Supervised Topic Modeling for Question Labeling

Junjia He
Shanghai Jiao Tong University
Shanghai, China
edfward@gmail.com

Kenny Q. Zhu
Shanghai Jiao Tong University
Shanghai, China
kzhu@cs.sjtu.edu.cn

ABSTRACT

Multi-label classification is a very common problem in Machine Learning community. It requires given an instance the model should be able to generate corresponding labels according to the instance's features. Meanwhile, topic model tries to understand the latent semantics of an observed document, which could be used to describe the similarity of documents, avoid the influences of polysemy or synonyms, or cluster documents in a meaningful way. Nowadays using topic model to do multi-label classification on documents gathers popularity gradually.

This paper attempts to tackle the labeling problem on online Question-Answer sites such as *Stack Overflow*, and tries to establish a new model improving current efforts on adapting unsupervised topic model to supervised one. Specifically, this paper believed that the question documents in online QA sites should be modeled according to both its title and body but in a different way. And we proposed a new supervised topic model integrating the title information (called *qinfo* in the corresponding context) and proved in certain conditions the new model could have better performances.

Keywords

Multi-label classification; Topic model; Probabilistic graphical model;

1. INTRODUCTION

This paper focuses on the question labeling issue on Internet Q&A site. Currently most if not all questions are labeled by the users themselves or the website editors, which requires additional human resources and cannot guarantee the accuracy. Therefore this paper hopes to model the question texts (hereinafter collectively referred to as "documents", including the question's title and body) accurately and extract necessary features to label the documents as a multi-label classification problem. Taking into account that a considerable number of documents have already been manually labeled, we choose supervised learning to better utilize those information. On the other hand, since interdisciplinary document classification is too difficult, we restrict the target documents in a specific area (like programming or mathematical problems), such that the number of labels are under control.

Since the underlying task is a multi-label classification problem for large-scale documents, the probabilistic topic model seems a feasible idea. This paper provides attempts to improve the topic model to better express target documents which contain titles in addition to ordinary document bodies, trying to acquire knowledge beyond simple texts by analyzing its latent semantics to improve the labeling results.

The multi-label classification now becomes a popular topic in Machine Learning communities, such as gene prediction or news article tagging. Most approaches fall into two categories: problem transformation methods and algorithm adaptation methods. The former approach tries to transform the problems to a set of binary classification problems while the latter tries to adapt the existing algorithms for single-label classification to ones that are capable of predicting multi labels, such as Multi-Label k Nearest Neighbors (MLkNN) [16] based on kNN algorithm and Back-Propagation Multi-Label Learning (BPMLL) [15] based on back-propagation algorithm.

Meanwhile the topic model aims to understand the texts by modeling its hidden semantics, which could be used to describe the text similarity, to avoid the impact of polysemy or synonyms, or to cluster the texts. Those latent semantics summarize and abstract the semantics of the documents, representing a underlying form of the document content. Afterwards, along with the development of statistical language models, latent semantic is interpreted as a probability distribution over the dictionary while every document corresponds to a probability distribution on the semantic space. Based on this notion, Probabilistic Latent Semantic Analysis (pLSA)[7] and Latent Dirichlet Allocation (LDA)[4] have been proposed to depict the topic structure behind texts.

Also notice that question documents in Q&A sites have some other interesting properties, and we'd like to emphasize the functionality of a document's title. Usually users on those sites tend to describe their problems succinctly yet informative, for example they would very likely tell how the problems behave, or in what software, or what the wrong message is. In addition, since the document set we used mainly focused on programming or mathematics, many question titles would contain crucial keywords (like some named entities), making the titles more valuable for labeling. For example if the noun "stemming" tends to appear in the title with the label *nlp*, for all other questions having "stemming" in the title without label *nlp*, the topic assignment for words in those documents could also have *nlp* as their choice.

In this paper, we propose a model based on L-LDA for documents having structures like *title-body-labels*, in which we treat titles separately and model them in another way. At first we be-

lieve the key information in document titles would be named entities (along with some useful nouns) which we would refer to as *qinfo* hereafter in this paper. Then we consider *qinfo* is sampled independently from words, and there is another distribution called *qinfo*-label distribution which is used to describe the probabilistic relations between *qinfo* and topics. Based on this idea, we implemented a prototype model and showed that the results have some relative advantages over baseline models and aforementioned L-LDA.

2. APPROACH

In this section we introduce some background knowledge on the topic model and its supervised variants. And later we propose our model utilizing the title information for question documents.

2.1 Supervised topic models

At first we show how to adapt the unsupervised topic model to a supervised one which could establish the correspondence between topics and labels as proposed in [11]. Then we extend the model to incorporate information provided by document titles.

