Reconciliation of Categorical Opinions from Multiple Sources

Adway Mitra
Computer Science and Automation
Indian Institute of Science
Bangalore, India
adway.cse@gmail.com

Srujana Merugu
Amazon SDC
Bangalore, India
srujana@gmail.com

ABSTRACT

Reconciling opinions from multiple sources on questions of interest to determine the correct answers is an important problem encountered in collaborative information systems such as Q & A forums and prediction markets. Our current work focuses on a widely applicable variant of the above problem where the opinions and answers are categorical-valued with the set of values possibly varying across questions. Most of the existing techniques are tailored only for binary opinions and cannot be effectively adapted for questions with categorical opinions. To address this, we propose a generic Bayesian framework for opinion reconciliation that can readily incorporate latent and observed attributes of sources and subjects. For the scenario of interest, we derive three specific model instantiations of the general approach (CTM, CTM-OSF, CTM-LSG), which respectively capture the latent source behavior, variations of source behavior across subject groups, and inter-source correlations. Empirical results on real-world datasets point to the relative superiority of the proposed models over existing baselines.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: [Learning]

General Terms

Design, Experimentation

Keywords

Opinion mining; Graphical models; Gibbs sampling

1. INTRODUCTION

Rapid advances in internet technologies have made it increasingly easy to solicit and exchange information resulting in the creation of massive collaborative information systems powered by user-generated content. Examples include online Q & A forums (e.g., Quora), prediction markets (e.g., Intrade), online diagnostic systems(e.g., interactiveMD), wiki-compilations (e.g., Wikipedia, Wikimapia). Harnessing this "wisdom of the crowd" requires an effective solution for integrating sparse opinions from multiple unreliable, and possibly malicious sources.

In a collaborative information system, there are multiple sources offering opinions on various subjects or questions of interest. Unlike a subjective question like "who is the best US president ever?", there is a unique correct answer for an objective question like "who won the 2012 US Presidential election?". In such scenarios, the goal of opinion integration is to determine this correct answer. Our current work focuses on such *Opinion Reconciliation (OR)*, where each question is associated with a unique correct answer.

Let $\{U_1, \cdots, U_i, \cdots, U_{N_u}\}$ and $\{S_1, \cdots, S_j, \cdots, S_{N_s}\}$ denote the multiple sources and subjects (questions) in the information system respectively. Each question S_j is associated with a single correct answer M_j and one or more opinions $\{O_{ij}\}_{i,j}$ from a subset of the sources $\{U_i\}_i$. Often, one might also have access to attributes of sources and subjects, denoted by X_i and Y_j respectively. *The opinion reconciliation (OR) problem can then be stated as follows: Given opinions $\{O_{ij}\}_{i,j}$, limited (or even none) observations on the correct answers $\{M_j\}_j$, source attributes $\{X_i\}_i$, and subject attributes $\{Y_j\}_j$, predict the unknown correct answer $M_j,\ \forall j,\ [j]_1^{N_s}$.*

There can be multiple variants of the above problem depending on the nature of the opinions and the correct answers, which could be real-valued (e.g., US GDP in dollars), ordinal (e.g., US credit rating), binary (e.g., Is US a monarchy?), categorical from a small known set (e.g., type of governance in US) , text phrases from a given vocabulary (e.g., US national anthem), set-valued (e.g., list of US presidents from New York), etc. Among all the variants, a particularly important formulation is the one where the space of possible answers and opinions for a question S_j comprises of a large number of categorical values or text phrases $\mathcal{O}_j$, which could vary across questions. Applications include information systems with heterogeneous questions permitting factoid style answers, which is fairly common in prediction markets, specialized Q & A forums, and diagnostic systems. Table 1 shows a toy example of such a scenario with multiple human sources, heterogeneous questions on various topics and opinions corresponding to text phrases. Existing approaches to opinion reconciliation can be broadly grouped as :

Axiomatic approaches. These include simple non-data-driven techniques assuming inter-source and inter-subject independence where the correct answer of a question is obtained in terms of various characteristics (mode, mean, median) of the distribution over the corresponding opinions.

