Modeling Noisy Annotated Data with Application to Social Annotation

Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda

Abstract—We propose a probabilistic topic model for analyzing and extracting content-related annotations from noisy annotated discrete data such as web pages stored using social bookmarking services. With these services, since users can attach annotations freely, some annotations do not describe the semantics of the content, thus they are noisy, i.e. not content-related. The extraction of content-related annotations can be used as a prepossessing step in machine learning tasks such as text classification and image recognition, or can improve information retrieval performance. The proposed model is a generative model for content and annotations, in which the annotations are assumed to originate either from topics that generated the content or from a general distribution unrelated to the content. We demonstrate the effectiveness of the proposed method by using synthetic data and real social annotation data for text and images.

Index Terms—Topic Models, Gibbs sampling, Text Modeling, Social Annotation, Noisy Data

1 Introduction

Recently there has been great interest in social annotation, also called collaborative tagging or folksonomy, created by users freely annotating objects such as web pages [1], photographs [2], blog posts [3], videos [4], music [5], and scientific papers [6]. Delicious, which is a social bookmarking service, and Flickr, which is an online photo sharing service, are two representative social annotation services, and they have succeeded in collecting huge numbers of annotations. Since users can attach annotations freely in social annotation services, the annotations include those that do not describe the semantics of the content, and are, therefore, not content-related [7]. For example, annotations such as 'nikon' or 'canon' in a social photo service often represent the name of the manufacturer of the camera with which the photographs were taken, or annotations such as '2008' or 'november' indicate when they were taken. Other examples of content-unrelated annotations include those designed to remind the annotator such as 'toread', those identifying qualities such as 'great', and those identifying ownership.

Content-unrelated annotations can often constitute noise if used for training samples in machine learning tasks, such as automatic text classification and image recognition. Although the performance of a classifier can generally be improved by increasing the number of training samples, noisy samples have a detrimental effect on the classifier. We can improve classifier performance if we can employ huge amounts of social annotation data from which content-unrelated annotations have been filtered out. Content-unrelated annotations may also constitute noise in information retrieval. For example, a user may wish to retrieve a photograph of a Nikon camera rather than a photograph taken by a Nikon camera.

In this paper, we propose a probabilistic topic model for analyzing and extracting content-related annotations from noisy annotated data, which we call the *noisy annotation topic model* (NATM). A number of methods for automatic annotation have been proposed [8], [9], [10], [11], [12]. However, they implicitly assume that all annotations are related to content, and to the best of our knowledge, no attempt has been made to extract content-related annotations automatically. The extraction of content-related annotations can improve the performance of machine learning and information retrieval tasks. The NATM can also be used for the automatic generation of content-related annotations.

The NATM is a generative model for content and annotation. It first generates content, and then generates the annotations. We assume that each annotation is associated with a latent variable that indicates whether or not it is related to the content, and the annotation originates either from the topics that generated the content or from a content-unrelated general distribution depending on the latent variable. The inference can be achieved based on collapsed Gibbs sampling.

Intuitively speaking, our approach considers an annotation to be content-related when it is almost always attached to objects in a specific topic. On the other hand, an annotation that is attached to objects in various topics is considered to be content-unrelated. Even if annotations are nominally the same, some may be related to the content, and others may not. For example, the annotation 'nikon' attached to a photograph about a camera made by Nikon is related to the content, on the other hand one attached to a photograph taken with a Nikon camera is not. To deal with this situation, the NATM models the relevance of each annotation to the content by considering both the annotation text and the

• *T. Iwata, T. Yamada, and N. Ueda are with NTT Communication Science Laboratories, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan.*

relationship between the content and annotation.

The NATM is based on topic models. A topic model is a hierarchical probabilistic model, in which a document is modeled as a mixture of topics, and where a topic is modeled as a probability distribution over words. Topic models are successfully used for a wide variety of applications including information retrieval [13], [14], collaborative filtering [15], [16], and visualization [17] as well as for modeling annotated data [9]. The NATM is an extension of correspondence latent Dirichlet allocation (Corr-LDA) [9], which is a generative topic model for content and annotation. Since the Corr-LDA assumes that all annotations are related to the content, it cannot be used for separating content-related annotations from content-unrelated annotations.

The extraction of content-related annotations can be considered a binary classification problem. However, as regards real social annotation data, the annotations are not explicitly labeled as content related/unrelated. Therefore, we cannot use supervised binary classifiers such as the support vector machine (SVM) [18]. The NATM is an unsupervised model, which can extract content-related annotations without content relevance labels.

In the rest of this paper, we assume that the given data are annotated document data, in which the content of each document is represented by words appearing in the document, and each document has both content-related and content-unrelated annotations. The NATM is applicable to a wide range of discrete data with noisy annotations. These include annotated image data, where each image is represented with visual words [19], and annotated movie data, where each movie is represented by user ratings.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work. In Section 3, we formulate the proposed noisy annotation topic model, and describe its inference procedures. We also present procedures in a partially explicit relevance setting and procedures for using the proposed model for the preprocessing when training different types of classifiers. In Sections 4 and 5, we demonstrate the effectiveness of the proposed model by using synthetic data and real social annotation data, respectively. Finally, we provide concluding remarks and a discussion of future work in Section 6.

2 Related work

Recently a number of models have been proposed for automatic annotation especially for images [8], [9], [10], [11], [12]. However, since they do not model the relevance between content and annotation, they cannot be used for extracting content-related annotations. These automatic annotation methods are based on supervised classifiers, in which all annotations of training samples are considered even if they are unrelated to the content. This means they cannot be employed for the automatic generation of content-related annotations.

Topic models for social annotation have been proposed [20], [21], which model the relationships between objects, annotations, and users. In contrast to these models, the NATM does not require user information. There are some social annotation data in which user information is unavailable. Therefore, the proposed model is applicable to wider range of data sets than the methods that require user information. In [20], entropy is used as an indicator of the ambiguity of the annotation, where the entropy represents how uniformly the annotation is attached over topics. Since ambiguous annotations imply that they are attached to documents covering a wide range of topics independent of content, this entropy-based method may also be used for extracting content-related annotations. However, the NATM has three advantages over the entropy-based methods. First, even if annotations have the same name, the NATM is able to consider some of them to be related to the content and others not by taking account of the relevance of each annotation to the content. Second, the entropy-based method requires some ad-hoc entropy thresholds for classifying whether or not each annotation is content-related. On the other hand, the NATM does not require any thresholds because the classification is explicitly achieved by inferring a latent variable that represents the relevance of each annotation to the content. Third, the NATM simultaneously models content and annotations with their relevance in one probabilistic framework. On the other hand, since the entropy-based method finds content-unrelated annotations via post-processing after the inference, errors accumulated in the inference cannot be corrected in the extraction process.

Topic models with a background distribution [22], [23] assume that words are generated either from a topic-specific distribution or from a corpus-wide background distribution. Although they are generative models for documents without annotations, the NATM is related to it in the sense that data may generated from a topic-unrelated distribution depending on a latent variable. There are other topic models that generate a word depending on a latent binary variable. For example, citation influence models [24] are also related to the proposed model because they assume that a word is generated according to topic proportions of the document or those of a cited document depending on the latent binary variable. In short, the citation influence models are topic models for generating words in a content using the citation information, whereas the proposed model is a topic model for generating both words in a content and annotations from the same topics. Therefore, the aim is different from ours. [25] recently introduced a probabilistic model designed to understand scene images, objects and associated noisy annotations. Their model uses a switch variable that decides whether a annotation is visually relevant or not. The switch variable depends on an object in the content, and visually irrelevant annotations are generated from the scene dependent distribution. Therefore, the visually irrelevant annotations do not

TABLE 1
Notation

Symbol	Description
D	number of documents
W	number of unique words
T	number of unique annotations
K	number of topics
N_d	number of words in the dth document
M_d	number of annotations in the dth document
w_{dn}	nth word in the dth document, $w_{dn} \in \{1, \cdots, W\}$
z_{dn}	topic of the nth word in the dth document, $z_{dn} \in \{1, \cdots, K\}$
t_{dm}	mth annotation in the dth document, $t_{dm} \in \{1, \cdots, T\}$
c_{dm}	topic of the mth annotation in the dth document, $c_{dm} \in \{1, \cdots, K\}$
r_{dm}	relevance to the content of the mth annotation of the dth document, $r_{dm} = 1$ if relevant, $r_{dm} = 0$ otherwise

have obvious visual correspondences, but they relate to the content. Since they focus on image analysis, their model is not appropriate for text such as social bookmark data.

