# Identifying New Categories in Community Question Answering Archives: A Topic Modeling Approach

Yajie Miao * , Chunping Li * , Jie Tang + , Lili Zhao *
Tsinghua National Laboratory for Information Science and Technology (TNList)
* School of Software, Tsinghua University
+ Department of Computer Science and Technology, Tsinghua University
yajiemiao@gmail.com, {cli, jietang}@tsinghua.edu.cn, zhaoll07@mails.tsinghua.edu.cn

## ABSTRACT
Community Question Answering (CQA) services have evolved into a popular way of information seeking and providing. User-posted questions in CQA are generally organized into hierarchical categories. In this paper, we define and study a novel problem which is referred to as New Category Identification (NCI) in CQA question archives. New Category Identification is primarily concerned with detecting and characterizing new or emerging categories which are not included in the existing category hierarchy. We define this problem formally, and propose both unsupervised and semi-supervised topic modeling methods to solve it. Experiments with a ground-truth set built from Yahoo! Answers show that our methods identify and interpret new categories effectively.

## Categories and Subject Descriptors
H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.*

## General Terms
Algorithms, Performance, Experimentation.

## Keywords
Community Question Answering, Topic Modeling, New Category.

## 1. INTRODUCTION
With the blooming of Web 2.0, user-generated contents (UGC) such as Wikipedia, YouTube and Flickr begin to flourish. One type of UGC sites is the Community Question Answering (CQA) services, which enable users to post or answer questions on various subjects. Among CQA websites, Yahoo! Answers is now becoming the most popular portal. Since its launch in 2005, Yahoo! Answers has attracted millions of users, and has stored a tremendous number of community questions in its database. As the volume of these questions is growing to an intractably huge size, how to manage them efficiently and effectively has become an increasingly important research issue.

In Yahoo! Answers, the community questions are organized in the form of hierarchical categories. However, the maintenance of this category hierarchy highly relies on human efforts, and usually its structures remain unchanged in a fairly long period. Consequently, the current categories are definitely unable to capture newly-arising topics which are attracting intensive public attention. Questions belonging to neither of the existing categories would be assigned by users into the pseudo "*Other*" category, e.g., "Other-Internet" in the "Internet" domain. These accumulating *Other questions* bring difficulties and inconvenience to both users and CQA service providers. In this paper, we study extensively the problem of New Category Identification (NCI) in CQA, which aims to find potential categories not included currently. This problem subsumes interesting applications in that the CQA category structures can be enriched and refined continuously.

Recently, there has been a growing amount of research [2, 3, 4, 5, 8, 9, 10, 11] on Community Question Answering. However, there are yet no mechanisms with which we can find new categories in the CQA question archives. In this study, we formulate the novel NCI problem as a topic modeling issue. We first adapt PLSA, a basic algorithm in the context of topic modeling, to New Category Identification. Essentially, PLSA is an unsupervised method, i.e., we are unaware of the meaning of the generated categories, even after the model has been fully estimated. We develop semi-supervised topic modeling methods to solve this challenge. Specifically, we cast the prior knowledge about specific categories into PLSA in a probabilistic manner and fit the model to the question collection with Maximum A Posterior (MAP) estimation. With the model estimated, we can then naturally reach the categories to be identified.

The rest of the paper is organized as follows. In Section 2, we formally define the problem of New Category Identification. After that, we present the methods in Section 3. Then, we show and discuss the experimental results in Section 4. Finally, we have the conclusion and future work in Section 5.

## 2. PROBLEM DEFINITION
In Yahoo! Answers, one community question usually consists of three parts, i.e., the *subject* (a brief statement of the question), the *content* (additional detailed descriptions of the question) and the *answers* posted by other users. We define the *associated text* of a community question as the concatenation of its subject, content and answers.

Formally, we let $D$ denote the domain we are interested in, and $Q = \{q_1, q_2, \cdots\}$ denote a collection of Other questions in $D$. When considering the associated text, each question $q \in Q$ can be described as a word vector as follows.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*CIKM'10*, October 26-30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10...$10.00.

$$q = \{c(w_1, q), c(w_2, q), \cdots, c(w_{|V|}, q)\}, \quad (1)$$

where $c(w_h, q)$ is the number of occurrences of word $w_h$ in the associated text of question $q$, and $V$ is the whole set of words in the collection $Q$.