We describe each document $d \in \{1, \cdots, D\}$ as a multinomial distribution $\theta^{(d)}$ on labels, and each label (same as the topic in conventional LDA's terms) $\beta \in \{1, \cdots, K\}$ as another multinomial distribution over words. Noted that K in standard LDA is supposed to be provided by users, while here it's the number of unique labels in the document set. Then document generation process is as following:

1. For every label k, sample a multinomial label-word distribution $\beta_k \sim Dir(\cdot|\eta)$;
2. For each document d, sample a multinomial document-label distribution $\theta^{(d)} \sim Dir(\cdot|\alpha^{(d)})$, where entries in $\alpha^{(d)}$ are non-zero if and only if corresponding labels are observed;
3. In document d, do following steps to generate every word:
   (a) sample a label z for this word position from $\theta^{(d)}$;
   (b) sample a word w for this word position from β_z.

2.2 Incorporating question titles

First we want to emphasize the functionality of a document's title, especially for those on Q&A websites. Users on those sites tend to describe their problems succinctly yet informative, for example they would very likely tell how the problems behave, or in what software, or what the wrong message is. In addition, since the document set we used mainly focused on programming or mathematics, many question titles would contain crucial keywords (like some named entities), making the titles more valuable for making decisions.

In this paper, we proposed a model called Q-DLA (Question-LDA) based on L-LDA for documents having structures like *title-body-labels*, in which we treated titles separately and model them in another way. Our model is depicted using graphical model notation in Figure 1. In addition to original L-LDA's parameters we have random variables describing the title information and its corresponding relations to labels, thus the topic mixture for each document is dependent not only on the Dirichlet prior and observed labels but also on observed titles.

At first we believe the key information in document titles would be named entities (along with some useful nouns) which we would refer to as *qinfo* hereafter in this paper. While all other parameters are the same as in L-LDA, to better illustrate Figure 1, we classify the random variables into 3 categories:

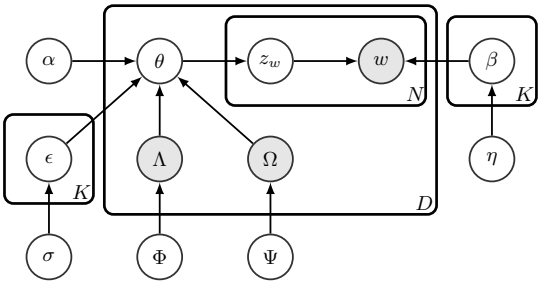


Figure 1: Graphical model of Labeled Q-LDA: both observed labels Λ and *qinfo* Ω would influence the topic mixture θ

- **Variables of standard LDA**: α, η are the hyper-parameters of Dirichlet distributions as the prior for label-word distribution β and document-label distribution θ. z is label assigned to every word and w is the observed word;
- **Variables describing labels**: Φ is the Bernoulli prior generating observed labels Λ;
- **Variables describing *qinfo***: Ω denotes the observed *qinfo* in the title sampled by the Bernoulli prior Ψ, and ϵ denotes the *qinfo*-label distribution sampled by the prior $Dir(\cdot|\sigma)$.

And the complete document generation process of our model is:

1. For every label k, sample label-word distribution $\beta_k \sim Dir(\cdot|\eta)$;
2. For every *qinfo* i, sample *qinfo*-label distribution $\epsilon_i \sim Dir(\cdot|\sigma)$;
3. For each document d,
   (a) sample document-label distribution $\theta'^{(d)} \sim Dir(\cdot|\alpha^{(d)})$, where entries in $\alpha^{(d)}$ are non-zero if and only if corresponding labels are observed;
   (b) sample the *qinfo* for this document $\Omega^{(d)} \sim Bernoulli(\cdot|\Psi)$, and denote them as $Q = \{i|\Omega_i^{(d)} = 1\}$;
   (c) generate the final label mixture from the restricted label distribution and *qinfo*-incurred label distribution $\theta^{(d)} = \mathcal{F}(\theta'^{(d)}, \sum_{i=1}^{|Q|} \epsilon_{Q_i})$.
4. In document d, do following steps to generate every word:
   (a) sample a label z for this word position from $\theta^{(d)}$;
   (b) sample a word w for this word position from β_z.