The NATM is related to the partial label learning problem [26], in which each training sample is labeled with a set of possible labels, one of which is correct, when we consider a content-related annotation to be the correct label. However, the partial label learning assumes that there is only one correct label per sample. Since there may be multiple content-related annotations, methods for the partial label learning cannot be used for our purpose.

3 Proposed Method

3.1 Noisy Annotation Topic Model

Suppose that, we have a set of D documents, and each document consists of a pair of words and annotations $(\boldsymbol{w}_d, \boldsymbol{t}_d)$, where $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ is the set of words in the dth document that represents the content, and $\boldsymbol{t}_d = \{t_{dm}\}_{m=1}^{M_d}$ is the set of assigned annotations, or tags. The vocabulary of words and that of annotations can be different. For example, words and annotations can be discrete visual features and tags, respectively, and they can be written in different languages. Our notation is summarized in Table 1.

The proposed noisy annotation topic model (NATM) first generates the content, and then generates the annotations. The generative process for the content is the same as that of basic topic models, such as latent Dirichlet allocation (LDA) [13]. Each document has topic proportions $\boldsymbol{\theta}_d$ that are sampled from a Dirichlet distribution. For each of the N_d words in the document, a topic z_{dn} is chosen from the topic proportions, and then word w_{dn} is generated from a topic-specific multinomial distribution $\boldsymbol{\phi}_{z_{dn}}$. In the generative process for annotations, each annotation is assessed as to whether or not it is related to the content. In particular, each annotation is associated with a latent variable r_{dm} with a value $r_{dm} = 0$ if annotation t_{dm} is unrelated to the content; $r_{dm} = 1$ otherwise. If the annotation is not related to the content, $r_{dm} = 0$, annotation t_{dm} is sampled from a general topic-unrelated multinomial distribution $\boldsymbol{\psi}_0$. If the annotation is related to the content, $r_{dm} = 1$, annotation t_{dm} is sampled from a topic-specific multinomial distribution $\boldsymbol{\psi}_{c_{dm}}$, where c_{dm} is the topic for the annotation. Topic c_{dm} is sampled given topics $\boldsymbol{z}_d = \{z_{dn}\}_{n=1}^{N_d}$ that have previously generated content. This means that topic c_{dm} is generated from a multinomial distribution, in which $P(c_{dm} = k) = \frac{N_{kd}}{N_d}$, where N_{kd} is the number of words assigned to topic k in the dth document.

In summary, the NATM assumes the following generative process for a set of annotated documents $\{(\boldsymbol{w}_d, \boldsymbol{t}_d)\}_{d=1}^{D}$,

1) Draw relevance probability
   $\lambda \sim \text{Beta}(\eta)$
2) Draw content-unrelated annotation probability
   $\boldsymbol{\psi}_0 \sim \text{Dirichlet}(\gamma)$
3) For each topic $k = 1, \cdots, K$:
   a) Draw word probability
      $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\beta)$
   b) Draw annotation probability
      $\boldsymbol{\psi}_k \sim \text{Dirichlet}(\gamma)$
4) For each document $d = 1, \cdots, D$:
   a) Draw topic proportions
      $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha)$
   b) For each word $n = 1, \cdots, N_d$:
      i) Draw topic for word
         $z_{dn} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$
      ii) Draw word
         $w_{dn} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{dn}})$
   c) For each annotation $m = 1, \cdots, M_d$:
      i) Draw topic for annotation
         $c_{dm} \sim \text{Multinomial}(\{\frac{N_{kd}}{N_d}\}_{k=1}^{K})$
      ii) Draw relevance
         $r_{dm} \sim \text{Bernoulli}(\lambda)$
      iii) Draw annotation
         $$t_{dm} \sim \begin{cases} \text{Multinomial}(\boldsymbol{\psi}_0) & \text{if } r_{dm} = 0 \\ \text{Multinomial}(\boldsymbol{\psi}_{c_{dm}}) & \text{otherwise} \end{cases}$$

where α, β and γ are Dirichlet distribution parameters, and η is a beta distribution parameter. Figure 1 shows a graphical model representation of the NATM, where shaded and unshaded nodes indicate observed and latent variables, respectively.

Each latent relevance variable r_{dm} is drawn from a Bernoulli distribution with parameter λ, where λ represents the probability that an annotation is related to the content. We assume that λ is generated according to a beta distribution because it is conjugate to a Bernoulli distribution, and the inference can be efficiently performed based on collapsed Gibbs sampling by integrating out the parameter λ. We use conjugate Dirichlet priors for the multinomial parameters in the proposed model as used in the LDA.

Topics for annotations are drawn proportional to the number of topics assigned in the content. Therefore, annotations tend to be assigned the same topics with the

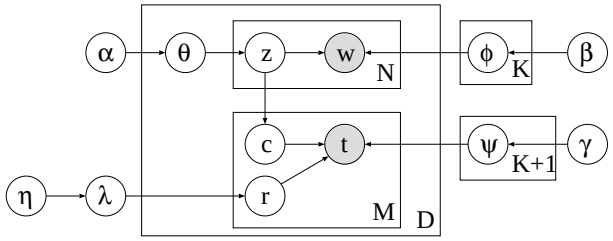


Fig. 1. Graphical model representation of the proposed noisy annotation topic model (NATM) with content relevance.

content, and a topic for content and the corresponding topic for annotation have similar meaning.

Intuitively speaking, the proposed model assumes that content-related annotations have the same topics with the ones assigned to the words in the content. Therefore, a content-related annotation is assigned a topic c that is generated from topics in the content. The proposed model assumes that content-unrelated annotations are independent from the topics in the content, and they are generated by a general topic-unrelated multinomial distribution.

As with the Corr-LDA, the NATM first generates the content and then generates the annotations by modeling the conditional distribution of latent topics for annotation given the topics in the content. Therefore, it achieves a comprehensive fit of the joint distribution of content and annotations and finds superior conditional distributions of annotations given content [9].

The joint distribution on words, annotations, topics for words, topics for annotations and relevance is described as follows:

$$\begin{aligned} &P(\boldsymbol{W},\boldsymbol{T},\boldsymbol{Z},\boldsymbol{C},\boldsymbol{R}|\alpha,\beta,\gamma,\eta) \\ &= P(\boldsymbol{Z}|\alpha)P(\boldsymbol{W}|\boldsymbol{Z},\beta)P(\boldsymbol{T}|\boldsymbol{C},\boldsymbol{R},\gamma)P(\boldsymbol{R}|\eta)P(\boldsymbol{C}|\boldsymbol{Z}), \end{aligned} \tag{1}$$

where $\boldsymbol{W} = \{\boldsymbol{w}_d\}_{d=1}^D$, $\boldsymbol{T} = \{\boldsymbol{t}_d\}_{d=1}^D$, $\boldsymbol{Z} = \{\boldsymbol{z}_d\}_{d=1}^D$, $\boldsymbol{C} = \{\boldsymbol{c}_d\}_{d=1}^D$, $\boldsymbol{c}_d = \{c_{dm}\}_{m=1}^{M_d}$, $\boldsymbol{R} = \{\boldsymbol{r}_d\}_{d=1}^D$, and $\boldsymbol{r}_d = \{r_{dm}\}_{m=1}^{M_d}$. We can integrate out multinomial distribution parameters, $\{\boldsymbol{\theta}_d\}_{d=1}^D$, $\{\boldsymbol{\phi}_k\}_{k=1}^K$ and $\{\boldsymbol{\psi}_{k'}\}_{k'=0}^K$, because we use Dirichlet distributions for their priors, which are conjugate to multinomial distributions. The first term on the right hand side of (1) is calculated by $P(\boldsymbol{Z}|\alpha) = \prod_{d=1}^D \int P(\boldsymbol{z}_d|\boldsymbol{\theta}_d)P(\boldsymbol{\theta}_d|\alpha)d\boldsymbol{\theta}_d$, and we have the following equation by integrating out $\{\boldsymbol{\theta}_d\}_{d=1}^D$,

$$P(\boldsymbol{Z}|\alpha) = \left(\frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K}\right)^D \prod_d \frac{\prod_k \Gamma(N_{kd}+\alpha)}{\Gamma(N_d+\alpha K)}, \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function. Similarly, the second term is given as follows,

$$P(\boldsymbol{W}|\boldsymbol{Z},\beta) = \left(\frac{\Gamma(\beta W)}{\Gamma(\beta)^W}\right)^K \prod_k \frac{\prod_w \Gamma(N_{kw}+\beta)}{\Gamma(N_k+\beta W)}, \tag{3}$$