Our basic idea is to perform topic modeling on the collection $Q$ and group Other questions into distinct topics. After topic modeling, the topical structures of the collection $Q$ can be represented as

$$Q = \{G_1, G_2, \cdots, G_k\}, \quad (2)$$

where each question group $G_i$ corresponds to an underlying topic in $Q$. Then qualified question groups are selected out as categories that we have identified. We define three criteria for group selection. These criteria can be found in a more detailed version[1]of this paper.

# 3. IDENTIFYING METHODS

## 3.1 Unsupervised Topic Modeling

Probabilistic latent semantic analysis (PLSA) [1] has been applied to topic modeling with promising results [6, 7, 12]. For the NCI problem, our idea is to use a unigram language model (a multinomial word distribution) to model a group (topic). To be consistent with previous literature, we still define the $k$ unigram language models as $\Theta = \{\theta_1, \theta_2, \cdots, \theta_k\}$ which capture individual groups. Then each word $w_h$ in question $q$ is generated from a two-stage process: first, a group $\theta_j$ is chosen conditionally for the question according to $\pi_{q,j}$; second, the word $w_h$ is generated from $\theta_j$ according to the conditional probability $p(w_h \mid \theta_j)$. From a statistical perspective, the question collection $Q$ is the observed data, and its log-likelihood is described as

$$\begin{aligned} L(Q \mid \Lambda) &= \sum_{q \in Q} \sum_{w_h \in V} \{c(w_h, q) \times \log \sum_{j=1}^{k} [\pi_{q,j} \cdot p(w_h \mid \theta_j)]\} \\ &= \sum_{q \in Q} \sum_{w_h \in V} \{c(w_h, q) \times \log \sum_{j=1}^{k} [\pi_{q,j} \cdot p(w_h \mid \theta_j)]\}, \end{aligned} \quad (3)$$

where $\Lambda$ represents the set of model parameters, $\pi_{q,j}$ actually measures the conditional probability of choosing $\theta_j$ given $q$.

We perform Maximum Likelihood Estimation using the EM algorithm to estimate the model. The latent variable $z_{q,w_h}$ is defined as the group from which the word $w_h$ in question $q$ is generated. During the estimation process, the model parameters are updated iteratively as follows.

**E-step:** $$p(z_{q,w_h} = j) = \frac{\pi_{q,j}^{(n)} p^{(n)}(w_h \mid \theta_j)}{\sum_{j'=1}^{k} \pi_{q,j'}^{(n)} p^{(n)}(w_h \mid \theta_{j'})}$$

**M-step:** $$\pi_{q,j}^{(n+1)} = \frac{\sum_{w_h \in V} c(w_h, q) p(z_{q,w_h} = j)}{\sum_{j'}^{k} \sum_{w_h \in V} c(w_h, q) p(z_{q,w_h} = j')}$$

$$p^{(n+1)}(w_h \mid \theta_j) = \frac{\sum_{q \in Q} c(w_h, q) p(z_{q,w_h} = j)}{\sum_{w_h' \in V} \sum_{q \in Q} c(w_h', q) p(z_{q,w_h'} = j)} \quad (4)$$

where $p(z_{q,w_h} = j)$ represents the probability that the word $w_h$ in question $q$ is generated from the $j^{th}$ group.

When the iterative estimation process converges, we assign question $q$ into the group which has the largest $\pi_{q,j}$, and the group mapping function is

$$group(q) = \arg\max_j (\pi_{q,j}). \quad (5)$$

In the prior-PLSA method, we adopt this group assignment strategy as well.

## 3.2 Semi-supervised Topic Modeling

In many application scenarios, users indeed know what potential categories they are interested in. For example, we want to know whether "Twitter" should be a new category in the "Internet" domain. In this case, it would be nice if we establish "Twitter" as a predefined facet and guide topic modeling with this prior knowledge. From PLSA, we see that the estimation results are language models whose elements are $p(w_h \mid \theta_j)$. Therefore, it is natural to also input prior knowledge as language models. Specifically, we may want to input $\bar{\theta}_j$ as the prior topic model for topic $j$ given by the user. For "Twitter", $\bar{\theta}_j$ assign high probability to words like "tweet", "follow", "user", "message", etc. In practice, it may be infeasible to build these prior distributions manually. In our detailed version, we also formulate how to get prior knowledge automatically from Wikipedia.

