

A Bayesian Technique for Estimating the Credibility of Question Answerers

Byron Dom*
Yahoo! Inc,
Santa Clara, CA

Deepa Paranjpe
Yahoo! Inc,
Santa Clara, CA

Abstract

We address the problem of ranking question answerers according to their credibility, characterized here by the probability that a given question answerer (user) will be awarded a *best answer* on a question given the answerer's question-answering history. This probability (represented by θ) is considered to be a hidden variable that can only be estimated statistically from specific observations associated with the user, namely the number b of best answers awarded, associated with the number n of questions answered.

The more specific problem addressed is the potentially high degree of uncertainty associated with such credibility estimates when they are based on small numbers of answers. We address this problem by a kind of Bayesian smoothing. The credibility estimate will consist of a mixture of the overall population statistics and those of the specific user. The greater the number of questions asked, the greater will be the contribution of the specific user statistics relative to those of the overall population.

We use the *Predictive Stochastic Complexity* (PSC) as an accuracy measure to evaluate several methods that can be used for the estimation. We compare our technique (*Bayesian Smoothing* (BS)) with *maximum a priori* (MAP) estimation, *maximum likelihood* (ML) estimation and *Laplace smoothing*.

1 Introduction

Online communities that allow users to ask and answer questions have recently become very popular. Unlike automated question answering systems like *Ask*¹ or *Brainboost*², these online community portals provide a platform where users ask questions and other users in the community can answer those questions. *Yahoo! Answers*³, *Amazon's AskVille*⁴ and *Blurtit*⁵ are some examples of such online Q&A services. The growth of such services has been phenomenal (Yahoo! Answers, for example, has over 100 million users, about 220 million answers and more than 20 million questions). For such an online community, identifying reliable/credible users for expert detection or ranking becomes impor-

tant.

In Yahoo! Answers, a question asker chooses a best answer from the list of answers to his question. If the asker does not choose it, the best answer is chosen by voting. An incentive for users to answer questions or choose a best answer or to vote is given in the form of *points*. We are particularly interested in computing the answering credibility of users and thus the total points for a user is clearly not a reliable estimate of his(her) credibility. One way to characterize the answering credibility of users is by the proportion of their answers that are awarded *best answer* by the question asker or by voting by the community of Yahoo! Answers users. best answers to the total number of answers. However, For shorter answering histories, however, this estimate can be unreliable.

We characterise answering credibility by the probability that a given question answerer(user) will be awarded a best answer on a question in the category/set of questions being considered. This probability (represented by θ) is considered to be a hidden variable so that we can only estimate it statistically from specific observations associated with the user (the number of best answers awarded b associated with the number of questions answered n). The more specific problem that we address is the potentially high degree of uncertainty associated with such credibility estimates when they are based on small numbers of answers.

We address the problem by a kind of Bayesian smoothing where the credibility estimate consists of a mixture of the overall population statistics and those of the specific user. The greater the number of questions answered, the greater will be the contribution of the specific user statistics relative to those of the overall population. At the extremes:

- For $n = 0$ the only statistics we have are those of the overall population and those will dominate our estimate for small n .
- For large $n(n \rightarrow \infty)$ however our estimate will be completely dominated by the specific user statis-

*Contact Author

¹www.ask.com

²www.brainboost.com

³www.answers.yahoo.com

⁴www.askville.com

⁵www.blurtit.com

tics, making the overall population statistics irrelevant.

This technique can be applied to any set of questions and answers. An obvious example would be to apply it to a single category so that users would have category-specific credibility ratings.

2 Related Work

The problem of ranking objects (in general) has been of interest to many researchers. The authors of [2], [15] and [19] have developed techniques for object ranking for the purpose of better searching of web documents. The authors of [18] present an information flow modeling based on diffusion rate for prediction and ranking users in a social network. The authors of [10] and [11] have looked at the problem of ranking users in a Question Answering community such as *Yahoo! Answers*. Here the authors have presented an analysis of the link structure of a general-purpose question answering community to discover authoritative users. Their technique is inspired by the HITS web-page ranking technique. In [6] the authors apply several graph-based ranking algorithms to the problem of finding experts based on e-mail collections. Yet another application of graph based ranking algorithms for identifying experts in an E-learning platform is presented in [20] while the authors of [13] study social networks of researchers.

Our use of a mixture of Beta distributions as a prior is somewhat related to the works of [9] and [14]. The use of Bayesian inferencing in [12] and [7] is somewhat similar to the way we've used Bayesian inferencing for our problem. In [12], the authors use a Bayesian procedure for inference and prediction on event frequencies pointing out its advantages over ML based procedures. This paper is an extension of a previous paper by the first author (Lee and Savabala 1987), which describes a Bayesian approach with conjugate-type Beta-family priors for suitably transformed parameters in the Beta-Binomial. In the paper that we cite, the authors use numerical integration for the same. They applied this Bayesian estimation method in a real-world example of predicting viewership based on "loyalty" for new programs and showed it to be effective and the results to be better than the ones shown in their previous paper. The major contribution claimed by the authors comes by way of the numerical integration that they used for the Bayesian estimation.

In [7], the authors have addressed computational efficiency in the same problem and solution addressed in the paper by Lee and Lio (above). They do this

by providing closed-form Bayesian inference for the Beta-Binomial model. They approximate the likelihood by a polynomial expansion and use the well-known Pearson Type VI distribution. They also approximate the prior density by a polynomial expansion. These approximations can be made arbitrarily accurate with very little computational expense.