where N_{kw} is the number of times word w has been assigned to topic k, and $N_k = \sum_w N_{kw}$. The third term is given as follows,

$$P(\boldsymbol{T}|\boldsymbol{C},\boldsymbol{R},\gamma) = \left(\frac{\Gamma(\gamma T)}{\Gamma(\gamma)^T}\right)^{K+1} \prod_{k'} \frac{\prod_t \Gamma(M_{k't}+\gamma)}{\Gamma(M_{k'}+\gamma T)}, \tag{4}$$

where $k' \in \{0,\cdots,K\}$, and $k'=0$ indicates irrelevance to the content. $M_{k't}$ is the number of times annotation t has been identified as content-unrelated if $k'=0$, or as content-related topic k' if $k' \neq 0$, and $M_{k'} = \sum_t M_{k't}$. The Bernoulli parameter λ can also be integrated out because we use a beta distribution for the prior, which is the conjugate prior of a Bernoulli distribution. The fourth term is given as follows,

$$P(\boldsymbol{R}|\eta) = \frac{\Gamma(2\eta)}{\Gamma(\eta)^2}\frac{\Gamma(M_0+\eta)\Gamma(M-M_0+\eta)}{\Gamma(M+2\eta)}, \tag{5}$$

where M is the number of annotations, and M_0 is the number of content-unrelated annotations. The fifth term is given as follows,

$$P(\boldsymbol{C}|\boldsymbol{Z}) = \prod_d \prod_k \left(\frac{N_{kd}}{N_d}\right)^{M'_{kd}}, \tag{6}$$

where M'_{kd} is the number of annotations that are assigned to topic k in the dth document.

3.2 Inference

The inference of the latent topics $\boldsymbol{Z}$ given content $\boldsymbol{W}$ and annotations $\boldsymbol{T}$ can be efficiently computed using collapsed Gibbs sampling [27]. Given the current state of all but one variable, z_j, where $j=(d,n)$, the assignment of a latent topic to the nth word in the dth document is sampled from,

$$\begin{aligned} &P(z_j = k|\boldsymbol{W},\boldsymbol{T},\boldsymbol{Z}_{\backslash j},\boldsymbol{C},\boldsymbol{R}) \\ &\propto \frac{N_{kd\backslash j}+\alpha}{N_{d\backslash j}+\alpha K}\frac{N_{kw_j\backslash j}+\beta}{N_{k\backslash j}+\beta W}\left(\frac{N_{kd\backslash j}+1}{N_{kd\backslash j}}\right)^{M'_{kd}}, \end{aligned} \tag{7}$$

where $\backslash j$ represents the count when excluding the nth word in the dth document. Given the current state of all but one variable, r_i, where $i=(d,m)$, the assignment of either relevance or irrelevance to the mth annotation in the dth document is estimated as follows,

$$\begin{aligned} P(r_i=0|\boldsymbol{W},\boldsymbol{T},\boldsymbol{Z},\boldsymbol{C},\boldsymbol{R}_{\backslash i}) &\propto \frac{M_{0\backslash i}+\eta}{M_{\backslash i}+2\eta}\frac{M_{0t_i\backslash i}+\gamma}{M_{0\backslash i}+\gamma T}, \\ P(r_i=1|\boldsymbol{W},\boldsymbol{T},\boldsymbol{Z},\boldsymbol{C},\boldsymbol{R}_{\backslash i}) &\propto \frac{M_{\backslash i}-M_{0\backslash i}+\eta}{M_{\backslash i}+2\eta}\frac{M_{c_it_i\backslash i}+\gamma}{M_{c_i\backslash i}+\gamma T}. \end{aligned} \tag{8}$$

The assignment of a topic to a content-unrelated annotation is estimated as follows,

$$P(c_i = k|r_i=0,\boldsymbol{W},\boldsymbol{T},\boldsymbol{Z},\boldsymbol{C}_{\backslash i},\boldsymbol{R}_{\backslash i}) \propto \frac{N_{kd}}{N_d}, \tag{9}$$

and the assignment of a topic to a content-related annotation is estimated as follows,

$$P(c_i = k|r_i=1,\boldsymbol{W},\boldsymbol{T},\boldsymbol{Z},\boldsymbol{C}_{\backslash i},\boldsymbol{R}_{\backslash i}) \propto \frac{M_{kt_i\backslash i}+\gamma}{M_{k\backslash i}+\gamma T}\frac{N_{kd}}{N_d}. \tag{10}$$

One iteration of the Gibbs sampler corresponds to sample topics for each word and for each annotation, and the relevance of each annotation in the given documents.

The parameters α, β, γ, and η are estimated by maximizing the joint distribution (1). The following updating rules for maximizing the joint distribution are derived by using the bounds as described in [28],

$$\alpha^{(\text{new})} \leftarrow \alpha \frac{\sum_d \sum_z \Psi(n_{zd}+\alpha) - DK\Psi(\alpha)}{K\left(\sum_d \Psi(N_d+\alpha K) - D\Psi(\alpha K)\right)}, \quad (11)$$

$$\beta^{(\text{new})} \leftarrow \beta \frac{\sum_z \sum_w \Psi(n_{zw}+\beta) - KW\Psi(\beta)}{W\left(\sum_z \Psi(n_z+\beta W) - K\Psi(\beta W)\right)}, \quad (12)$$

$$\gamma^{(\text{new})} \leftarrow \gamma \frac{\sum_t \sum_{z'} \Psi(m_{z't}+\gamma) - (K+1)T\Psi(\gamma)}{T\left(\sum_{z'} \Psi(m_{z'}+\gamma T) - (K+1)\Psi(\gamma T)\right)}, \quad (13)$$

$$\eta^{(\text{new})} \leftarrow \eta \frac{\Psi(m_0+\eta) + \Psi(m-m_0+\eta) - 2\Psi(\eta)}{2\left(\Psi(m+2\eta) - \Psi(2\eta)\right)}, \quad (14)$$

where $\Psi(\cdot)$ is a digamma function defined by $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$. These update rules can find a local optimum solution for the parameters.

By iterating Gibbs sampling with (7), (8), and (9) or (10), and maximum likelihood estimation with (11), (12), (13), and (14), we can infer latent topics for words and annotations as well as the relevance to their content while optimizing the parameters.

Let $\{r_{dm}^{(s)}\}_{s=1}^{S}$ be S sampled values of the relevance variable r_{dm} after the burn-in period. The relevance probability of the mth annotation in the dth document can be estimated by,

$$\hat{r}_{dm} = \frac{1}{S}\sum_{s=1}^{S} r_{dm}^{(s)}. \quad (15)$$

The inferred model can also predict content-related annotations given content without annotations. In particular, the probability of content-related annotation t in the dth document can be calculated as follows,

$$P(t|d,\mathcal{D}) = \sum_k \hat{\theta}_{dk}\hat{\psi}_{kt}, \quad (16)$$

where $\hat{\theta}_{dk} = \frac{N_{kd}}{N_d}$ is the point estimate of the topic proportions for annotations, and $\hat{\psi}_{kt} = \frac{M_{kt}+\gamma}{M_k+\gamma T}$ is the point estimate of the annotation multinomial distribution.

3.3 Partially explicit relevance setting

In the above discussion, we assumed that none of the annotations were labeled as content related/unrelated, or relevance information was implicit. However in practice, the content relevance labels for some annotations might be available. For example, some social annotations can be manually labeled by experts. Therefore, we consider a partially explicit relevance setting, in which we have annotations labeled with related or unrelated as well as unlabeled annotations. The NATM can deal directly with labeled annotations. In an implicit relevance setting, relevance variables r are assumed to be hidden for all annotations as in Figure 1. In a partially explicit setting, relevance variables r are assumed to be observed and fixed for explicit relevance annotations, and hidden for implicit relevance annotations. Thus, Gibbs sampling of r for each explicit relevance annotation is unnecessary. The other inference procedures in a partially explicit setting are the same as those in an implicit setting. In particular, the inference can be performed by iterating Gibbs sampling with (7) for each word, (8) for each unlabeled annotation, (9) or (10) for each explicit and implicit annotation, and maximum likelihood estimation with (11), (12), (13), and (14). By using the explicit relevance information, the NATM can be inferred more precisely.