Discriminative Meta-learning. These approaches involve learning a functional mapping (e.g., linear model or decision tree) from the opinions and features of subjects and

CIKM '13 San Francisco, California, USA

http://dx.doi.org/10.1145/2505515.2507844 .

sources to the correct answer through supervision. But this form of supervised learning is not effective when the opinions are very sparse and supervision is highly limited.

Trust Propagation. There is a large body of work on trust propagation over graphs that allow one to estimate the "reputation" (e.g., trustworthiness or correctness) of sources (represented by nodes). Truthfinder algorithm [5] adapts these ideas to OR by considering a bipartite graph over facts (subject-opinion pairs) and sources, with an edge corresponding to a positive assertion on a fact by a source (missing edges are negative assertions). The likelihood of a source making true assertions and the probability of a fact being true are iteratively estimated in terms of each other, but the algorithm is not guaranteed to converge.

Bayesian Models Galland et al. [2] propose a generative model for binary opinions based on source-specific probability of error and not-opinionating as well as subject-specific probability of error and not-opinionating, with the opinion by a source on a subject being determined by the interaction of these parameters. Unfortunately, the algorithms proposed in the paper are based on heuristics not related to the generative model and are not guaranteed to converge. Rayakar et al. [4] and Zhao et al. [7] both propose similar generative models for binary opinions that take into account source-specific probabilities of Type-1 and Type-2 errors, but use different inference approaches. Both works have also been extended to real-valued opinions [6]. A more advanced approach is the Multi-Source Sensing (MSS) model [3], which considers latent groups of sources and models error probabilities as property of each source group.

However, most of the existing techniques are focused on the scenarios where the correct answer M_j and the opinions O_{ij} are binary variables [7, 5, 2]. To some extent, these techniques can be adapted to handle other scenarios involving categorical, textual and set-valued answers by transforming the original subject space to one that permits binary opinions/answers. For example, the question "What is the type of US government?" with possible answers {"democracy", "monarchy", "oligarchy"} can be alternately represented as three questions: {"Is US a democracy?", "Is US a monarchy?", "Is US an oligarchy?" }, each of which permits a binary yes/no answer independently. However, these techniques cannot effectively exploit the implicit mutual exclusivity in the categorical valued variables.

This work is an attempt to directly address the opinion reconciliation problem for heterogeneous questions with a large number of categorical opinions which requires taking into account various practical issues shown in Table 1:

Variations in source behavior. In Table 1, we observe that Mr. A is more accurate than other users, indicating that majority vote may not suffice, and source expertise and reliability needs to be accounted for.

Variations in same source's expertise across topics. In the example, Prof. C is an expert in math and chemistry, but ignorant in history pointing to the need for topic-specific modeling of expertise.

Highly limited supervision. The correct answer is often known only for a small subset of questions and one needs to use this supervision to calibrate the expertise of sources.

Opinion Sparsity. Each source provides opinions on only a small subset of questions, which makes it difficult to employ traditional meta-learning techniques.

Textual Variations. Certain opinions are minor variations of the correct answer (eg. typos), which need to be treated differently from an outright wrong answer.

To address some of the above challenges, we propose a Bayesian framework for opinion generation that jointly models the source behavior and subject-specific correct answers as latent variables and make the following contributions:

1) A generic framework for opinion reconciliation via Bayesian modeling that can incorporate latent and observed attributes of sources and subjects. Existing Bayesian approaches such as the LTM [7] can be shown to be special cases.

2) To reconcile categorical-valued opinions, we propose three instantiations (CTM, CTM-OSF, CTM-LSG) of the generic approach to capture the latent source behavior, variations across subject groups, and inter-source correlations.

3) Empirical evaluation of predictive performances of the proposed models and multiple baselines on real-world datasets, which points to the relative efficacy of the proposed models.