On the language model $\theta_j$, we define a Dirichlet prior $Dir(\{1 + \mu_j \cdot p(w_h \mid \bar{\theta}_j)\}_{w_h \in V})$ using $\bar{\theta}_j$, where the factor $\mu_j$ indicates how strong our confidence is on the prior $\bar{\theta}_j$. This prior determines the probability of a specific setting of $\theta_j$ and therefore is called a *distribution of distribution*. Dirichlet distribution is a conjugate prior for multinomial distribution and we will see the advantage of this conjugacy in parameter estimation. Then the probability of $\theta_j$ can be formulated as

$$p(\theta_j) = \frac{1}{\mathrm{B}(\{1 + \mu_j p(w_h \mid \bar{\theta}_j)\}_{w_h \in V})} \cdot \prod_{w_h \in V} p(w_h \mid \theta_j)^{\mu_j p(w_h \mid \bar{\theta}_j)} \quad (6)$$

where the beta function $\mathrm{B}(\{1 + \mu_j p(w_i \mid \bar{\theta}_j)\}_{w_i \in V})$ is a constant normalizing factor which can be expressed into combination of gamma functions. In order to keep the form of the prior uncluttered, we omit this factor. Then the prior for the whole parameter set $\Lambda$ is

$$p(\Lambda) = \prod_{j=1}^{k} p(\theta_j) \propto \prod_{j=1}^{k} \prod_{w_h \in V} p(w_h \mid \theta_j)^{\mu_j p(w_h \mid \bar{\theta}_j)} \quad (7)$$

With the prior defined above, we turn to Bayesian inference to maximize the posterior probability of the parameters $\Lambda$ after we have been give the observed data $Q$, rather than maximize the likelihood of $Q$. For parameter estimation, we need to find a set of parameters $\hat{\Lambda}$ as

[1] http://dm.thss.tsinghua.edu.cn:8080/yajiem/cikm-detailed.pdf

$$\hat{\Lambda} = \arg\max_{\Lambda} p(\Lambda \mid Q) = \arg\max_{\Lambda} p(Q \mid \Lambda) p(\Lambda). \quad (8)$$

Then for the log-likelihood value, we have

$$L(\Lambda \mid Q) = L(Q \mid \Lambda) + L(\Lambda) + const, \quad (9)$$

where $L(Q \mid \Lambda)$ remains the same as Equation (3), $L(\Lambda)$ is the log-likelihood for the prior $p(\Lambda)$, and *const* is a constant value.

We can use the Maximum A Posterior (MAP) estimator to obtain the parameters. As revealed by Equation (7), $L(\Lambda)$ is independent of the latent variable $z_{q,w_h}$ and the topic distribution $\pi_{q,j}$. So the introduction of the prior only affects the estimation of $p(w_h \mid \theta_j)$. The overall MAP estimation is performed by rewriting the updating formula for $p(w_h \mid \theta_j)$. For completeness, we give the updating formulas as follows.

**E-step:** $$p(z_{q,w_h} = j) = \frac{\pi_{q,j}^{(n)} p^{(n)}(w_h \mid \theta_j)}{\sum_{j'=1}^{k} \pi_{q,j'}^{(n)} p^{(n)}(w_h \mid \theta_{j'})}$$

**M-step:** $$\pi_{q,j}^{(n+1)} = \frac{\sum_{w_h \in V} c(w_h, q) p(z_{q,w_h} = j)}{\sum_{j'}^{k} \sum_{w_h \in V} c(w_h, q) p(z_{q,w_h} = j')}$$

$$p^{(n+1)}(w_h \mid \theta_j) = \frac{\sum_{q \in Q} c(w_h, q) p(z_{q,w_h} = j) + \mu_j \cdot p(w_h \mid \bar{\theta}_j)}{\sum_{w_h' \in V} \sum_{q \in Q} c(w_h', q) p(z_{q,w_h'} = j) + \mu_j} \quad (10)$$

Due to usage of a conjugate prior, we can see that the updating formula for $p(w_h \mid \theta_j)$ here has the similar form with that of PLSA. An instructive interpretation of this formula is: for each topic $\theta_j$, we observe an additional pseudo community question, whose size (number of words) is $\mu_j$ and whose word distribution follows the prior distribution $p(w_h \mid \bar{\theta}_j)$. Therefore, the conjugacy of the Dirichlet prior allows for a tractable and interpretable solution to the MAP estimation.