3.4 Combination with different classifiers

The NATM can be used for the prepossessing of classifier training. By filtering out content-unrelated annotations, we can improve classifier performance. The NATM can predict annotations given content without annotations by (16). However, different classifiers might achieve better classification performance than the topic model based method in some applications. For example, discriminative classifiers such as maximum entropy models [29] and support vector machines [30] usually perform better when training data are abundant than generative classifiers such as naive Bayes models and topic models, and this has been experimentally confirmed in a text classification problem [31]. Thus, we present procedures that combine the advantage of the NATM in filtering out content-unrelated annotations and the advantage of different classifiers as regards high classification accuracy.

In general, a classifier is trained by minimizing the following empirical error function over the given samples,

$$E = \sum_{d=1}^{D}\sum_{m=1}^{M_d} J(\boldsymbol{x}_d, t_{dm}), \quad (17)$$

where $\boldsymbol{x}_d$ is the feature vector of the dth document's content, each sample is represented by a pair consisting of the feature vector and annotation $(\boldsymbol{x}, t)$, and the error function $J(\boldsymbol{x}, t)$ represents the error of the classifier given the sample. Typical error functions include negative log likelihood $J(\boldsymbol{x}, t) = -\log P(t|\boldsymbol{x})$, and 0-1 loss function, $J(\boldsymbol{x}, t) = 0$ if $f(\boldsymbol{x}) = t$ and $J(\boldsymbol{x}, t) = 1$ otherwise. If there are no content-unrelated annotations, the minimization of E will lead to the minimization of the expected error when we have sufficient numbers of samples. However, noisy samples, i.e., content-unrelated annotations, have a detrimental effect on training classifiers. Therefore, we propose the following weighted error function,

$$E_{\text{r}} = \sum_{d=1}^{D}\sum_{m=1}^{M_d} \hat{r}_{dm} J(\boldsymbol{x}_d, t_{dm}), \quad (18)$$

where each sample is weighted by the relevance to the content $\hat{r}_{dm}$ that is calculated by (15). When using

the weights, content-unrelated annotations are less effective for training than content-related ones. Therefore, a detrimental effect of content-unrelated annotations can be eliminated. The classifier that is used should be capable of dealing with weighted samples. This is not a severe limitation of our approach because many common classifiers are able to learn with a weighted error. For example, to learn a classifier based on an exponential family, such as one with multinomial or Gaussian class-conditional distributions, we have only to calculate weighted sufficient statistics. The SVM with weights for each sample, which is called the fuzzy SVM [32] or weighted SVM [33], has also been proposed. With feature selection methods, features that are useful for improving classifier performance are extracted. This approach selects samples that are useful for training classifiers instead of selecting useful features.

We describe the procedure for training maximum entropy models with the proposed weighted framework using the NATM as an example. The maximum entropy model, which is also called the logistic regression or log-linear model, is a discriminative model, and it estimates a probability distribution that maximizes entropy under the constraints in the given samples. This model has been used in various research fields such as text classification [29] and collaborative filtering [34]. The maximum-entropy distribution of annotation t given feature vector $\boldsymbol{x}$ is represented as follows:

$$P(t|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\mu}_t^\top \boldsymbol{x})}{\sum_{t'=1}^{T} \exp(\boldsymbol{\mu}_{t'}^\top \boldsymbol{x})}, \tag{19}$$

where $\boldsymbol{\mu}_t$ is an unknown parameter vector for annotation t, and where $\boldsymbol{\mu}_t^\top$ represents the transpose of $\boldsymbol{\mu}_t$. When we use a negative log likelihood for the error function and a Gaussian prior for $\boldsymbol{\mu}_t$ with mean $\mathbf{0}$ and covariance $\nu^{-1}\boldsymbol{I}$ [35], the weighted error function becomes:

$$\begin{aligned} E_{\mathrm{r}}^{\mathrm{ME}} &= -\sum_{d=1}^{D}\sum_{m=1}^{M_d} \hat{r}_{dm} \left(\boldsymbol{\mu}_{t_{dm}}^\top \boldsymbol{x}_d - \log \sum_{t'=1}^{T} \exp(\boldsymbol{\mu}_{t'}^\top \boldsymbol{x}_d) \right) \\ &\quad + \frac{\nu}{2} \sum_{t=1}^{T} \parallel \boldsymbol{\mu}_t \parallel^2 . \end{aligned} \tag{20}$$

We can estimate the unknown parameters $\{\boldsymbol{\mu}_t\}_{t \in \boldsymbol{T}}$ by minimization via the quasi-Newton method [36]. The global optimality of the estimate is guaranteed due to the concavity of the weighted error function.

4 Experiments with Synthetic Content-Unrelated Annotations

4.1 Data

We evaluated the NATM quantitatively by using labeled text data from the 20 Newsgroups corpus [37] [1] and adding synthetic content-unrelated annotations. The corpus contains about 20,000 articles categorized into 20 discussion groups. We considered these 20 categories as content-related annotations, and we also randomly attached dummy categories to training samples as content-unrelated annotations. We created four types of training data, 20News-DP, 20News-DU, 20News-SP, and 20News-SU. In 20News-DP and 20News-DU, the dummy content-unrelated annotations were different from the content-related annotations. On the other hand, in 20News-SP and 20News-SU, the dummy content-unrelated annotations were chosen from the content-related annotations. In some real social annotation data, nominally same annotations can be both content-related and unrelated. To evaluate the NATM in this situation, we constructed 20News-SP and 20News-SU. 20News-DP and 20News-SP were used for evaluating the NATM when analyzing data with different probabilities of adding content-unrelated annotations, and 20News-DU and 20News-SU were used with different numbers of unique content-unrelated annotations. Specifically, in the 20News-DP data, the number of unique content-unrelated annotations was set at 20, and the probability of adding content-unrelated annotations per document was set at $\{0.05, \cdots, 1.0\}$. In the 20News-DU data, the number of unique content-unrelated annotations was set at $\{1, \cdots, 20\}$, and the probability of adding content-unrelated annotations per document was set at 1.0. In the 20News-SP data, the number of unique content-unrelated annotations was set at 10 and they were selected from the content-related annotations of 20 categories, and the probability of adding content-unrelated annotations per document was set at $\{0.05, \cdots, 1.0\}$. In the 20News-SU data, the number of unique content-unrelated annotations was set at $\{1, \cdots, 20\}$, and the probability of adding content-unrelated annotations per document was set at 1.0. We omitted stop-words and words that occurred only once. The vocabulary size was 52,647. We sampled 100 documents from each of the 20 categories, for a total of 2,000 documents. We used 10 % of the samples as test data. The synthetic content-unrelated annotation data are summarized in Table 2.

4.2 Perplexity

We evaluated the predictive performance of each method using the following perplexity of held-out content-related annotations given the content,

$$\mathrm{Perplexity} = \exp\left(-\frac{\sum_d \sum_{m=1}^{M_d^{\mathrm{test}}} \log P(t_{dm}^{\mathrm{test}}|d, \mathcal{D})}{\sum_d M_d^{\mathrm{test}}} \right), \tag{21}$$

where M_d^{test} is the number of held-out annotations in the dth document, t_{dm}^{test} is the mth held-out annotation in the dth document, and $\mathcal{D}$ represents the training samples. A lower perplexity represents higher predictive performance. In the NATM, we calculated the probability of content-related annotation t in the dth document given the training samples using (16). Note that no content-unrelated annotations were attached to the test samples.

1. Available at the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/

TABLE 2
Summary of the synthetic content-unrelated annotation data.

	20News-DP	20News-DU	20News-SP	20News-SU
unrelated annotations are different from related annotations	yes	yes	no	no
number of unique content-unrelated annotations	20	1,$\cdots$,20	10	1,$\cdots$,20
probability of adding content-unrelated annotations	0.05, $\cdots$, 1	1	0.05, $\cdots$, 1	1

We compared the proposed NATM with the correspondence latent Dirichlet allocation (Corr-LDA). The Corr-LDA [9] is a topic model for words and annotations, where all of the annotations are considered to be relevant to the content. For the NATM and the Corr-LDA, we set the number of latent topics, K, at 20, and estimated latent topics and parameters by using collapsed Gibbs sampling and the fixed-point iteration method, respectively.

Figure 2 shows the average perplexities over 100 experiments on the four data sets. The perplexities of the NATM were lower than those of the Corr-LDA in all of the data sets. This result indicates that the NATM can robustly predict content-related annotations with noisy training data. The perplexity achieved by the Corr-LDA was high because it does not take account of the relevance to the content and it considers all attached annotations to be content-related even if they are not. As the probability of content-unrelated annotations increases, the performance of the Corr-LDA deteriorated as shown in Figure 2 (a) and (c). On the other hand, the performance of the NATM provided low perplexity even when the probability of the content-unrelated annotations was increased. The low perplexity of the NATM in Figure 2 (c) and (d) shows that the NATM can appropriately model noisy data with the same name of content-related and unrelated annotations. In the 20News-SU data, the number of occurrence of each unrelated annotation decreases as the number of unique unrelated annotations increases, because the probability of adding unrelated annotations is fixed. Therefore, the estimated probability of generating a certain unrelated annotation becomes low, and the perplexity by the Corr-LDA for the held-out related annotations decreases in the 20News-SU.