The work described in [5] assumes a Bernoulli/Binomial probability model like that assumed by us in this paper. In that work, however, it is assumed that the Bernoulli parameter, θ , may change one or more times over the length of the string. The assumed form is piecewise-constant θ . An MDL-based objective function is defined and used to determine the optimal segmentation of binary strings into regions of constant θ . That work addresses the general problem of segmenting binary strings, rather than any particular application. The approach used in that work is not explicitly Bayesian, being based on the Minimum Description Length (MDL) principle.

In [1], the authors present a framework for identifying high quality content within social networking forums such as Y! Answers. Their framework tries to combine evidence from different sources of information that can be tuned automatically for a given social media type and quality definition. This framework can be further enhanced using information about the quality of users i.e. user ranking that we propose in this paper.

3 Our Contributions

We believe the following to be our contributions in this paper.

1. A *prior* distribution, $\pi(\theta)$ that is formed by a kind of Bayesian averaging over the appropriate population of users. The details of this will come in the later sections.
2. The above-mentioned prior distribution is used in conjunction with the observed data (the number of questions answered and the number of best answers awarded) to obtain a posterior estimate of a user's credibility.

This general approach can be applied to many problems of this general form. The things that vary will be:

- The fundamental quantity being used as a measure. This corresponds to the best-answer probability θ in this example.
- The functional form of the probability distribution relating directly observable quantities (e.g. n_B

given n in this case) to the fundamental quantity (e.g. θ in this case.). In this case the probability distribution is the binomial distribution of 4.1. In this case the functional form is a direct theoretical consequence of the particular problem. In other problems there may be some latitude in choosing the probability distribution (its functional form) to be used.

- The functional form of the prior distribution ($\pi(\theta)$ in this case) of the fundamental quantity. In this example, the preferred form is the mixture of Beta distributions.
- As mentioned above the measure may be a function of the question category and possibly other question attributes.
- This analysis assumes a constant measure θ , whereas it will be a function of time - $\theta(t)$ - in general.

4 Description of our Approach

Our proposed measure is based on the following idealization. We assume that each Yahoo! Answers question answerer \mathbf{a} has an associated attribute $\theta_{\mathbf{a}}$, which is the probability that an answer given by \mathbf{a} will be chosen by the question asker as a best answer. We will assume that $\theta_{\mathbf{a}}$ is independent of other factors such as the category of the question, the asker and so on. In such a scenario, if \mathbf{a} answers n questions, the probability that b of those answers are chosen as best answers is given by the binomial distribution:

$$(4.1) \quad p(b|n, \theta) = \binom{n}{b} \theta^b (1 - \theta)^{n-b}$$

The figure of merit we use will be one of a number of possible estimates of or associated with $\theta_{\mathbf{a}}$. One example is the maximum-likelihood estimate $\hat{\theta}_{\mathbf{a}}$:

$$(4.2) \quad \hat{\theta}_{\mathbf{a}} \equiv \frac{b}{n}$$

This has the following associated problem, however. For small n the estimates, while being unbiased, are unreliable in the sense that their standard deviation σ is given by:

$$(4.3) \quad \sigma(\hat{\theta}_{\mathbf{a}}; n, \theta_{\mathbf{a}}) = \sqrt{\frac{\theta_{\mathbf{a}}(1 - \theta_{\mathbf{a}})}{n}}$$

and for small $\theta_{\mathbf{a}}$ there is a significant possibility of obtaining $b = 0$ and therefore $\hat{\theta}_{\mathbf{a}} = 0$.

To deal with this and to get better estimates we adopt the Bayesian statistical paradigm in which one assumes a *prior* distribution $\pi(\theta_{\mathbf{a}})$. Using this, we form the posterior distribution $p(\theta_{\mathbf{a}}|b, n)$, given by:

$$(4.4) \quad p(\theta_{\mathbf{a}}|b; n) = \frac{p(b|\theta_{\mathbf{a}}; n)\pi(\theta_{\mathbf{a}})}{p(b; n)}$$

5 Estimates of $\theta_{\mathbf{a}}$

Here we define four commonly used forms of parameter estimates $\hat{\theta}_{\mathbf{a}}$.

From this point on we will drop the subscript “ \mathbf{a} ” with the understanding that θ is associated with question answerer \mathbf{a} .

1. ML (*maximum likelihood*) estimate $\hat{\theta}_{ML}$:

$$(5.5) \quad \hat{\theta}_{ML} \triangleq \arg \max_{\theta} p(b|\theta; n)$$

In this particular problem in which the Binomial distribution applies:

$$(5.6) \quad \hat{\theta}_{ML} = \frac{b}{n}$$

2. MAP (*maximum a posteriori*) estimate $\hat{\theta}_{MAP}$:

$$(5.7) \quad \hat{\theta}_{MAP} \triangleq \arg \max_{\theta} p(\theta|b; n)$$

3. The formula (4.4) is for a probability density, which we can integrate over a finite interval to get a finite (non-infinitesimal) probability. For example:

$$(5.8) \quad \gamma = p(\theta > \theta_{\gamma}|b, n) \equiv \int_{\theta_{\gamma}}^1 p(\theta|b, n) d\theta,$$

where $p(\theta > \theta_{\gamma}|b, n)$ is the conditional probability that $\theta > \theta_{\gamma}$ given the observed values b and n .

4. Another quantity of potential interest that can be computed using (4.4) is the posterior expectation $\tilde{\theta}$ of θ .

$$(5.9) \quad \tilde{\theta} \triangleq E_n(\theta|b, n) \equiv \int_0^1 \theta p(\theta|b, n) d\theta$$

The fact that the parameter θ is interpreted as a probability means that $\tilde{\theta}$ is the probability that the next answer \mathbf{a} gives will be rated a best answer, given \mathbf{a} 's history (i.e. (n, b)) and assuming the prior $\pi(\theta)$.