In another words, while L-LDA tries to restrict the topic distribution using observed labels as the mask, knowledge about *qinfo* Ω would provide more information to give every word in that document more freedom of assigning topics. The topic distribution $\theta^{(d)}$ is drawn like this:

$$\text{step 1:} \quad \boldsymbol{\theta'}^{(d)} = (\theta_{l_1}, \cdots, \theta_{l_{M_d}})^T \sim Dir(L^{(d)} \times \boldsymbol{\alpha}) \quad (1)$$

$$\text{step 2:} \quad \boldsymbol{\theta}^{(d)} = \mathcal{F}(\boldsymbol{\theta'}^{(d)}, \sum_{i=1}^{|Q|} \epsilon_{Q_i}). \quad (2)$$

There are two points worth noticing: in Equation 1, $L^{(d)} \times \boldsymbol{\alpha}$ means the masked label distribution by observed labels, and in Equation 2, we did not specify the exact relations between the restricted $\theta'^{(d)}$ after step 1 from L-LDA and the accumulated distributions using observed *qinfo* after renormalization, since a specific

function of those two would make the model harder to solve using Gibbs sampling. To illustrate our intuition that *qinfo* from titles would indeed improve the classification results, we only need to prove they are somehow relevant and the model parameters should reflect such relevance, and we did this when inferring the model parameters in Gibbs sampling.

To illustrate why our model works, we selected a question as the example.

```
Title: Stemming text in java
Body : im searching for a possibility to stemm
strings in java. First I wanted to do it with
lucene but all the examples I found in the web...
Label: lucene, stemming
```

After pre-calculating the *qinfo*-label distribution of *stemming*, label *nlp* has a relatively high probability in its distribution. So in the question "Stemming text in java" with labels *lucene* and *stemming* shown above, instead of assigning only those two labels for each word and sampling iteratively, Q-LDA would also assign labels *nlp* since in its title the *qinfo* "Stemming" would contribute such label choice. In this way, the label-word distribution β after inference could avoid being skewed by expanding the label choice for each word.

2.3 Learning the model

The learning algorithm for the full model is difficult since the way Ω affects $\theta^{(d)}$ is intended to be indefinite, therefore we used a simplified model as the prototype, where the *qinfo*-label distribution is observed. This can be achieved by simply counting the co-occurrences of *qinfo* and labels, then normalize them to be the desired *qinfo*-label distribution.

Thus, in Gibbs sampling process, when determining target topics for a particular document to sample, besides ones specified by labels $\Lambda^{(d)}$, the model also samples topics added from $\sum_{i=1}^{|\Omega^{(d)}|} \epsilon_{\Omega_i^{(d)}}$. And we added those topics in a very intuitive way - first from the *qinfo* we added the corresponding distributions together and renormalized them, then from this distribution we pick ones with highest probability.

In this way, the topic distribution for every document is not only influenced by the labels but also titles. Furthermore, the pre-calculated *qinfo*-label distributions in a certain degree could describe the dependencies among labels, thus integrating them in the model would probably improve the word-label assignment process.

2.4 Multi-label classification

We take the same strategy as L-LDA to do classification, that is when encountering new documents without labels we do the Gibbs sampling process like normal LDA instead of enumerating all possible label assignments and pick the one with maximum posterior probability. This method is reasonable since it does not only reduce the computational cost but also approximate the original way - sampling from all labels is similar to trying all label assignments.

3. IMPLEMENTATION DETAILS

In this part we give a general view on how we implemented Q-LDA. Since our model is based on L-LDA, we modified the corresponding algorithm in Stanford Topic Modeling Toolbox.

3.1 Preprocessing

As section 2.3 noted, instead of learning the full model during the training, we simplified the problem by making *qinfo*-topic distribution ϵ observed. We did this in two steps:

1. We first parsed all document titles and then extracted the named entities using NER tools, and attached them as *qinfo* with every document together with its labels both serving as a document's features;

2. Next we count the co-occurrences between *qinfo* and labels, thus having a probability distribution of every *qinfo* in terms of labels.

3.2 Sampling

After acquiring *qinfo*-topic distribution ϵ, we can start the Gibbs sampling process. The major difference is that before deciding what topics to sample for a document, we calculated another distribution from all of its *qinfo*: that is, we numerically summed the distribution vectors of each *qinfo* and then renormalize it. And from this distribution we pick a certain number of topics (usually $max(|Q^{(d)}|, 5)$, as a empirical parameter) with highes probability in addition to the restricted topics specified by observed labels Λ. And then everyting follows the same as L-LDA.

4. EVALUATION

We extracted 50,000 questions from *Stack Overflow* with titles, bodies and labels as our experiment dataset. To demonstrate our model's effectiveness on question texts, we excluded codes appeared in the body of questions and corresponding labels of programming languages. The frequency distribution of the dataset is indicated in Figure 2(a) and 2(b).

We compare the results with two other models, one is "one-vs-all" SVM classifiers without parameter tunning and Labeled LDA from Stanford Topic Modeling Toolbox. And the results are indicated in Table 1.