4.3 Extracting content-related annotations

We evaluated the performance in terms of extracting content-related annotations. We considered extraction as a binary classification problem, in which each annotation is classified as either content-related or content-unrelated. For the evaluation measurement, we used F-measure, which is the harmonic mean of precision and recall.

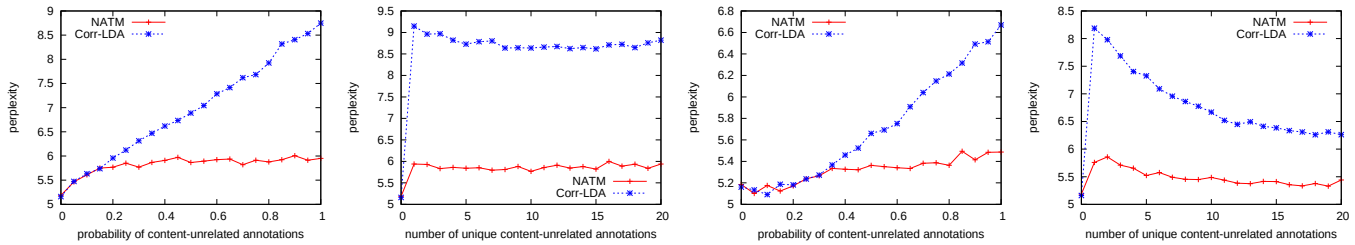
We compared the NATM with a entropy-based method and a baseline method. With the entropy-based method, we estimated the topic probability for each annotation $P(c|t)$ using the Corr-LDA, and we used the entropy for an indicator of the ambiguity of the annotation as is used in [20]. In particular, $P(c|t)$ is calculated by using the Bayes rule $P(c|t) \propto P(c)P(t|c)$, where $P(c)$ and $P(t|c)$ are obtained by the inference of the Corr-LDA. The entropy is given by $-\sum_{c=1}^{K} P(c|t) \log P(c|t)$. Then, we sorted annotations according to their entropies in descending order, and the annotations that are allocated before the maximum gap of the entropies are estimated as content-unrelated, and those after the maximum gap are estimated as content-related. With the baseline method, the annotations are considered to be content-related if any of the words in the annotations appear in the document. For example, when the category name is 'comp.graphics', if 'computer' or 'graphics' appears in the document, it is considered to be content-related. We assume that the baseline method knows that content-unrelated annotations do not appear in any document. Therefore, the precision of the baseline method is always one, because there are no false positive samples. Note that this baseline method does not support image data, because words in the annotations never appear in the content.

F-measures for the four 20News data sets are shown in Figure 3. A higher F-measure represents higher classification performance. The NATM achieved high F-measures with a wide range of ratios of content-unrelated annotations. The F-measures for 20News-DP and 20News-DU achieved by the NATM exceeded 0.89, and the F-measure without unrelated annotations was one. This result implies that the NATM can flexibly handle cases with different ratios of content-unrelated annotations. The F-measures for 20News-SP and 20News-SU are lower than those of 20News-DP and 20News-DU because annotations with the same name can be both content-related and unrelated in 20News-SP and 20News-SU. The F-measures achieved by the entropy-based method were lower than those by the NATM in the most cases except for 20News-DU. The F-measures achieved by the baseline method were low because annotations might be related to the content even if the annotations did not appear in the document. On the other hand, the NATM considers that annotations are related to the content when the topic, or latent semantics, of the content and the topic for the annotations are similar even if they did not appear in the document.

Figure 4 shows the ratio of the number of content-related annotations to that of all annotations. The ratio can be estimated by

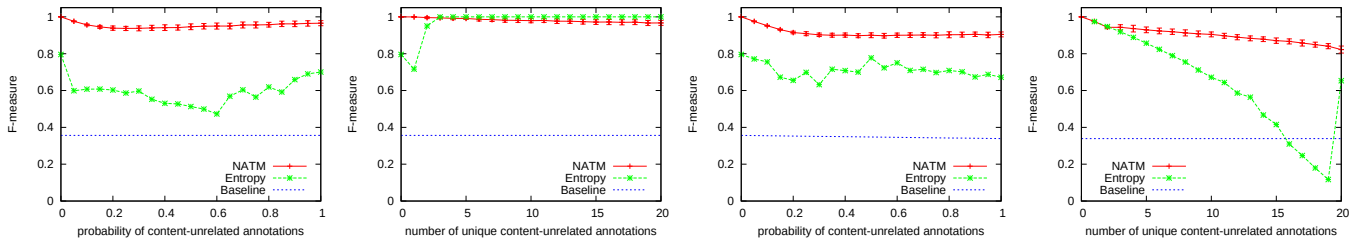
$$\hat{\lambda} = \frac{M - M_0 + \eta}{M + 2\eta}, \tag{22}$$

with the NATM. Here, M is the number of annotations, M_0 is the number of content-unrelated annotations, and



(a) 20News-DP (b) 20News-DU (c) 20News-SP (d) 20News-SU

Fig. 2. Perplexities of the held-out content-related annotations in 20News data.



(a) 20News-DP (b) 20News-DU (c) 20News-SP (d) 20News-SU

Fig. 3. F-measures of content relevance in 20News data.

η is the parameter of the beta distribution. The estimated ratios were close to the true ratios even though the NATM is unsupervised.

4.4 Scalability

The time complexity of one iteration of our Gibbs sampling is $O(DK(N+M))$, where D denotes the number of documents, K denotes the number of topics, and N and M are the average number of words and annotations in a document, respectively. We experimentally evaluated the scalability of the proposed model using a computer with a Corei7 965 3.2GHz CPU and a 12GB memory. Figure 5 (a) shows the average computational time with different numbers of documents while fixing the number of topics at 20, and Figure 5 (b) shows the average computational time with different numbers of topics while fixing the number of documents at 2,000. We performed 100 experiments using a 20News-DP data set, where we set the probability of content-unrelated annotations at 1, and the number of iterations at 1,000. The computational time is linear against the number of documents, and the number of topics. These results are consistent with the theoretical computational complexities.

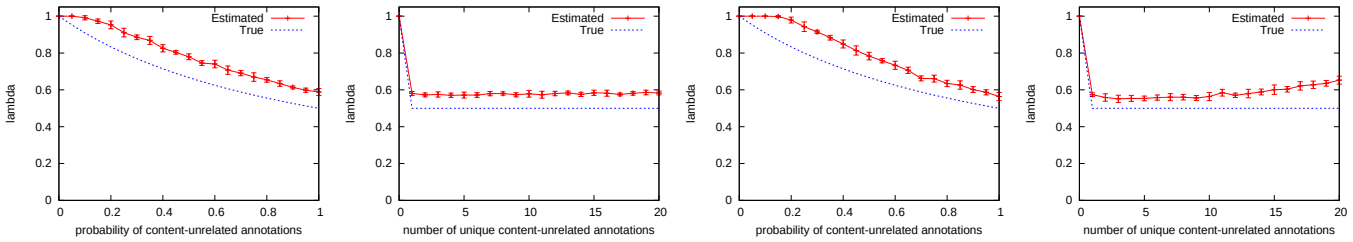
4.5 Analysis on parameter estimation

The proposed model has four parameters α, β, γ and η. We can find a local optimum solution by using (11), (12), (13), and (14). Here, we experimentally investigate sensitiveness of the parameter estimation for the initial condition. We performed 100 experiments using a 20News-DP data set, where we set the probability of content-unrelated annotations at 1. The initial values for the parameters are randomly chosen from a uniform distribution from 0 to 1. Table 3 shows the average estimated parameters and their standard deviations. The standard deviations were small for all the parameters. This result indicates that the proposed inference procedure is robust to the initialization even though they are local optima.

TABLE 3
Average estimated parameters and their standard deviations in 20News-DP data.