Four possible answerer/expertise ratings (figures of merit) are these estimates of θ : $\hat{\theta}_{ML}$ of (5.6), $\hat{\theta}_{MAP}$ of (5.7), θ_γ of (5.8) and $\tilde{\theta}$ of (5.9). An obvious variation of this is to use similar statistics of the score s given the best answer by the answerer rather than using θ . The four rating measures are types of statistical estimates of θ . With the exception of (5.6), these estimates are formulated within the Bayesian statistical paradigm. The third of these, θ_γ , is a kind of *interval estimate* where the upper limit of the interval is $\theta = 1$. That is, we are saying that θ lies within the interval $[\theta_\gamma, 1]$ with probability γ . This is an instance of the Bayesian concept of a *credible interval* (CI), which is the counterpart of a *confidence interval* in classical (frequentist) statistics.

6 Estimating the prior $\pi(\theta)$

6.1 A uniform prior: $\pi(\theta) = 1$ A common practice in estimation problems such as this is to assume a uniform prior: $\pi(\theta) = 1$. This is taken to correspond to no a priori knowledge about the distribution of θ values. In this special case the joint distribution $p(b, \theta|n)$ is equal to the conditional distribution $p(\theta|b, n)$, namely:

$$p(b, \theta|n) = p(\theta|b, n) = \binom{n}{b} \theta^b (1 - \theta)^{n-b}$$

The marginal distribution $p(b|n)$ is obtained by integrating $p(b, \theta|n)$ over θ , which in this case yields:

$$p(b|n) = \binom{n}{b} B(b+1, n-b+1),$$

where the denominator is the *Beta function* and the posterior $p(\theta|b, n)$ is given by:

$$(6.10) \quad p(\theta|b, n) = \frac{p(b, \theta|n)}{p(b|n)} = \frac{\theta^b (1 - \theta)^{n-b}}{B(b+1, n-b+1)},$$

which is a *Beta* distribution with parameters $(b+1)$ and $(n-b+1)$.

6.2 A mixture of Beta distributions In this situation (ranking question answerers on Yahoo! Answers) we can do better than assuming a uniform prior. We can construct a prior based on user data. To get a closed form for our prior, we might try using the Beta distribution as our prior, or we might try fitting a mixture of Beta distributions. The Beta distribution is *conjugate* to the Binomial and therefore gives a closed form when integrated with it.

Using a single Beta distribution, our prior would thus be:

$$(6.11) \quad \pi(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

where α and β are parameters that must be greater than zero and $B(\alpha, \beta)$ is the *Beta* function.

Another convenient form for representing our prior is a mixture of Beta distributions, which is obviously of the form:

$$(6.12) \quad \pi(\theta; \alpha, \beta, \eta) = \sum_{i=1}^m \frac{\eta_i}{B(\alpha_i, \beta_i)} \theta^{\alpha_i-1} (1 - \theta)^{\beta_i-1},$$

where $\eta = \{\eta_i | i = 1, 2, \dots, m\}$ are the *mixing coefficients* and here α and β represent the parameter vectors $\{\alpha_i | i = 1, 2, \dots, m\}$ and $\{\beta_i | i = 1, 2, \dots, m\}$.

The forms (6.11) and (6.12) can be fit to user (b, n) data by various methods.

6.3 Prior construction by smoothing user data

In this technique we estimate the “true” population distribution $p(\theta)$ by smoothing the raw $\hat{\theta} = \frac{b}{n}$ data using a posterior distribution $p(\theta|b, n)$ constructed assuming a uniform prior $\pi(\theta) = 1$. This method is similar to *kernel density estimation* [See [17] for example.] except that the (b, n) -specific posterior is used in place of a single *kernel* for all points.

For every user (represented by (b, n)) we compute the posterior $p(\theta|b, n)$. Then we sum up all of these posteriors (for all of the users) to get an estimate of the overall population distribution $p(\theta)$. From (4.4) we know that if we assume a uniform prior, then an observation of (b, n) implies a posterior that is the following Beta distribution

$$(6.13) \quad P(\theta|b, n) = \frac{\theta^b (1 - \theta)^{n-b}}{B(b+1, n-b+1)}$$

In this application we have such data on the entire set of users. For those users, we will use the following notation:

1. ν_i : The number of questions answered by user/answerer \mathbf{a}_i .
2. γ_i : The number of best answers received by user/answerer \mathbf{a}_i .

For the i^{th} user, the resulting posterior is of the same form as (6.13). That is:

$$P(\theta|\gamma_i, \nu_i) = \frac{\theta^{\gamma_i} (1 - \theta)^{\nu_i - \gamma_i}}{B(\gamma_i + 1, \nu_i - \gamma_i + 1)}$$

Using this,

$$\pi(\theta) = \frac{1}{N} \sum_{i=1}^N \frac{\theta^{\gamma_i} (1-\theta)^{\nu_i - \gamma_i}}{B(\gamma_i + 1, \nu_i - \gamma_i + 1)}$$

For a given user for which we have the observation (b, n) , using this prior gives the following posterior:

$$(6.14) \quad P(\theta|b, n) = \frac{1}{N} \frac{1}{B(b+1, n-b+1)} \sum_{i=1}^N \frac{\theta^{b+\gamma_i} (1-\theta)^{n+\nu_i-(b+\gamma_i)}}{B(\gamma_i + 1, \nu_i - \gamma_i + 1)}$$

It is worth noting that (6.14) is a special case of (6.12), a Beta mixture.

