calibration and review supplied by power contributors.

Casual and power contributors also differ in platform usage, as seen in Table 3. Not only is the crowdsour-cing population more engaged Facebook users, they tend to use desktop more than mobile.

These differences are all relevant to designing new crowdsourcing interfaces, understand what motivates contributors, and even to the prediction backend.

4 Feature classes and training data

In this section, we build a classifier to recognize accurate claims coming from first-time contributors using a logistic regression model incorporating demographic features.

We consider three families of features relevant to evaluating the quality of a claim coming from a new contributor:

- *Contributor features*, including demographics and patterns of Facebook usage as well as interests and behaviours obtained through the work in section 6.

Table 3: Average number of days active on platform in the last 7 days compared to the average user

Platform	Power contributors	Casual contributors
Desktop	$+233\%$	$+94\%$
Mobile	$+37\%$	$+37\%$

- *Place features*, including popularity for this Place and categorization (into an in-house taxonomy). Adding this information might enable the model to learn a proxy for question difficulty and expertise of demographic groups. For instance we might learn that zip codes are harder to come by in a particular country, or that younger users are better in filling information for pubs. To capture such interactions between features, we incorporate bigram features corresponding to each pair of covariates.
- *Claim features*, including features characterizing the connection between the contributor and the Place in question, the direction of the claim (agree or disagree), whether the claim is a suggestion or a vote, etc.

One of the hurdles we had to overcome was obtaining labeled data for training and testing. We use two kinds of labeled data sets, and they represent different tradeoffs between coverage and quality.

One set of labels (*the gold set*) is derived from in-house manual annotation work collected over a period of time, as well as high-confidence automated predictions. We estimate its accuracy above 95% but it incorporates various biases, including towards being mostly positive.

A second set (*the inferred set*) can be obtained using our crowdsourcing inference algorithm described in section 2 and considering the labels it has deduced. It can have a lower accuracy but it is more evenly distributed over the Places database. We will omit consideration of these differences here, due to space.

Another direction that has been studied previously in [3] is assuming values that have passed a *test of time*, by being untouched for a long period of time to be correct. This is a worthwhile extension for us to consider.

We also included small manually rated sets of claims randomly sampled from new users' contributions to obtain an unbiased set to be used for validation.

In training we combine the gold set and the inferred data sets, assigning them weights w_g and w_i respectively. We treat the ratio of these weights as a training parameter to be chosen via cross-validation.

In the next section we will explore how these features and training sets affect the performance of the model.

5 Results

In figure 2 we show the precision - recall curve obtained by our classifier in the validation set, comparing to test sets extracted from the gold set and the inferred set. We observe that the validation set performs worse than the others and attribute that to biases in the training sets. The model still shows considerable improvement compared to the baseline of 57% precision of new claims, as measured on the validation set.

We also studied how the choice of the weights given to the gold set and the inferred set samples in the stochastic gradient descent impacts performance. The trade-off of accuracy and coverage of the training data affects the precision–recall curve on our validation set. We searched for a ratio of w_g and w_i that optimized the performance on the unbiased validation set, and found empirically that giving the inferred set labels half the weight of gold set samples yields the best results.

It is important to clarify that the precision of the new contributor quality prediction is not the precision of the resulting knowledge base; this model primarily behaves as a filter to select potentially good contributions from new contributors (without the benefit of prior history features) and give

them a weight in the inference algorithm that then combines claims from both new and experienced contributors.

Finally, to understand the relevance of the three families of features described in the previous section, we perform a feature ablation study. The results are in figure 3. The most relevant set of features are the claim features, which model the interaction between the place and the user, but all three sets contribute to achieve the precision levels we are targeting (above 90%) with a reasonable recall.

6 Applications to growth

Besides the main purpose of this work, we would also like to use demographic information to grow our crowdsourcing communities. [1] described an interesting approach to leverage advertisement targeting systems to engage accurate contributors.

Here we discuss an extension to bootstrap such an approach in the setting of the Facebook ad targeting framework. We can use simple classification methods like linear discriminant analysis [7] applied to *anonymized* Facebook data to identify targeting features that distinguish the crowdsourcing population from the average Facebook population. Once such features are discovered, we can leverage Facebook's targeting capabilities to reach potential new contributors.

To see how this works, consider the following: for each user u in the set of all Facebook users U we can compose a binary feature vector v_u containing demographic and geographic features but also interests in the form of Facebook Pages, apps, and Groups they connect to or interact with. This is a high dimensional vector, so that whatever density estimation we perform must have low complexity.

If $C \subseteq U$ is the subset of potentially highly engaged and accurate crowd contributors, and f_1 is the probability density function of $v_u : u \in C$ and f_2 the probability density function of $v_u : u \in U$, then we can calculate the regression function as

$$r(u) = \mathbb{P}(u \in C | v_u) = \frac{\pi_1 f_1(v_u)}{\pi_2 f_2(v_u)}$$

where π_1 is the size of C and π_2 the size of U.

We can estimate a sample C by selecting engaged crowd contributors who respond correctly to calibration questions. Assuming f_1 and f_2 as Gaussians with means

$$\mu_1 = \frac{1}{\pi_1} \sum_{u \in C} v_u \text{ and } \mu_2 = \frac{1}{\pi_2} \sum_{u \in U} v_u$$

and equal covariance matrix Σ, then $r(u) \propto \exp(v_u^t \Sigma^{-1}(\mu_1 - \mu_2))$.