Question	User	Opinion
Materials used in namesake displays in computer monitors and HDTVs Answer: (known) liquid crystal Category: chemistry	Mr. A Ms. B Prof. C Mr. D	lcd liqiud crystal liquid crystal cathode ray
Found by solving $det(A - I) = 0$ Answer:(unknown) eigenvalues Category: math	Ms. B Mr. F Prof. C	prime prime numbers eigenvalues
Architect of the first theory of commmunism Answer:(unknown) Karl Marx Category: history	Ms. B. Prof. C. Mr. A	Karl Marx Lenin Marx

Table 1: Toy example of categorical opinions

2. SOLUTION APPROACH

In this section, we describe our generic Bayesian framework for opinion reconciliation, and present three models for categorical opinions. The graphical model encodes two assumptions: (a) Opinions $\{O_{ij}\}_{ij}$ are independent of each other given $\{M_j, X_i, Y_j\}$, (b) Dependencies among sources and subjects are captured entirely in terms of X_i and Y_j.

2.1 Generic Bayesian Opinion Reconciliation

The dependencies between the opinions of a source and the correct answer can be succinctly expressed in terms of other variables of interest, such as the source expertise, which are often unobserved or partially observed. An effective solution strategy is to simultaneously infer these latent variables as well as the precise form of the dependencies, in addition to inferring the primary target variable, the correct answer of a subject M_j. A natural mechanism is to encode the dependencies between the different variables of interest (O_{ij}, M_j, X_i, Y_j) in the form of a joint probability distribution that can be factored into conditional distributions amenable for learning. Figure 1 shows a graphical model corresponding to such a factorization. Here X_i^{lat} and X_i^{obs} denote the latent and the observed features of source U_i such that $X_i = [X_i^{lat}, X_i^{obs}]$. Similarly, Y_j^{lat} and Y_j^{obs} denote the latent and observed features of source S_j such that $Y_j = [Y_j^{lat}, Y_j^{obs}]$. The priors allow one to encode domain knowledge as well as data constraints.

The above framework provides an elegant way to model some of common factors relevant to opinion generation such as source expertise, source bias, difficulty of a question, inter-source correlations. Though Figure 1 depicts a specific directionality for the dependence between latent and observed source and subject attributes, e.g., between Y_j^{lat} and Y_j^{obs}, the appropriate directionality depends on which of the conditional probabilities (e.g., $p(Y_j^{lat}|Y_j^{obs})$ or $p(Y_j^{obs}|Y_j^{lat})$) is more learnable given the nature of the variables.

2.2 Categorical Truth Model & variants

We now consider the specific scenario where M_j, O_{ij} are both categorical values with support sets $\mathcal{M}$ and $\mathcal{O}$ respectively. We propose three different models, where the conditional probability $p(O_{ij}|M_j, X_i, Y_j)$ can be viewed as *con-*

fusion profile parametrized by X_i and Y_j. This confusion profile is essentially a set of $|\mathcal{M}|$ distributions on the $|\mathcal{O}|$ simplex. In the simple scenario where the variables M_j and O_{ij} are binary, it reduces to the distribution of Type I and Type II errors as done in [7], but can capture more intricate dependencies among categorical variables in general. The three models are described below.

CTM. This model attempts to capture hidden source behavior such as expertise and common mistake patterns in terms of a source-specific confusion profile θ_i, which can be viewed as a latent source-specific feature X_i^{lat}. The priors on the correct answer M_j and each of components in $X_i^{lat} = \theta_i$ are assumed to be Dirichlet-Multinomial and Dirichlet respectively. The generative process is as follows:

$$\begin{aligned} \phi \sim Dir(\beta); \theta_{im} \sim Dir(\alpha_m),\ [i]_1^{N_u}, [m]_1^{|\mathcal{M}|}, \\ M_j \sim \phi; O_{ij} \sim Mult(\theta_{iM_j}),\ [i]_1^{N_u},\ [j]_1^{N_s}. \end{aligned} \quad (1)$$