6.4 Prior-estimation approach used in this work

In the work described here the following combination of the techniques described in Sections 6.2 and 6.3 was used.

1. The (b, n) data from the complete population were used in the smoothing approach described in Section 6.3 to construct a prior.
2. A two-component Beta mixture as described in Section 6.2 was fit to the smoothed prior. This was done for tractability reasons using a variation of the EM algorithm-based technique used in [9, 8].

The results of applying the first of these operations and of applying both of them in sequence are shown in Figure 1 with a histogram of raw $\hat{\theta}_{ML}$ ((b/n) values) from user data. The fitted two-component Beta-mixture distribution was used as the prior in performing all the calculations reported here. This is a special case of (6.12), with the following parameter values: $\alpha_1 = 1.4856$, $\beta_1 = 12.9861$, $\alpha_2 = 1.5956$, $\beta_2 = 2.3595$, $\eta_1 = 0.4873$, and $\eta_2 = 0.5127$.

The associated distribution of values of n (the total number of questions answered) for the total Yahoo!-Answers user population is plotted in Figure 2 and the average values of $\hat{\theta}_{ML}$ for all observed values of n are plotted in Figure 3

7 Maximum *A Posteriori* Estimation of θ

7.1 MAP Estimate for the Beta-Binomial Distribution The *maximum a posteriori* estimate for θ is defined as:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(\theta|b, n) = \arg \max_{\theta} \frac{p(b, \theta|n)}{p(b|n)} \quad (7.15)$$

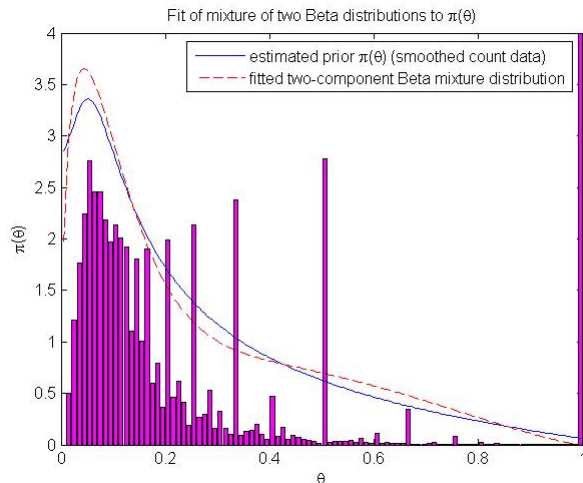


Figure 1: The *prior* obtained by smoothing raw count data plus a two-component Beta mixture distribution fitted to that smoothed count data. The bar graph is a 100-bin histogram of raw $\hat{\theta}_{ML} = (b/n)$ values from user data. The bar-graph values were scaled to fit on the plot with other continuous distributions. The first (smallest θ : $\theta = 0$) histogram bin is omitted because it would overwhelm the other bin values in the plot. The large spikes in the bar graph correspond to the small number of possible values of $\hat{\theta}_{ML}$ for small n (i.e. simple rational numbers like $1/2$, $1/3$, $2/3$ and so on).

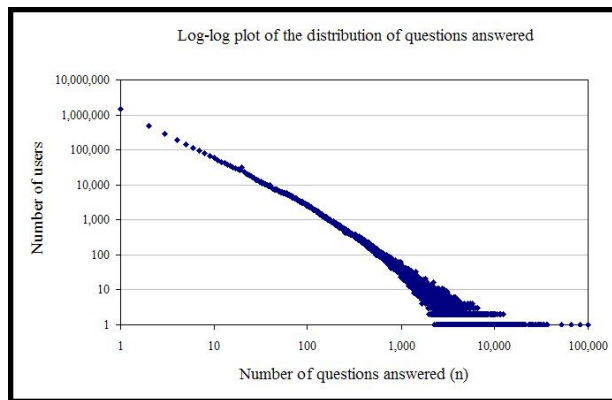


Figure 2: A plot showing the distribution of values of n (the total number of questions answered), which exhibits the power-law-like behavior common in such distributions.

We thus find $\hat{\theta}_{MAP}$ by setting the derivative of this to zero and solving for θ :

$$\frac{dp(\theta|b, n)}{d\theta} = 0$$

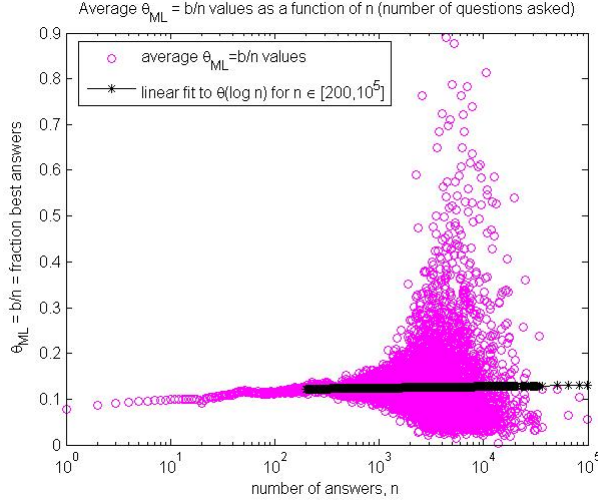


Figure 3: A plot of the average value of $\hat{\theta}_{ML}$ for every value of n . The high variability of $\hat{\theta}_{ML}$ at large n is consistent with the distribution $p(n)$ shown in Figure 2.