After the model is estimated, we can directly reach the categories that we expect to identify, without any indication from representative words. For instance, $\bar{\theta}_j$ carries our prior knowledge about "Twitter". Then the question group $G_j$, corresponding to $\theta_j$, is exactly on the topic "Twitter". For convenience, this semi-supervised method is called *prior-PLSA*.

# 4. EXPERIMENTS

## 4.1 Dataset and Preprocessing

With the APIs[2] provided by Yahoo! Developer Network, we create an inclusive dataset by downloading 6000 questions from "Other-Internet" in the "Internet" domain. These questions have been issued over a period from January to April 2010. We only focus on the *resolved* questions, meaning questions that have been given their best answers. For preprocessing, we perform *document frequency* feature selection on the vocabulary: those words which appear in less than three questions are removed.

[2] http://developer.yahoo.com/answers/

## 4.2 Sample Results

We first run the unsupervised method on the dataset. In Table 1, we present the sample results of three groups that are generated by PLSA and filtered with the three criteria. The table shows the top 10 questions that are ranked according to $\pi_{q,j}$. We can discover and interpret the three categories in a meaningful way. The first group is about Twitter and the second one is about eBay. The third group is talking about Lockerz, a website which was launched in early 2010. These three categories do not currently exist under the "Internet" domain.

Although some questions are also noisy or misclassified, most of the top 10 questions are assigned to the right categories, which is sufficient to help us recognize the underlying semantics of the categories.

These three categories are taken as *target categories* in our experiments. In principle, we are able to identify any categories with semi-supervised methods, only if we can obtain the appropriate prior knowledge. However, in order for performance comparison, we also identify the three target categories when running prior-PLSA. Since $\mu_j$ represents the size of a pseudo question, we heuristically set $\mu_j$ to the average size of the questions in $Q$, which is 179 in our dataset.

**Table 1. Sample results of PLSA.**

| Questions |
|---|
| Funniest usernames!!!!!? |
| How do I send a "tweet" directly to someone on Twitter? |
| how do i tweet jls on twitter?? |
| If I have a protected Twitter acct, can celebrities I follow see what I tweet? |
| On Twitter if you mention someone (@someone), will they see your me…? |
| How Can I send a direct message on twitter? |
| If you use Twitter, I need your help? |
| what, in twitter, does "&apos" mean? |
| What is a subliminal tweet? |
| what are some username ideas? |
| Milhouse? In my meme? |
| i just sold something on ebay and now im confused on what to do next? |
| If your item doesn't sell on ebay, do you get the listing fee back? |
| Can you ask an eBay seller to hold an item for you? |
| eBay seller to cancel transaction for defective item, what do i do? |
| If I sell on eBay, will they take the shipping cost from the amount the…? |
| on ebay auctions, when bidding has finished if i find the bid too low do i …? |
| Ebay: am i likely to get an unpaid item strike? |
| ebay payment problem? |
| shipping costs confusion on ebay? |
| swagbucks referrals...? |
| When Lockerz re-stocks how long until they run out? |
| Waffles.fm invite for a what.cd invite? |
| How many prizes can I redeem in Lockerz general redemption? |
| dose any one know sites like swagbucks? |
| When is the Lockerz March 2010 Redemption date?!? |
| Why can't I redeem my PTZ on Lockerz? |
| When does Lockerz get more prizes in? |
| question about lockerz? |
| Lockerz.com -How to redeem prizes? |

## 4.3 Performance Evaluation

In order to quantitatively evaluate the methods, we ask volunteers (graduate students from our department) to annotate the results of the three target categories. In particular, annotators judge and then mark each community question as "relevant" or

"irrelevant" to the category it belongs to. To ensure these standard outputs to be precise and consistent, we give rigorous annotation guidelines to the annotators. In cases when the questions cannot be determined directly, e.g., "How long will it take?", the annotators are required to browse these questions' pages for final judgment. After obtaining the human annotated results, we evaluate the methods with two metrics, i.e., *Hit Number* and *Weighted Precision*.

Hit Number aims to measure how many relevant questions are absorbed into each target category, which gives basic perspective into the inner structure of each category. Weighted Precision is the precision of each category when we consider the weight of each question. Formally, if we denote a target category as $C_i$ and the relevant questions in it as $\tilde{C}_i$, then Weighted Precision is defined as

$$WP_i = \sum_{q' \in \tilde{C}_i} \pi_{q',i} \Big/ \sum_{q \in C_i} \pi_{q,i} \ . \qquad (11)$$

The intuition behind Weighted Purity is that the important questions contribute more to the *purity* of the category, whereas the performance will be given penalty if irrelevant questions are given high importance.