CTM with Observed Subject Features (CTM-OSF). This model attempts to capture the variations in source behavior across observed categories of subjects. In this scenario, each subject is associated with a categorical observed attribute $Y_j^{ob} \in \{1, \cdots, N_{sf}]$ and and each source is associated with $|N_{sf}|$ confusion profiles corresponding to each of the subject categories. The generative process is given by:

$$\begin{aligned} \phi_a \sim Dir(\beta),\ [a]_1^{N_{sf}}, \\ \theta_{iam} \sim Dir(\alpha_m), [i]_1^{N_u},\ [m]_1^{|\mathcal{M}|},\ [a]_1^{N_{sf}}, \\ M_j \sim \phi_{Y_j}; O_{ij} \sim Mult(\theta_{iY_jM_j}),\ [i]_1^{N_u},\ [j]_1^{N_s}. \end{aligned} \quad (2)$$

CTM with Latent Source Groups (CTM-LSG). Rather than model the confusion profiles at an individual source level, this model assumes that each source U_i belongs to a hidden group $G_i \in \{1, \cdots, N_{sg}\}$, and associates each group with a confusion profile. The generative process includes assignment of group indices to the sources.

$$\begin{aligned} \phi \sim Dir(\beta), \theta_{km} \sim Dir(\alpha_m), [k]_1^{N_{sg}}, [m]_1^{|\mathcal{M}|}, \\ \psi \sim Dir(\gamma), G_i \sim \psi,\ [i]_1^{N_s}, \\ M_j \sim \phi_{Y_j}; O_{ij} \sim Mult(\theta_{G_i,M_j})\ [i]_1^{N_u},\ [j]_1^{N_s}. \end{aligned} \quad (3)$$

2.2.1 Inference

The generative process for CTM, as described above results in the following joint distribution:

$$\begin{aligned} p(O, M, \theta, \phi) &\propto \prod_{k=1}^{|\mathcal{M}|} \phi_k^{\beta_k - 1} \prod_{i=1}^{N_u} \prod_{k=1}^{|\mathcal{M}|} \prod_{l=1}^{|\mathcal{M}|} \theta_{ikl}^{\alpha_{kl}-1} \prod_{j=1}^{N_s} \prod_{k=1}^{K} \phi_k^{\delta(M_j,k)}, \\ &\times \prod_{i=1}^{N_u} \prod_{j=1}^{N_s} \prod_{k=1}^{|\mathcal{M}|} \prod_{l=1}^{|\mathcal{M}|} \theta_{ikl}^{\delta(M_j,k)\delta(O_{ij},l)}, \\ &\propto \prod_{k=1}^{|\mathcal{M}|} \phi_k^{n_k+\beta_k-1} \prod_{i=1}^{N_u} \prod_{k=1}^{|\mathcal{M}|} \prod_{l=1}^{|\mathcal{M}|} \theta_{ikl}^{m_{ikl}+\alpha_{kl}-1}. \end{aligned} \quad (4)$$

Here m_{ikl} is the number of times source i has provided opinion l to a subject whose correct answer is k, and n_k is the number of subjects which have k as the correct answer. On integrating out θ and ϕ, the Gibbs sampling equation is:

$$p(M_j = k | M_{-j}, O) \propto (n_k^{-j} + \beta_k) \prod_{i=1}^{N_u} \prod_{l=1}^{|\mathcal{M}|} (m_{ikl}^{-j} + \alpha_{kl}), \quad (5)$$

where n_k^{-j} and m_{ikl}^{-j} are n_k and m_{ikl} respectively, without considering subject j. In practice, while sampling M_j, we restrict ourselves to only those values in $\mathcal{M}$ which have been used in at least one of the opinions available on subject j. The inference steps for CTM-OSF and CTM-LSG are similarly derived, but are skipped here for brevity.

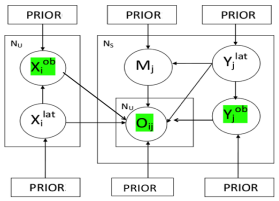

Figure 1: Graphical model for generic Bayesian opinion generation. Observed variables are marked in green.