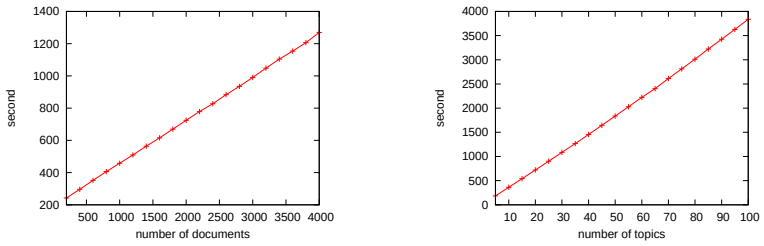
variable	average $\pm$ standard deviation
α	0.0909 $\pm$ 0.0003
β	0.0143 $\pm$ 0.0001
γ	0.0201 $\pm$ 0.0006
η	16.7975 $\pm$ 0.8553

We experimentally evaluated the importance of updating parameters. Figure 6 shows the perplexities of the held-out content-related annotations in 20News-DP data when we fixed a parameter. For example, in Figure 6 (a), α was fixed with the value at the x-axis, and the other parameters, β, γ and η, were estimated by maximizing the joint likelihood. The horizontal line shows the perplexity when the all parameters were estimated. The perplexities changed depending on the values of parameters α, β and γ, and the proposed model achieved low perplexities by estimating these parameters. This result indicates that estimating these parameters is important for the performance. The value of η did not influence on the performance. Note that the range of the y-axis in Figure 6 (d) is narrow compared with the other figures. This is because η is a hyper-parameter for a Bernoulli distribution, which has only one parameter. Figure 7 shows the F-measures in 20News-DP data when we fixed a parameter. The F-measures shows the same tendency of the perplexities in Figure 6. The parameter settings that



(a) 20News-DP (b) 20News-DU (c) 20News-SP (d) 20News-SU

Fig. 4. Estimated content-related annotation ratios in 20News data.



(a) different numbers of documents (b) different numbers of topics

Fig. 5. Computational time (second) of the proposed model with (a) different numbers of documents and (b) different numbers of topics.

achieved lower perplexities achieved relatively higher F-measures.

4.6 Partially explicit relevance setting

In Section 3.3, we described an extension of the NATM for a partially explicit relevance setting, where some of the annotations were explicitly labeled with related or unrelated. We evaluated the NATM in the partially explicit setting using 20News data sets. We used two sets of training data 20News-Dpart and 20News-Spart. In 20News-Dpart, for each document, we randomly attached a content-unrelated annotation that was different from the content-related annotations. The number of unique content-unrelated annotations was set at 20. In 20News-Spart, for each document, we randomly attached a content-unrelated annotation that was chosen from the content-related annotations. The number of unique content-unrelated annotations was set at 10. For both of the data sets, the probability of being labeled was set at $\{0, 0.05, \cdots, 1.0\}$. Figures 8, 9 and 10 show the perplexities, F-measures and the estimated content-related annotation ratios in the partially explicit setting, respectively. As the proportion of labeled annotations increases, the performance for all of the measurements was improved. This result indicates that the use of explicitly labeled annotations is important for the modeling if relevance information is available.

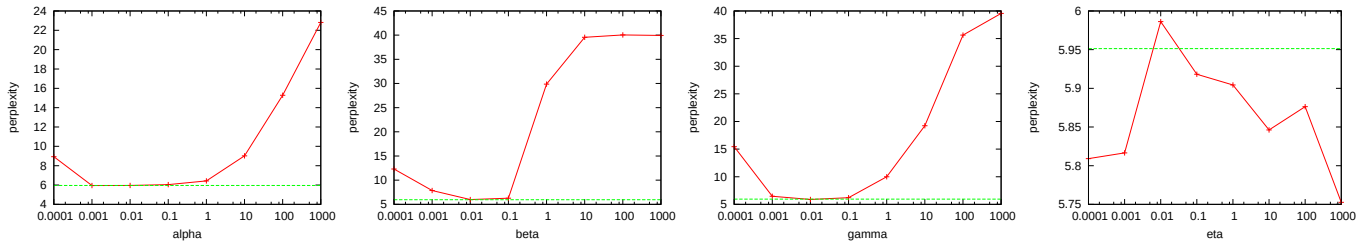
4.7 Combination with different classifiers

We evaluated the NATM when using it for the preposessing when training different types of classifiers as described in Section 3.4. The task is text classification using 20News data sets. For the classifier, we used a maximum entropy model (MaxEnt) that has confirmed its effectiveness in text classification [29]. We used the same four data sets as in Section 4, 20News-DP, 20News-DU, 20News-SP, and 20News-SU. The classifier performance was evaluated in terms of classification accuracy. Figure 11 shows the result. By combining the NATM and maximum entropy models, the accuracies were better in the presence of noise compared with those without using the NATM. The accuracy of the NATM when not combined with maximum entropy models was 0.50 for data without noise. In terms of classification accuracy, the NATM alone is worse than the maximum entropy model, which is a discriminative classifier whose effectiveness for text classification tasks has been confirmed in many applications. However, by using the NATM for preprocessing with high performance classifiers, content-unrelated annotations are filtered out, and the classification performance is improved. Since the proposed combination framework in (18) is general as described in Section 3.4, we can select high performance classifiers that depend on the application and the type of given data. The perplexities were high when there was one unique content-unrelated annotation in Figure 11 (b) and (d). Because, in this case, all the samples are labeled with a content-unrelated annotation, and the discriminative classifier is trained to attach the content-unrelated annotation.

5 Experiments with Real Social Annotations

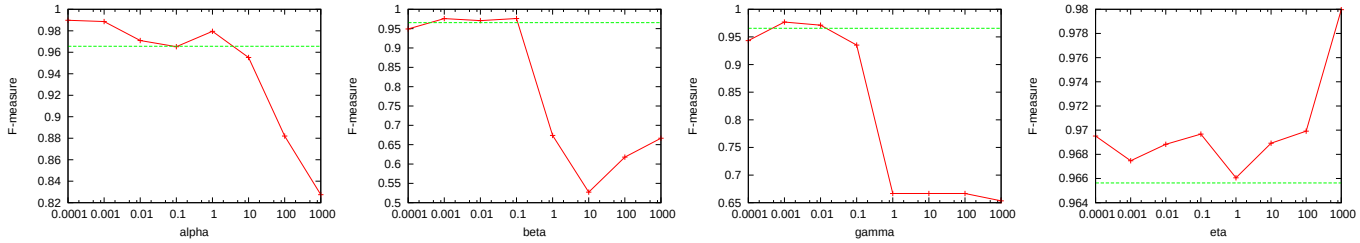
5.1 Data

We analyzed the following three sets of real social annotation data taken from two social bookmarking services



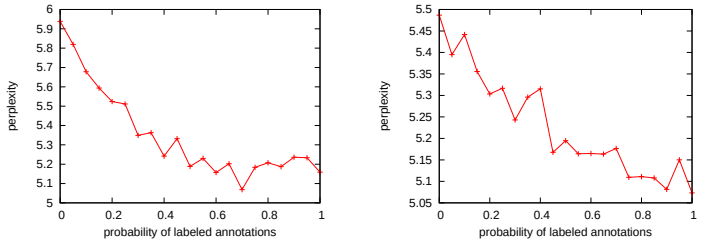
(a) α (b) β (c) γ (d) η

Fig. 6. Perplexities of the held-out content-related annotations in 20News-DP data with a fixed parameter. The horizontal line shows the perplexity when the all parameters were estimated.



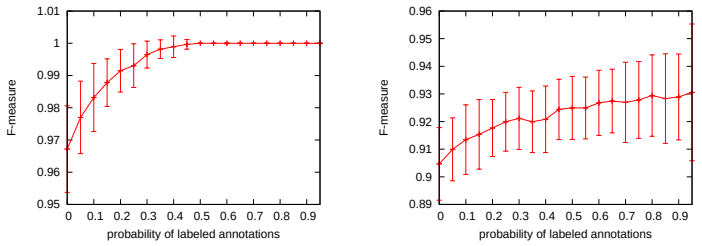
(a) α (b) β (c) γ (d) η

Fig. 7. F-measures in 20News-DP data with a fixed parameter. The horizontal line shows the F-measure when the all parameters were estimated.



(a) 20News-Dpart (b) 20News-Spart

Fig. 8. Perplexities of the held-out content-related annotations with different probabilities of labeled annotations in a partially explicit setting.