In practice we assume Σ to be diagonal, thus assuming the features to be independent from each other to make the time complexity linear on the number of user–feature connections on our sparse graph.

The features with highest absolute value in the weight vector $\Sigma^{-1}(\mu_1 - \mu_2)$ are most relevant in this discrimination. Note that because we are interested in the features that distinguish the two sets, we can use anonymized data in this analysis.

Some of the features that appeared using this method, common among crowd contributors, were affinity to technology applications or entities such as *Duolingo*, *IMDB*, *Foursquare*, or *NASA*. We also found behaviors indicating technology early adoption, and engaging in local-related activities such as traveling, eating out or going to the theater.

The Facebook advertising system can be leveraged to target capaigns to people with the specific interests and behaviors we found to be relevant. Furthermore, having created an ad with these targeting features, we can set up the advertising system to automatically improve targeting performance by creating a feedback loop with the advertising system by using conversion pixels. There are many metrics we can optimize for in terms of conversions, combining volume and accuracy.

We have conducted small tests using this idea and evaluated incoming claims using calibration questions. In our experiments, we find that our ads-based targeting reaches more accurate contributors compared to other organic channels that reach the Graph Editor, showing a 6% improvement over the next best channel.

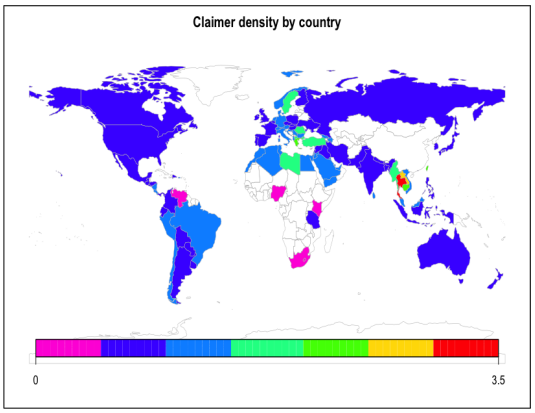

Figure 1: Heatmap of CD

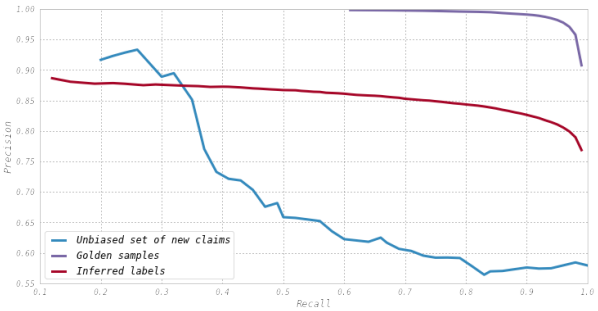

Figure 2: Precision – Recall curve on different test sets

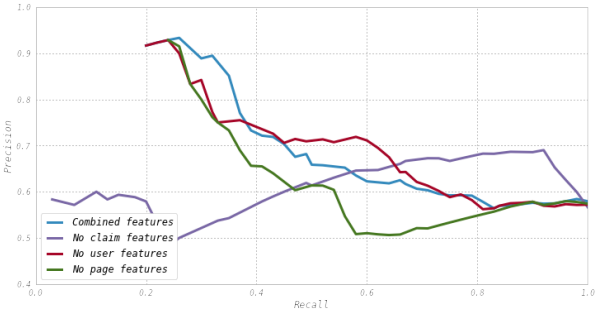

Figure 3: Feature group ablation test

Acknowledgments

I would like to acknowledge Annie Liu, Kedar Bellare, Onur Ismael Filiz and Venky Iyer for their invauluable feedback and advise.

References

[1] P. Ipeirotis and E. Gabrilovich (2014) Quizz: Targeted Crowdsourcing with a Billion (Potential) Users. *www*

[2] D. R. Karger, S. Oh, and D. Shah. (2011) Iterative learning for reliable crowdsourcing systems. *Advances in Neural Information Processing Systems 24, pages 19531961*

[3] C. How Tan, E. Agichtein, P. Ipeirotis and E. Gabrilovich (2014) Trust, but Verify: Predicting Contribution Quality for Knowledge Base Construction and Curation. *WSDM*

[4] Q. Liu, J. Peng and A. Ilher (2012) Variational Inference for Crowdsourcing *NIPS*

[5] A.N. Kolmogorov (1933) Sulla determinizione empirica di una legge di distribuzione *Giorn. Ist. Ital. Attuari , 4 pp. 8391*

[6] N.V. Smirnov (1938) On estimating the discrepancy between empirical distribiution curves for two independent samples *Byull. Moskov. Gos. Univ. Ser. A , 2 : 2 pp. 314*

[7] R. A. Fisher (1936). The Use of Multiple Measurements in Taxonomic Problems *Annals of Eugenics 7 (2): 179188*

[8] N. Vesdapunt, K. Bellare and D. Nilesh (2014) Crowdsourcing algorithms for entity resolution *Proceedings of the VLDB Endowment 7 (12)*