Note that $p(\theta|b, n) = p(\theta, b|n)/p(b|n)$. Using this relationship and multiplying (7.15) through by $p(b|n)$ yields:

$$\frac{\partial p(b, \theta|n)}{\partial \theta} = 0$$

Substituting the binomial distribution of (4.1) for $p(b|\theta, n)$ and assuming a prior $\pi(\theta)$ that is a single Beta distribution (i.e. (6.11)) with parameters α and β , we obtain:

$$\hat{\theta}_{MAP} = \frac{(b + \alpha - 1)}{(n + \alpha + \beta - 2)}.$$

7.2 MAP Estimate for the Beta-Binomial Mixture Distribution To find $\hat{\theta}_{MAP}$ for our Beta mixture, as in the previous section, we need to solve the following:

$$(7.16) \quad \frac{\partial p(b, \theta|n)}{\partial \theta} = 0$$

In this case however, $p(b, \theta|n)$ is a mixture distribution of the form of (6.12).

Finding the MAP estimate for such a mixture of Beta-Binomials must be done numerically. In this work we used a simple Newton's-method iteration, which for this problem, leads to the following update equation:

$$(7.17) \quad \hat{\theta}_{t+1} = \hat{\theta}_t - \left(\frac{\partial p(x^n, \theta)}{\partial \theta} / \frac{\partial^2 p(x^n, \theta)}{\partial \theta^2} \right) \Big|_t$$

8 Evaluation: The Accuracy of Estimates of θ

In this discussion θ will represent the *true* value of the credibility parameter and $\hat{\theta}$ will stand for any of a number of possible estimates of θ based on n and b .

The primary goal of our technique is to mitigate the deleterious effects of sampling error when estimating θ from (n, b) count data, especially for small n . Estimates of θ might be applied in various ways, but the key point is that it is a measure of how much a user (question asker) should trust a new answer given by the associated user/answerer.

8.1 Predictive Stochastic Complexity An appropriate means of evaluating the results of these estimation techniques would follow the principle of *Prequential Probability*[3, 4], which uses measures that depend on how well a probability model predicts the next outcomes in sequences, based on the past elements of those sequences. In the application we are considering the answering histories of users are just such sequences. In fact we can treat them as binary sequences where the 0's represent answers that didn't receive a *best-answer* award from the asker and the 1's represent answers that did receive a best answer award.

Let x be a binary variable corresponding to a user's answer to a question. It takes a value only after the question has been resolved:

$$\begin{aligned} x = 1 &\Rightarrow \text{answer selected as best answer} \\ x = 0 &\Rightarrow \text{answer not selected as best answer} \end{aligned} \quad (8.18)$$

When considering the answering history of a particular user, let x_k represent the best-answer status of the user's k^{th} answer. The sequence $\{x_k | k = 1, 2, \dots, n\}$ thus represents the user's answer/best-answer history. We represent this sequence by x^n and for $k < n$, x^k represents the length- k prefix subsequence starting with x_1 . Using this notation, the prequential probability we need can be written as:

$$p(x_k | x^{k-1}) = \frac{p(x_k, x^{k-1})}{p(x^{k-1})} = \frac{p(x^k)}{p(x^{k-1})},$$

where obviously $p(x_k, x^{k-1}) \equiv p(x^k)$, being merely two notational representations of the same thing.

The expression for $p(x^{k-1})$ in terms of the prior $\pi(\theta)$ is:

$$(8.19) \quad p(x^{k-1}) = \int_0^1 p(x^{k-1} | \theta) \pi(\theta) d\theta$$

In this $p(x^{k-1}|\theta)$ is the *Bernoulli sampling* distribution:

$$(8.20) \quad p(x^{k-1}|\theta) = \theta^{b_{k-1}} (1-\theta)^{k-1-b_{k-1}},$$

where b_k is the number of best answers in the first k answers of the user's answering history. From the definition (8.18) we can see that:

$$b_{k-1} \equiv \sum_{j=1}^{k-1} x_j$$

Because the Bernoulli parameter θ of (8.20) is itself a probability, when $x_k = 1$, $p(x_k|x^{k-1})$ turns out to be equivalent to the posterior expectation θ_{k-1} proposed in our original treatment as one possible estimate of θ . I.e.:

$$(8.21) \quad \tilde{\theta}_{k-1} \triangleq E_n(\theta|b_{k-1}, k-1) \equiv \int_0^1 \theta p(\theta|b_{k-1}, k-1) d\theta$$

If $x_k = 0$, on the other hand,

$$p(x_k|x^{k-1}) = 1 - \tilde{\theta}_{k-1}$$

For this discussion we temporarily assume that the prior $\pi(\theta)$ is a *Beta distribution*:

$$(8.22) \quad \pi(\theta|\alpha, \beta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta)$ is the *Beta function*. Substituting (8.20) and (8.22) into (8.19) yields:

$$(8.23) \quad \begin{aligned} p(x^{k-1}) &= \frac{\int_0^1 \theta^{b_{k-1}+\alpha-1} (1-\theta)^{(k-1)-b_{k-1}+\beta-1} d\theta}{B(\alpha, \beta)} \\ &= \frac{B(b_{k-1} + \alpha, (k-1) - b_{k-1} + \beta)}{B(\alpha, \beta)} \end{aligned}$$

Likewise, we have:

$$(8.24) \quad p(x^k) = \frac{B(b_k + \alpha, k - b_k + \beta)}{B(\alpha, \beta)},$$

which means that:

$$(8.25) \quad \begin{aligned} p(x_k|x^{k-1}) &\equiv \frac{p(x^k)}{p(x^{k-1})} \\ &= \frac{B(b_k + \alpha, k - b_k + \beta)}{B(b_{k-1} + \alpha, (k-1) - b_{k-1} + \beta)} \end{aligned}$$