We choose the hyperparameters α and β based on the data. α_{kl} is set proportional to the number of times l is provided as opinion on a subject where the most frequent opinion is k. We set β_k to be proportional to the number of times k is provided as opinion. In the presence of supervision, β_k is boosted proportional to the number of times it occurs as the correct answer for the supervised subjects.

3. EMPIRICAL EVALUATION

In this section, we present empirical results comparing the performance of the proposed models (CTM, CTM-OSF and CTM-LSG) relative to the state-of-the-art methods on real-world datasets in supervised and unsupervised settings.

3.1 Experimental Set-up

Datasets: We consider two datasets comprising of categorical-valued opinions: (a) *Quizmaster Dataset* [1], which contains questions on 11 different topics (physics, chemistry, history, literature, etc.) and the *Hubdub Dataset* [2] which contains questions pertaining to the outcome (winner, victory margin) of upcoming sports matches. For both the datasets, each question has a single correct answer, and the users (sources) attempt a variable number of questions. In the Quizmaster dataset, opinions also have typos and linguistic issues, which was addressed via string-matching and normalization as a preprocessing step though in principle, it could be incorporated into the confusion profile. Table 2 provides the details of the datasets. To evaluate the techniques in the presence of supervision, we also created a subset of the QuizMaster dataset (Quizmaster2), where we randomly select 80% of the subjects and their associated opinions. The remaining 20% is kept aside for supervision.

Algorithms: We consider the following baselines: Voting, TruthFinder (TF) [5], 3-Estimates [2], and LTM [7]. For LTM, TF and 3-Estimates, all opinions are transformed into binary-valued facts (question-opinion pairs), and each fact is assigned a score. The opinion corresponding to the fact with the best score is selected as the predicted correct answer for a subject. We consider two versions of LTM: (a) LTM-1 which corresponds to the original LTM and may infer more than one opinion as the correct answer for a question since each question-opinion pair is considered independently and inferred as true/false, and (b) LTM-2, which explicitly chooses a single fact per subject.

In the presence of supervision, we also consider an additional baseline algorithm *Discriminative* based on discriminative modeling. As with TF [5] and 3-estimate, we construct question-opinion pairs, which can be associated with a binary label of TRUE ("opinion is correct answer for ques-

Dataset	#Subjects	#Sources	#Opinions	#Distinct opinions per subject
Quizmaster	6076	458	33243	min 1, max 22
Quizmaster2	4876	447	26841	min 1, max 22
Hubdub	357	447	3051	min 1, max 6

Table 2: Details of experimental datasets.

Method	QuizMaster	Hubdub
Maximum	(5681)	(357)
Voting	5317	236
TF	5348	239
3-Est	5340	215
LTM1	3846	171
LTM2	5242	158
CTM	5513	239
CTM-OSF	**5523**	-
CTM-LSG	5508	**240**

Table 3: Number of correct answers found by different models on Categorical Data.

tion") or FALSE. We also construct features based on the opinion distribution and learn a generalized linear model.

We compare the newly proposed methods: CTM, CTM-OSF and CTM-LSG, against the above methods. In case of the quizmaster dataset, the topics for each question (11 topics such as physics, chemistry, history) can be used as subject features (Y_j^{ob}) in CTM-OSF. Since the Hubdub dataset does not have such features, we do not evaluate CTM-OSF on this data.

Metrics: All the algorithms except LTM1 output one answer for each subject. As performance metrics, we evaluate the number of predicted answers that match the Ground Truth. We note that, in the Quizmaster dataset, none of the opinions are correct in 395 out of the 6076 questions, and so in these 395 questions the correct answer may never be found, and hence, the maximum number of correct predictions achievable on this dataset is 5681.