We compare the two methods on the Twitter, eBay and Lockerz target categories in Table 2. Note that the performance of topic modeling depends on initialization of parameters. For fairness, the two methods share the identical set of initial parameter values. The prior knowledge gives predefined descriptions of the target categories, with which prior-PLSA can adjust the estimation process to make the parameter values close to the prior distributions. This eventually enables prior-PLSA to absorb more relevant questions into each category, and the Hit Number of prior-PLSA is therefore generally larger than that of the basic PLSA.

Also, prior-PLSA consistently outperforms PLSA on Weighted Precision. The introduced prior knowledge acts to "shape" the basic structures of the formed target categories. Due to the restriction and shaping of prior knowledge, the absorbed questions mostly conform to the overall features of the target categories. Therefore, prior-PLSA results in more high-quality categories and performs more effectively for the New Category Identification problem.

**Table 2. Performance evaluation of various methods.**

| Methods | Categories | Hit Number | Weighted Precision |
|---|---|---|---|
| unsupervised | Twitter | 130 | 0.7687 |
| | eBay | 234 | 0.7745 |
| | Lockerz | 175 | 0.7388 |
| semi-supervised | Twitter | 136 | 0.7878 |
| | eBay | 265 | 0.8033 |
| | Lockerz | 184 | 0.7819 |

# 5. CONCLUSION AND FUTURE WORK

In this paper, we study the novel problem of New Category Identification. We give the formal description of this problem, and propose both unsupervised and semi-supervised topic modeling methods to solve it. The results show that our methods perform effectively in finding and interpreting potential categories in CQA. For future work, we will try to test our methods on other domains in Yahoo! Answers. Also, we consider extending this problem to other social media websites such as blogs and online forums.

# 6. ACKNOWLEDGEMENTS

This work was supported by National Natural Science Funding of China under Grant No. 90718022 and National 863 Plans Project under Grant No. 2009AA01Z410. Also, Jie Tang was supported by National High-tech R&D Program (No. 2009AA01Z138).

# 7. REFERENCES

[1] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of SIGIR'99*, pages 50-57, 1999.

[2] J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A Framework to Predict the Quality of Answers with Non-Textual Features. In *Proceedings of SIGIR'06*, pages 228-235, 2006.

[3] P. Jurczyk, and E. Agichtein. Discovering Authorities in Question Answer Communities by Using Link Analysis. In *Proceedings of CIKM'07*, pages 919-922, 2007.

[4] Y. Liu, J. Bian, and E. Agichtein. Predicting Information Seeker Satisfaction in Community Question Answering. In *Proceedings of SIGIR'08*, pages 483-490, 2008.

[5] Y. Liu, N. Narasimhan, V. Vasudevan, and E. Agichtein. Is This Urgent? Exploring Time-Sensitive Information Needs in Collaborative Question Answering. In *Proceedings of SIGIR'09*, pages 712-713, 2009.

[6] Y. Lu, and C. Zhai. Opinion Integration Through Semi-supervised Topic Modeling. In *Proceedings of WWW'08*, pages 121-130, 2008.

[7] Y. Lu, C. Zhai, and N. Sundaresan. Rated Aspect Summarization of Short Comments. In *Proceedings of WWW'09*, pages 131-140, 2009.

[8] M. Qu, G. Qiu, X. He, C. Zhang, H. Wu, J. Bu, and C. Chen. Probabilistic Question Recommendation for Question Answering Communities. In *Proceedings of WWW'09*, pages 1229–1230, 2009.

[9] K. Wang, Z. Ming, and T. S. Chua. A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services. In *Proceedings of SIGIR'09*, pages 187-194, 2009.

[10] X. J. Wang, X. Tu, D. Feng, and L. Zhang. Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning. In *Proceedings of SIGIR'09*, pages 179-186, 2009.

[11] X. Xue, J. Jeon, and W. B. Croft. Retrieval Models for Question and Answer Archives. In *Proceedings of SIGIR'08*, pages 475-482, 2008.

[12] C. Zhai, A. Velivelli, and B. Yu. A Cross-Collection Mixture Model for Comparative Text Mining. In *Proceedings of SIGKDD'04*, pages 743-748, 2004