Combining (8.23), (8.24) and (8.25) we obtain:

$$\begin{aligned} p(x_k = 0|x^{k-1}) &= \frac{[(k-1) - b_{k-1} + \beta]}{(k-1 + \alpha + \beta)} \\ p(x_k = 1|x^{k-1}) &= \frac{(b_{k-1} + \alpha)}{(k-1 + \alpha + \beta)} \end{aligned}$$

The *Predictive Stochastic Complexity*[16] of the sequence x^n is defined as:

$$(8.26) \quad \mathcal{S}(x^n) \triangleq - \sum_{k=1}^n \log p(x_k|x^{k-1})$$

While not obvious from this notation, this is defined relative to the particular probability model used to construct $p(x_k|x^{k-1})$. In this case that model is composed of the Bernoulli sampling distribution of (8.20) and the prior of (8.22).

It is straightforward to generalize (8.25) to a prior that is the following mixture of Beta distributions.

$$(8.27) \quad \pi(\theta; \alpha, \beta, \eta) = \sum_{i=1}^m \frac{\eta_i}{B(\alpha_i, \beta_i)} \theta^{\alpha_i-1} (1-\theta)^{\beta_i-1},$$

In this case of the Beta-mixture prior (8.27) generalizes (8.25) to:

$$(8.28) \quad p(x_k|x^{k-1}) \equiv \frac{\sum_{i=1}^m \frac{\eta_i}{B(\alpha_i, \beta_i)} B(b_k + \alpha_i, k - b_k + \beta_i)}{\sum_{i=1}^m \frac{\eta_i}{B(\alpha_i, \beta_i)} B(b_{k-1} + \alpha_i, (k-1) - b_{k-1} + \beta_i)}$$

8.2 Notes

- 1. Estimation Error in $\hat{\theta}$:** While our primary means of evaluation will be based on $\mathcal{S}(x^n)$ (PSC), defined in (8.26), we will also compute and present simply the error in predicting θ , i.e. $(\hat{\theta} - \theta)$ and $|\hat{\theta} - \theta|$. The absolute difference is more meaningful as an accuracy measure because positive and negative errors can cancel each other in the average signed difference. We include it, however, as a measure of bias. Rather than absolute difference, one might also use the RMS error.
- 2. Why $\hat{\theta}_{ML}$ cannot be used in computing $\mathcal{S}(x^n)$ (PSC):** The *maximum likelihood* (ML) estimate $\hat{\theta}_{ML}$ can't be evaluated using PSC for two reasons.
 - (a) The estimate for $n = 0$, which would be used to encode x_1 in the PSC measure, is undefined.
 - (b) Even if an arbitrary value such as $\hat{\theta}_{ML} = 0.5$ were used to encode x_1 , all histories will eventually result in an infinite code length because, when the prior history consists of all one value - either 0 or 1 - the corresponding estimate $\hat{\theta}_{ML}$ will equal either 0 or 1 and when

the first instance of the other value (1 or 0) is encountered, the associated code length will be $\log 0 = \infty$.

The estimation errors $(\hat{\theta}_{ML} - \theta)$ and $|\hat{\theta}_{ML} - \theta|$ will be computed and presented, however. In those results no value is required for $n = 0$. We simply start at $n = 1$ and the two possible values of $\hat{\theta}_{ML}$ at $n = 1$ are 0 and 1.

9 Experiments

We performed two nearly identical evaluation experiments on the following two data sets.

1. On synthetic data that matches the model assumption of a constant θ for each user and for which the θ values are drawn from $\pi(\theta)$ the two-component Beta mixture that was fit to user data as described in Section 6.4.
2. On real user data from the population used to construct the prior distribution $\pi(\theta)$ that we used.

9.1 Evaluation on Synthetic Data In this experiment we performed an evaluation by randomly drawing sequences x^n (answer/best-answer histories) from our population, computing $\mathcal{S}(x^n)$, summing all these \mathcal{S} values corresponding to our sample of sequences and computing the average of these $\bar{\mathcal{S}}$. We assume an infinite population described by our model ((8.20) & (6.11) or (6.12)) and use it as the generation model in our experiment.

1. Repeat the following N times, where N is an adequate sample size.
 - (a) Draw a value of n_i from $p(n)$.
 - (b) Draw a value of θ_i from $\pi(\theta)$.
 - (c) Draw a sequence $x_i^{n_i}$ from $p(x^{n_i}|\theta_i, n_i)$.
 - (d) Compute and accumulate $\mathcal{S}(x^{n_i})$.

2. Compute $\bar{\mathcal{S}}_N$:

$$(9.29) \quad \bar{\mathcal{S}}_N = \frac{1}{N} \sum_{i=1}^N \mathcal{S}(x_i^{n_i})$$

where $p(x_k|x^{k-1})$ is given by (8.28) and $\mathcal{S}(x^n)$ is given by (8.26). If the true value of θ were known, by *Shannon's first theorem* we know that the minimum expected code length would be obtained by encoding every "1" in x^n with $\log \theta$ bits and every "0" with

$\log(1 - \theta)$ bits. That is the string x^n would simply be encoded using the Bernoulli distribution:

$$p(x^n) = \theta^{b_n}(1 - \theta)^{n - b_n}$$

and its code length would be:

$$-\log p(x^n) = b_n \log \theta + (n - b_n) \log(1 - \theta).$$

9.1.1 Experimental Results on Synthetic Data

The results of our experiments on synthetic data are shown in Figures 4 and 5.