3.2 Results and Discussions

Unsupervised Setting: Table 3 presents the prediction results of various algorithms on the two datasets (Quizmaster and Hubdub) in the absence of supervision. The values for CTM-LSG correspond to $N_{sg} = 5$, but variations of N_{sg} did not significantly affect the prediction. In case of QuizMaster, the proposed methods are clearly superior to all the baselines while in case of Hubdub, these methods are superior to the Bayesian models, but comparable to Voting and TruthFinder. A possible reason for this is that source-specific confusion profiles can effectively capture the latent interactions in QuizMaster dataset. In case of Hubdub dataset, the representation of the opinions and correct answers (e.g., win by 5 points) may not encode the relevant semantics (Soccer Team A wins over Soccer Team B by 5 points or Hockey Team C wins over Hockey Team D by 5 points) which are not the same from a source perspective. This problem would have been alleviated in CTM-OSF in the presence of observed subject-specific features, which were not available in readily usable form.

Since most of the baselines are primarily meant for binary-valued opinions, we also transformed the categorical opinions to binary facts (subject-opinion pairs) and measured the prediction quality in each case, in terms of Precision and Recall. In case of TruthFinder, we obtained precision-recall values of (0.87, 0.58) while for 3-estimate, we obtained (0.85,0.94), with thresholds chosen so as to maximize the F-measure. For LTM-2 the values were (0.86,0.86). For CTM, CTM-OSF and CTM-LSG, these values are (0.91, 0.91). So it appears that most of the gain is coming from better utilization of the mutual exclusivity between categorical values.

Method	0%	25%	50%	100%
Voting	4280	4280	4280	4280
TF	3789	4268 ± 1.92	4287 ± 8.44	4319
Discrim	-	4226	4249	4249
3-Est	4253	4248 ± 16.63	4237 ± 31.4	4275
CTM	4429 ± 3.81	4434 ± 3.67	4434 ± 5.08	4437 ± 6.13
CTM-OSF	4433 ± 4.25	4438 ± 4.13	4443 ± 4.47	4449 ± 4.45
CTM-LSG	4427 ± 3.78	4430 ± 3.5	4429 ± 4.29	4429 ± 8.12

Table 4: Effects of Supervision on prediction accuracy on Quizmaster2 datset. Values and standard deviations computed over 10 runs each.

Effects of Supervision: Next, we study the effect of providing limited supervision in the form of correct answers to a few subjects being known. We choose these subjects randomly from 20% of Quizmaster dataset kept aside for supervision, and perform the predictions on the test partition (Quizmaster2). We consider 4 levels of supervision- 0%, 25%, 50% and 100% of the training subset. Table 4 shows the results pointing to the superior performance of the proposed models. However, the performance of the proposed methods is relatively invariant to the amount of supervision provided, unlike TF which clearly benifits from supervision.

4. CONCLUSION AND FUTURE WORK

We proposed a generic opinion reconciliation approach via Bayesian modeling that includes certain existing Bayesian models (e.g. Latent Truth Model [7]) as special cases. We presented three models (CTM, CTM-OSF, CTM-LSG) for categorical-valued opinions that elegantly capture the hidden source behavior, variations across subject groups, and inter-source correlations via appropriately chosen latent variables. Empirical results are encouraging and indicate that the proposed models are superior to existing state-of-the-art techniques based on trust propagation, discriminative learning, as well as Bayesian approaches designed for binary opinions. In future, we plan to explore CTM-variants that incorporate textual variations in the generative processes, as well as specialized models for subjects that involve a comparison among a pair of entities, e.g., match outcomes in Hubdub. We also plan to explore more efficient utility-based inference mechanisms that can scale to large web-scale datasets.

5. REFERENCES

[1] J. Boyd-Graber, B. Satinoff, H. He, and H. Daumé III. Besting the quiz master: crowdsourcing incremental classification games. In *EMNLP'12*, pages 1290–1301.

[2] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *3rd ACM ICWSDM'2010*, pages 131–140. ACM, 2010.

[3] G. Qi, C. Aggarwal, P. Moulin, and T. Huang. Learning from collective intelligence in groups. *arXiv preprint 1210.0954*, 2012.

[4] V. C. Rayakar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *JMLR*, 13:491–518, 2012.

[5] X. Yin, J. Han, and P. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE TKDE*, 20(6):796–808, 2008.

[6] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *QDB'12*, 2012.

[7] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *VLDB Endowment'2012*, 5(6):550–561.