Table 1: F_1 scores for three models

SVM	L-LDA	**Q-LDA**
0.540	0.613	**0.620**

Noted in our evaluation F_1 score is defined as

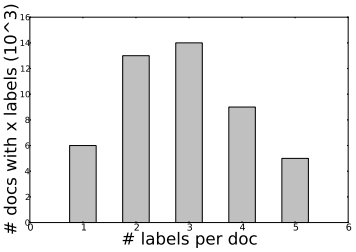
$$F_1 = \frac{1}{K}\sum_{\lambda=1}^{K}\frac{2 \times \text{true positive}}{2 \times \text{true positive} + \text{false negative} + \text{false positive}} \tag{3}$$

which is the macro averaged F_1 score for all labels.

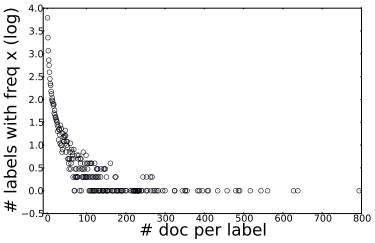
5. RELATED WORK

Traditional multi-label classification models which transform the problem to a series of binary classification problems like "one-vs-rest" SVMs [13] usually have a dramatic performance drop when the label set has a large size and a skewed frequency distribution. As [9] stated, the extremely rare categories introduced by the skewed distribution would make the SVM classifier unacceptable. Therefore when dealing with real world data, models which had good performances in traditional benchmark datasets are not satisfying.

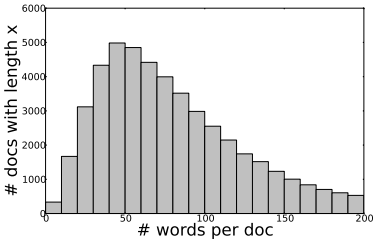
Therefore researchers have turned to generative approaches based on LDA [4] for document classification, while several adaptation have been made such that this model could be used in a supervised context [10, 3, 11, 12], where given a multi-labeled corpus the word-label distribution and document-label distribution could be inferred. Ramage et al. proposed Labeled-LDA (L-LDA)[11] trying to establish a one-to-one correspondence between topics the original LDA learned and the labels which are provided manually. They did this by restricting the document-topic distribution for each



(a) Number of documents that have L labels



(b) Number of labels that appear in D documents



(c) Number of documents with different sizes

Figure 2: Caption

document such that only topics specified by the document labels could be counted, while all others are masked. In this way, every word in this particular document could only be assigned topics corresponding to the labels, and the topic-label correspondence becomes obvious. On the other hand MM-LDA[12] could also be used for multi-label classification as it's not constrained to one label per document, but their learned topics don't correspond with the label set.

Models describing the dependencies among tags are usually another kind of extensions of LDA such as [1, 8, 2, 6]. Ghamrawi and McCallum [6] proposed a CRF model which could be used for multi-label classification, while the following work [5] utilized unlabeled data such the model became a semi-supervised learning one. On the other hand [14] used a hybrid generative-discriminative approach where separate classifiers have been trained in a Bayesian network and accumulated in the topological order.

6. CONCLUSION

We presented a multi-label classifier based on LDA which incorporates information provided in document title. The main idea is to expand the label choice for every word during the training phase, instead of being restricted solely by observed labels. Therefore we use the title information (we called *qinfo*) to provide information on label dependencies in a certain degree. We demonstrated such incorporation could improve the classification results. However there remains much work to do. Currently we didn't specify the relation between the *qinfo*-incurred distribution and the standard label distribution calculated by L-LDA but only followed the intuition that they are somehow related, therefore we could improve our algorithm by modeling their relations in a definite way and giving quantitative descriptions. This will be included in our future investigation.

7. REFERENCES

[1] D. Blei and J. Lafferty. Correlated topic models. *NIPS*, 18:147, 2006.

[2] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, volume 16, 2003.

[3] D. M. Blei and J. D. McAuliffe. Supervised topic models. In *NIPS*, volume 7, pages 121–128, 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[5] G. Druck, C. Pal, A. McCallum, and X. Zhu. Semi-supervised classification with hybrid generative/discriminative methods. In *KDD*, pages 280–289. ACM, 2007.

[6] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *CIKM*, pages 195–200. ACM, 2005.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM, 1999.

[8] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. 2006.

[9] T.-Y. Liu, Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter*, 7(1):36–43, 2005.

[10] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

[11] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, pages 248–256. Association for Computational Linguistics, 2009.

[12] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *WSDM*, pages 54–63. ACM, 2009.

[13] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *JMLR*, 5:101–141, 2004.

[14] M.-L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *KDD*, pages 999–1008. ACM, 2010.

[15] M.-L. Zhang and Z.-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *Knowledge and Data Engineering, IEEE Transactions on*, 18(10):1338–1351, 2006.

[16] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.