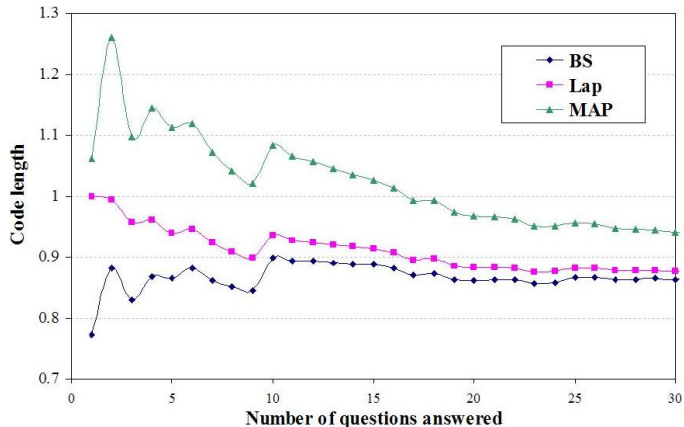


Figure 4: Prequential analysis of a sample of 1000 histories generated randomly using θ values drawn randomly from our two-component Beta mixture prior. Shown here: average Incremental (per question) Predictive Stochastic Complexity (PSC) vs. question number. The reason for the absence of $\hat{\theta}_{ML}$ here is given in Section 8.2.

In both Predictive Stochastic Complexity (PSC) and in predicting the true θ value the posterior-expectation estimate (represented by either $\hat{\theta}$ or $p(x_{k+1}|x^k)$ and labeled "BS" (for "Bayesian Smoothing") in the figures) did best. The MAP estimate was second best in PSC, but Laplace's rule (posterior expectation using a uniform prior) did second best at predicting the true θ value. Accuracy at predicting the true θ value is judged by the average absolute-difference results shown in figure 5-a. The signed-difference results are presented to show the bias in the various estimators. The maximum-likelihood (ML) estimator $\hat{\theta}_{ML}$ is known to be unbiased and this is verified in Figure 5-b, but in figure 5-a it can be seen that it is the worst in terms of absolute error.

9.2 Evaluation on Real User Data The experimental evaluation for user data is almost exactly the

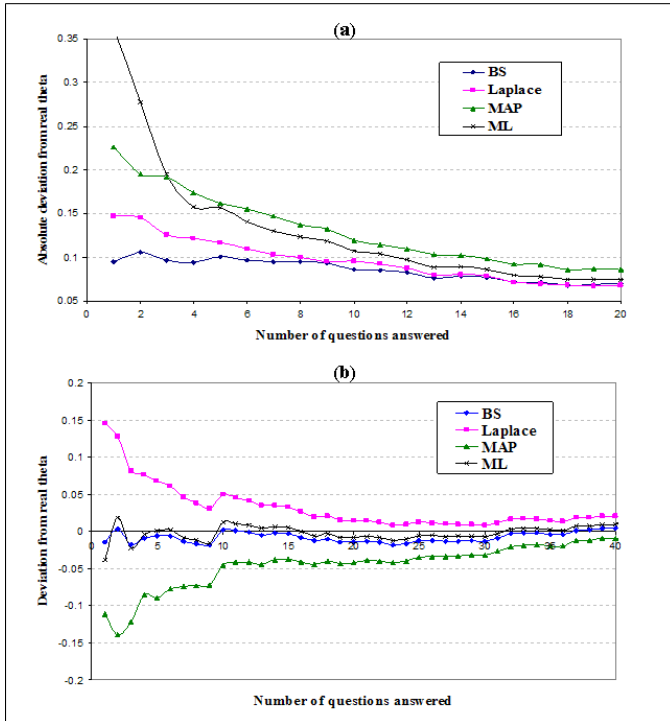


Figure 5: Prequential analysis of a sample of 1000 histories generated randomly using θ values drawn randomly from our two-component Beta mixture prior. Shown here: (a) absolute average difference, (b) average difference between $\hat{\theta}$ and θ .

same as for the synthetic data. A sample of 1000 user answering histories was drawn randomly from the total user population. Once drawn these we analyzed in exactly the same way as the synthetic histories. The results of our experiments on actual user data are shown in Figures 6 and 7.

The relative results on user data are almost exactly the same as on synthetic data. Once again, in both PSC and in predicting the true θ value the posterior-expectation estimate did best. In this experiment, however, the MAP estimate was second best in both PSC and in predicting θ . Another difference is that Laplace's rule did worst at predicting the true θ value, beaten even by $\hat{\theta}_{ML}$.

It is clear from these results, that of the estimation techniques we examined, the posterior expectation did best.

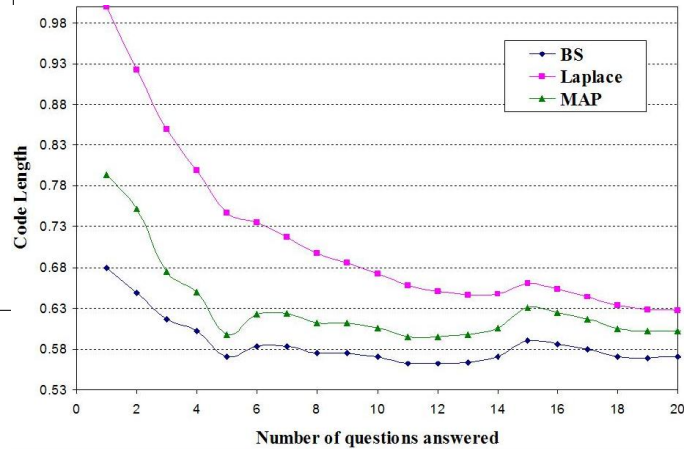


Figure 6: Prequential analysis of a sample of 1000 user histories drawn uniformly randomly from the total user population. Shown here: average Incremental (per question) Predictive Stochastic Complexity (PSC) vs. question number. The reason for the absence of $\hat{\theta}_{ML}$ here is given in Section 8.2.