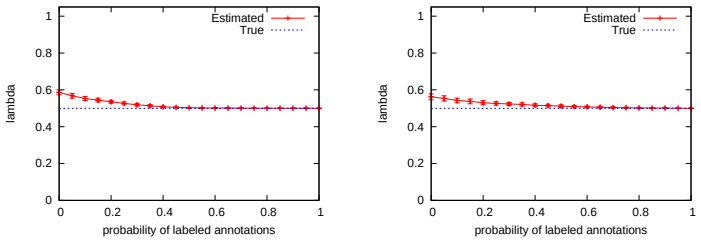


(a) 20News-Dpart (b) 20News-Spart

Fig. 9. F-measures of content relevance with different probabilities of labeled annotations in a partially explicit setting.

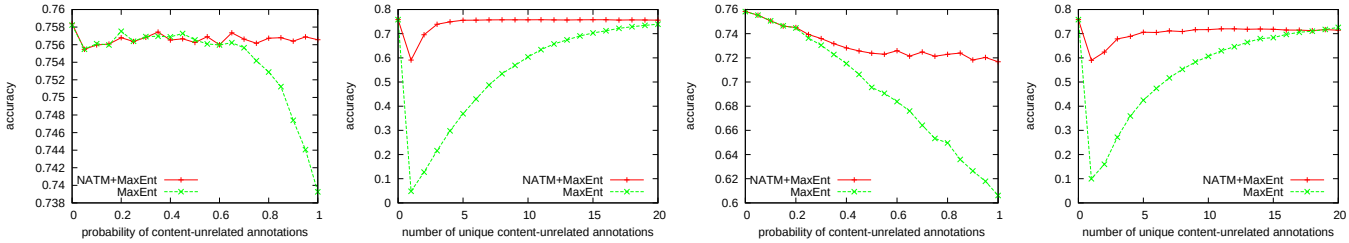
and a photo sharing service, namely Hatena, Delicious, and Flickr.

From the Hatena data, we used web pages and their annotations in Hatena::Bookmark, which is a social bookmarking service in Japan, which were collected using a similar method to that used in [20], [21]. Specifically, we first obtained a list of URLs of popular bookmarks for October 2008. We then obtained a list of users who had bookmarked the URLs in the list. Next, we obtained a new list of URLs that had been bookmarked by the users. By iterating the above process, we collected a set of web pages and their annotations. We omitted stop-words and words and annotations that occurred in fewer than ten documents. We omitted documents with fewer than ten unique words and also omitted those without annotations. The numbers of documents, unique words, and unique annotations were 39,132, 8,885, and 43,667, respectively. From the Delicious data, we used



(a) 20News-Dpart (b) 20News-Spart

Fig. 10. Estimated content-related annotation ratios with different probabilities of labeled annotations in a partially explicit setting.



(a) 20News-DP (b) 20News-DU (c) 20News-SP (d) 20News-SU

Fig. 11. Accuracies in 20News data with the NATM combined with the maximum entropy model.

web pages and their annotations that were collected using the same method as that used for the Hatena data. The numbers of documents, unique words, and unique annotations were 65,528, 30,274, and 21,454, respectively. From the Flickr data, we used photographs and their annotations provided in Flickr that were collected in November 2008 using the same method as that used for the Hatena data. We transformed photo images into visual words by using scale-invariant feature transformation (SIFT) [38] and k-means as described in [19]. We omitted annotations that were attached to fewer than ten images. The numbers of images, unique visual words, and unique annotations were 12,711, 200, and 2,197, respectively. For the experiments, we used 2,000 documents that were randomly sampled from each data set.

5.2 Results

The upper half of each table in Tables 4 and 5 shows probable content-unrelated annotations in the leftmost column, and probable annotations for some topics, which were estimated with the NATM using 50 topics. The lower half in Table 4 shows probable words in the content for each topic. With the Hatena data, we translated Japanese words into English, and we omitted words that had the same translated meaning in a topic. For content-unrelated annotations, words that seemed to be irrelevant to the content were extracted, such as 'toread', 'later', '*', '?', 'imported', '2008', 'nikon', and 'cannon'. Each topic has characteristic annotations and words, for example, Topic1 in the Hatena data is about economics, Topic2 is about cell-phone, and Topic3 is about music. Figure 12 shows some examples of the extraction of content-related annotations.

TABLE 7
Average frequency of the ten most probable content-unrelated annotations.

Data	NATM	Entropy
Hatena	4033.9	34.0
Delicious	10334.5	1386.3
Flickr	463.5	37.8

For the comparison, we show the ten highest entropy annotations estimated with the Corr-LDA using 50 topics, which were content-unrelated annotations estimated by the entropy-based method. Some annotations were content-unrelated, such as 'goodread', 'readthis' and 'reference' in Delicious, and 'canoneos20d' and '60mm' in Flickr. However, more content-related annotations were extracted as unrelated compared with the NATM. The entropy-based method was likely to extract low frequency annotations as shown in Table 7. Table 7 shows the average frequency of the ten most probable content-unrelated annotations with the NATM and the entropy-based method. This is because the variance of the estimated entropies for low frequency annotations are large with the entropy-based method. On the other hand, the proposed model is robust because the content-unrelated annotations are estimated in a Bayesian framework.

Figure 13 (a)(b)(c) shows the average perplexities over 100 experiments for held-out annotations in the three real social annotation data sets with different numbers of topics. Figure 13 (d) shows the result with Patent data as an example of data without content-unrelated annotations. The Patent data consist of patents published in Japan from January to March in 2004, to which Inter-

TABLE 4
The ten most probable content-unrelated annotations (leftmost column), and the ten most probable annotations for some topics (other columns), estimated with the NATM using 50 topics. Each column represents one topic. The lower half in (a) and (b) shows probable words in the content.

(a) Hatena

unrelated	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
toread	economics	mobile	music	oversea	science	programming
troll	finance	iPhone	perfume	international	technology	development
later	business	pc	sound	world	china	business
web	international	cell-phone	music-stream	military	science-technology	it
summary	japan	cell	content	korea	traffic	management
great	money	gadget	techno	war	tech	work
*	usa	hardware	audio	Japan	news	how-to-work
reference	biz	apple	sound	peninsula	sf	system-dev
?	stock	ipod	serious	east-asia	physics	technology
neta	market	network	CD	diplomacy	technology	dev
	yen	cell	music	japan	space	development
	year	yen	track	korea	year	system
	economics	handling	year	country	earth	web
	finance	usege	sound	person	experiment	information
	investment	on-board	CD	japanese	china	series
	japan	product	live	future	photo	technology
	market	phone	album	korea	change	change
	exchange	digital	listen	china	day	company
	bank	year	anonymous	military	world	people
	gold	pc	tour	say	earthquake	management

(b) Delicious

unrelated	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
reference	economics	iphone	food	environment	statistics	ruby
web	finance	mobile	recipe	science	data	programming
design	money	hardware	recipes	green	math	rails
imported	business	iPhone	cooking	sustainability	graph	php
tools	economy	ipod	Food	energy	visualisation	development
web2.0	financial	apple	dessert	home	visualization	opensource
toread	usa	games	Recipes	Technology	processing	framework
work	Finance	phone	Cooking	house	graphs	code
internet	recession	tech	baking	Environment	chart	python
cool	Money	gadget	cook	future	excel	rubyonrails
	money	iphone	1	energy	2	rails
	government	apple	recipe	green	1	php
	financial	2	recipes	rating	sample	web
	market	ipod	food	star	data	ruby
	economic	mobile	october	space	test	license
	crisis	game	cheese	solar	distribution	project
	credit	gps	2	power	size	django
	economy	phone	make	oil	population	mysql
	years	games	love	water	statistical	1
	business	blackberry	2008	system	probability	python

TABLE 5
The ten most probable content-unrelated annotations (leftmost column), and the ten most probable annotations for some topics (other columns), estimated with the NATM using 50 topics on Flickr.

unrelated	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
2008	music	night	autumn	beach	sky	family
canon	live	rock	park	water	clouds	portrait
nikon	paris	house	mountains	nature	spring	baby
bw	gig	park	leaves	bird	australia	friends
red	concert	mallory	canada	camping	lake	cute
nyc	show	coach	mountain	md	sunset	dof
blue	graffiti	inn	yellow	sun	soccer	black
color	fashion	creature	green	wildlife	d80	boy
sanfrancisco	bench	texas	river	backpacking	sea	lights
de	jg	concert	britishcolumbia	food	beach	dress

national Patent Classification (IPC) codes were attached by experts according to their content. The numbers of documents, unique words, and unique annotations (IPC codes) were 9,557, 104,621, and 6,117, respectively. With the Patent data, the perplexities of the NATM and the Corr-LDA were almost the same. On the other hand, with real social annotation data, the NATM achieved lower perplexities than the Corr-LDA, especially with

flash game games cool action fun flashgames free flashgame toread temp onlinegames online_game r
tools microsoft windows maintenance utilities software todo Microsoft update pc PC
barter swap imported business firefox:bookmarks
programming theory cs mathematics wikipedia in graph detection computer math reference cycle algorithms science
no_tag peru friends
blogging business businesses small Business Corporate no_tag blogs cookie
painting toread blogpost cool art sf graffiti publicart novelty sanfrancisco video
typo3 webdesign design gallery cms showcase portal gallerie inspiracion bestoftypo3 opensource via:mento.info css
online elearning courses
politics obama history newspapers election2008 reference news blogged photos global world History usa election

Fig. 12. Examples of content-related annotations in the Delicious data extracted by the NATM. Each row shows annotations attached to a document; content-unrelated annotations are shaded.

TABLE 6
The ten highest entropy annotations estimated with the Corr-LDA using 50 topics.

Hatena	XP, deferred, maikoh, PS, text editor, mouse, FirefoxAddOn, webdev, movable, technical tips
Delicious	relationship, newsletter, faster, stumble, goodread, Blogger, readthis, r, crossplatform, reference
Flickr	blonde, canoneos20d, 60mm, entertainment, 1855mm, scenic, advertising, macro, little, sunday

Hatena and Delicious data. This result implies that it is important to consider relevance to the content when analyzing noisy social annotation data. The perplexity of the Corr-LDA with social annotation data becomes worse as the number of topics increases because the Corr-LDA overfits noisy content-unrelated annotations. With the Flickr data, the difference in perplexities was small, because it is more difficult to extract features from image data than text data. In text data, words are used to represent the content, where most words have some semantics. On the other hand, some visual words in image data might not have semantics, and they might fail to represent the content.

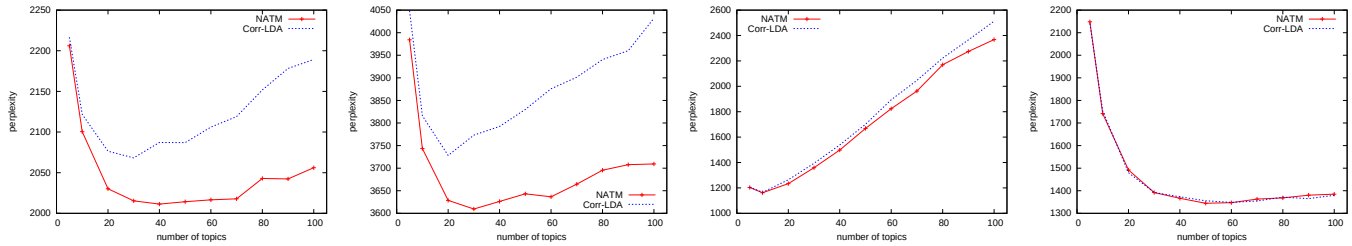
6 Conclusion

We have proposed a topic model for extracting content-related annotations from noisy annotated data. The proposed model can be applied in both implicit and partially explicit relevance settings, and it can also be used as the preprocessing for different classifiers as well as for modeling noisy annotated data. We have confirmed experimentally that the proposed method can extract content-related annotations appropriately, and can be used for analyzing social annotation data.

Although our results have been encouraging to date, we must extend our approach in a number of directions. First, we want to determine the number of topics automatically by extending the proposed model to a non-parametric Bayesian model such as the Dirichlet process mixture model [39]. Second, we want to incorporate user information into the model for modeling social annotation data. Third, a framework to deal with content-unrelated annotations can be used in models other than topic models. Finally, since the proposed method is theoretically applicable to various kinds of annotation data, we will confirm this in additional experiments.

References

[1] Delicious, http://delicious.com.
[2] Flickr, http://flickr.com.
[3] Technorati, http://technorati.com.
[4] YouTube, http://www.youtube.com.
[5] Last.fm, http://www.last.fm.
[6] CiteULike, http://www.citeulike.org.
[7] S. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, 2006.
[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.
[9] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 127–134.
[10] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *CVPR '04: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 1002–1009.
[11] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2003, pp. 119–126.
[12] J. Jeon and R. Manmatha, "Using maximum entropy for automatic image annotation," in *CIVR '04: Proceedings of the 3rd International Conference on Image and Video Retrieval*, 2004, pp. 24–32.
[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[14] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI '99: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.
[15] ——, "Collaborative filtering via Gaussian probabilistic latent semantic analysis," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2003, pp. 259–266.
[16] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior," in *IJCAI '09: Proceedings of 21st International Joint Conference on Artificial Intelligence*, 2009, pp. 1427–1432.
[17] T. Iwata, T. Yamada, and N. Ueda, "Probabilistic latent semantic visualization: topic model for visualizing documents," in *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008, pp. 363–371.
[18] V. N. Vapnik, *The nature of statistical learning theory*. Springer, 1995.
[19] G. Csurka, C. Dance, J. Willamowski, L. Fan, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.



(a) Hatena (b) Delicious (c) Flickr (d) Patent

Fig. 13. Perplexities of held-out annotations with different numbers of topics in social annotation data (a)(b)(c), and in data without content unrelated annotations (d).

[20] X. Wu, L. Zhang, and Y. Yu, "Exploring social annotations for the semantic web," in *WWW '06: Proceedings of the 15th International Conference on World Wide Web*. ACM, 2006, pp. 417–426.

[21] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles, "Exploring social annotations for information retrieval," in *WWW '08: Proceedings of the 17th International Conference on World Wide Web*. ACM, 2008, pp. 715–724.

[22] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating topics and syntax," in *In In Advances in Neural Information Processing Systems*, vol. 17, 2005, pp. 537–544.

[23] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 241–248.

[24] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *Proceedings of the 24th international conference on Machine learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 233–240.

[25] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA, USA: IEEE Computer Society, 2009, pp. 2036–2043.

[26] N. Nguyen and R. Caruana, "Classification with partial labels," in *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2008, pp. 551–559.

[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101 Suppl 1, pp. 5228–5235, 2004.

[28] T. Minka, "Estimating a Dirichlet distribution," M.I.T., Tech. Rep., 2000.

[29] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999, pp. 61–67.

[30] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[31] R. Raina, Y. Shen, A. Y. Ng, and A. McCallum, "Classification with hybrid generative / discriminative models," in *Advances in Neural Information Processing Systems*, 2004.

[32] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE. Transactions on Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.

[33] X. Yang, Q. Song, and Y. Wang, "Weighted support vector machine for data classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 5, pp. 961–976, 2007.

[34] T. Iwata, K. Saito, and T. Yamada, "Recommendation method for extending subscription periods," in *KDD '06: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Miniining*, 2006, pp. 574–579.

[35] S. F. Chen and R. Rosenfeld, "A Gaussian prior for smoothing maximum entropy models," Computer Science Department, Carnegie Mellon University, Tech. Rep. CMUCS–99–108, 1999.

[36] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 3, pp. 503–528, 1989.

[37] K. Lang, "NewsWeeder: learning to filter netnews," in *ICML '95: Proceedings of the 12th International Conference on Machine Learning*, 1995, pp. 331–339.

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[39] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

PLACE PHOTO HERE

Tomoharu Iwata received the B.S. degree in environmental information from Keio University in 2001, the M.S. degree in arts and sciences from the University of Tokyo in 2003, and the Ph.D. degree in informatics from Kyoto University in 2008. He is currently a research scientist at Learning and Intelligent Systems Research Group of NTT Communication Science Laboratories, Kyoto, Japan. His research interests include data mining, machine learning, information visualization, and recommender systems.

PLACE PHOTO HERE

Takeshi Yamada received the B.S. degree in mathematics from the University of Tokyo in 1988 and the Ph.D. degree in informatics from Kyoto University in 2003. He was a Leader of Emergent Learning and Systems Research Group of NTT Communication Science Laboratories and is currently a Senior Manager of NTT Science and Core Technology Laboratory Group. His research interests include Data Mining, Statistical Machine Learning, Graph Visualization, Metaheuristics and Combinatorial Optimization.

PLACE PHOTO HERE

Naonori Ueda received the B.S., M.S., and Ph D degrees in Communication Engineering from Osaka University, Osaka, Japan, in 1982, 1984, and 1992, respectively. In 1984, he joined the Electrical Communication Laboratories, NTT, Japan. In 1991, he joined the NTT Communication Science Laboratories. His current research interests include parametric and non-parametric Bayesian approach to machine learning, pattern recognition, data mining, signal processing, and cyber-physical systems. Currently, he is a director of NTT Communication Science Laboratories.