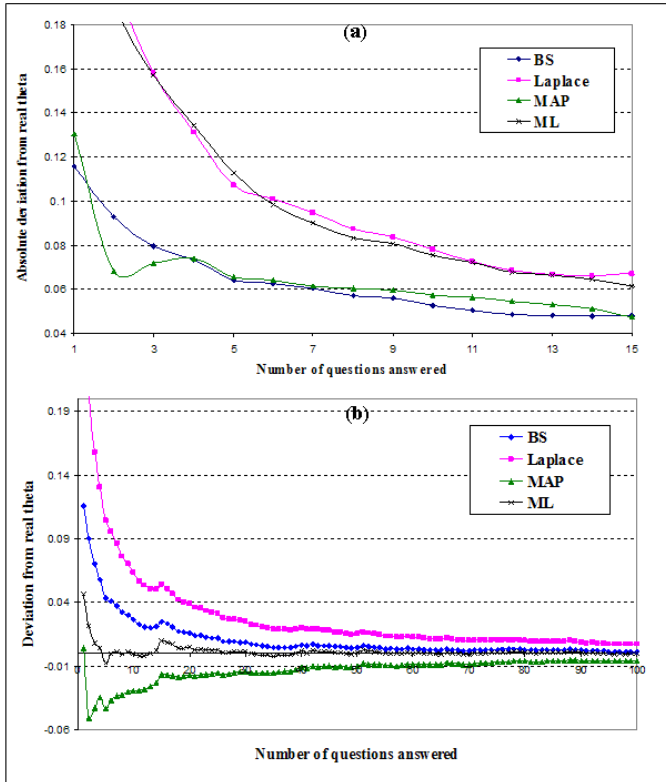


Figure 7: Prequential analysis of a sample of 1000 user histories drawn uniformly randomly from the total user population. Shown here: (a) average absolute difference, (b) average difference between $\hat{\theta}$ and θ .

References

- [1] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the First ACM International Conference on Web Search and Data Mining (WSDM 2008)*, Stanford University, Stanford, California, USA, February 2008. Association for Computing Machinery (ACM). Accepted.
- [2] Le Chen, Lei Zhang, Feng Jing, Ke-Feng Deng, and Wei-Ying Ma. Ranking web objects from multiple communities. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 377–386, New York, NY, USA, 2006. ACM Press.
- [3] A. Dawid. Prequential analysis, stochastic complexity, and Bayesian inference. In J. Bernardo, J. Berger, A. David, and F. Smith, editors, *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, pages 1019–1125, Oxford, U.K., 1992. Oxford Univ. Press.
- [4] A. Philip Dawid and Vladimir G. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5(1):125–162, 1999.
- [5] Byron Dom. MDL estimation with small sample sizes including an application to the problem of segmenting binary strings using Bernoulli models. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, June 1997. longer version: IBM Research Report RJ 9997 (89085).
- [6] Byron Dom, Iris Eiron, Alex Cozzi, and Yi Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 42–48, June 2003.
- [7] Philip J. Everson and Eric T. Bradlow. Bayesian inference for the Beta-Binomial distribution via polynomial expansions. *Journal of Computational & Graphical Statistics*, 11(1):202–207, March 2002.
- [8] Yuan Ji. private communication, November 2006.
- [9] Yuan Ji, Chunlei Wu, Ping Liu, Jing Wang, and Kevin R. Coombes. Applications of Beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, May 2005.
- [10] Pawel Jurczyk and Eugene Agichtein. Discovering authorities in question answer communities using link analysis. In *ACM Conference on Information and Knowledge Management (CIKM)*. ACM Press, 2007.
- [11] Pawel Jurczyk and Eugene Agichtein. Hits on question answer portals: exploration of link analysis for author ranking. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 845–846, New York, NY, USA, 2007. ACM Press.
- [12] Jack C. Lee and Y. L. Lio. A note on Bayesian estimation and prediction for the beta-binomial model. *Journal of Statistical Computation and Simulation*, 1997.
- [13] Juanzi Li, Jie Tang, Jing Zhang, Qiong Luo, Yunhao Liu, and Mingcai Hong. Eos: expertise oriented search using social networks. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1271–1272, 2007.
- [14] S. Lowe. The beta-binomial mixture model for word frequencies in documents with applications to information retrieval. In *Proceedings of Eurospeech-99*, volume 6, pages 2443–2446, September 1999.
- [15] Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen, and Wei-Ying Ma. Object-level ranking: Bringing order to web objects, 2005.
- [16] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *World Scientific Series in Computer Science*. World Scientific, Singapore, 1989.
- [17] David W. Scott. *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley-Interscience, 1992. ISBN 0-471-54770-0.
- [18] Xiaodan Song, Yun Chi, Koji Hino, and Belle L. Tseng. Information flow modeling based on diffusion rate for prediction and ranking. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 191–200, 2007.
- [19] Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma, and Liu Wenyin. Ranking user’s relevance to a topic through link analysis on web logs, 2002.
- [20] Wei Wei, Jimmy Lee, and Irwin King. Measuring credibility of users in an e-learning environment. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1279–1280, 2007.