Query Classification Based on Regularized Correlated Topic Model

Haijun Zhai*, Jiafeng Guo†, Qiong Wu†, Xueqi Cheng†, Huawei Sheng †, Jin Zhang†
*Department of Computer Science and Technology, University of Science & Technology of China, Hefei, China 230027
†Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China 100080
{zhaihaijun, guojiafeng, wuqiong, shenghuawei, zhangjin}@software.ict.ac.cn cxq@ict.ac.cn

Abstract

This paper addresses the problem of query classification (QC), which aims to classify Web search queries into one or more predefined categories. The state-of-the-art solution for QC is to employ a bridging classifier via an intermediate taxonomy. In this paper, we advanced the bridging method by leveraging probabilistic topic models. The topic model, referred as RCTM (Regularized Correlated Topic Model), is an extension of the conventional CTM (Correlated Topic Model). RCTM learns a topic model by leveraging weak supervision from existing annotated data rather than in an unsupervised fashion, and thus it can effectively address the problem in topic modeling while the topics are predefined. The experimental evaluations show that our QC approach outperforms other baseline methods.

1. Introduction

This paper addresses the problem of QC, which involves assigning a Web search query to one or more predefined categories. QC can be conducted based on different taxonomies, e.g., Broder's informational / navigational / transactional taxonomy [4], topical taxonomy [1], or geographical taxonomy [6]. In this paper, we focus on classifying queries into categories based on their topics, which has been highlighted by KDDCUP 2005. QC is potentially useful in many applications in web search, online advertisement and personalization. However, since queries are usually short (e.g., 2-3 words) and ambiguous, how to classify these queries has become a major research issue.

There is previous work focused on topical categorization of user queries. Beitzel et al. [1] proposed to combine a small seed manual classification with techniques from machine learning and computational linguistics for QC. However, it is very difficult and time-consuming to obtain enough training data for their approach. Shen et al. [7] conducted QC by mapping queries to target categories via an intermediate taxonomy. However, this solution is not flexible enough, as it needs to train new classifiers for each new set of target categories. Broder et al. [5] attempted QC by classifying the Web search results retrieved by queries, which also suffered from the retraining problem. A bridging classifier is further proposed in [8] which addressed the retraining problem. The basic idea is to judge the similarity between queries and target categories by their relationship to the intermediate taxonomy. However, such relationship is captured with a simple language model which is limited in its effectiveness.

Inspired by the work of [8], we advanced the bridging method by leveraging probabilistic topic models. Specifically, a classifier is built bridging on an intermediate taxonomy, in which a probabilistic topic model is proposed to capture the relationship from queries and target categories to the intermediate taxonomy.

The special challenge for the topic model here is that hidden topics are predefined by the intermediate taxonomy. Therefore, we proposed to use a novel topic model, referred as RCTM (Regularized Correlated Topic Model), to address this problem. RCTM is an extension of the conventional CTM (Correlated Topic Model) [2]. CTM is proposed based on the earlier topic model LDA (Latent Dirichlet Allocation) [3] to capture the correlation between different topics, which gives a better fit of corpus than LDA[2]. In contrast to CTM which is based on unsupervised learning, RCTM learns a topic model by leveraging weak supervision from existing annotated data.

The experimental evaluations are conducted based on KDDCUP 2005 data set. The results show that our QC method outperforms other baseline methods.

2. QC Framework

In this section, we describe our QC framework based on RCTM. The task of QC is to classify Web search queries into one or more predefined topical categories. The queries are collected from real search engines submitted by Web users. The meanings and intensions of the queries are subjective. The target categories can be defined based on different taxonomies according to real applications. The meanings/semantics of the target categories are defined by the labels in the taxonomy.

Since queries and target categories usually contain only a few words in the QC problem, how to enrich them becomes critical for the QC task [8]. In this paper, we also enrich queries and target categories through search engines. Specifically, queries and target categories are submitted to search engines. Then we can enrich queries and target categories by their returned search results. We extract the

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.91

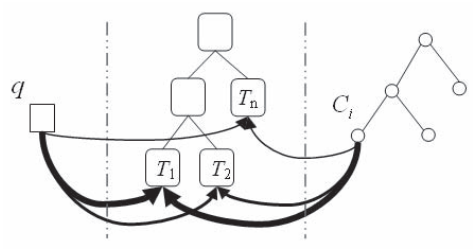

Figure 1. Graph of Classifiers based on an Intermediate Taxonomy

top N titles and snippets from the returned Web pages and assemble them together as the description documents for queries and target categories. It has been proved by most research that the snippets and titles are good descriptions for queries and target categories [8].

2.1. Classifier Based on Topic Models

In our QC framework a classifier is built bridging on an intermediate taxonomy, in which a probabilistic topic model is proposed to capture the relationship from queries and target categories to the intermediate taxonomy. The bridging idea can be illustrated as Fig. 1. The square in the left denotes the query to be classified. The tree in the right represents the target taxonomy. The rounded rectangles in the middle denote the intermediate taxonomy. The lines from a specific query/target category to the intermediate taxonomy denote the proportion of query/target category over the intermediate categories. For example, we can see that the proportion on T_1 is much bigger than that on T_n for the query q. Given a category C_i in the target taxonomy and a query to be classified q, we can judge the similarity between them by the distributions of their relationship to the intermediate taxonomy. The above idea can be explained under a probabilistic framework which has the following form

$$\begin{aligned} p(C_i|q) &= \sum_{T_i} p(C_i, T_i|q) \\ &= \sum_{T_i} p(C_i|T_i, q)p(T_i|q) \\ &\approx \sum_{T_i} p(C_i|T_i)p(T_i|q) \\ &= \sum_{T_i} \frac{p(T_i|C_i)p(C_i)}{p(T_i)} p(T_i|q) \\ &= p(C_i) \sum_{T_i} \frac{p(T_i|C_i)p(T_i|q)}{p(T_i)} \end{aligned} \quad (1)$$

Where $p(T_i)$ and $p(C_i)$ denote the prior distribution of intermediate category T_i and target category C_i. $p(T_i|C_i)$ and $p(T_i|q)$ reflect the relationship from queries and target categories to the intermediate taxonomy respectively. Therefore, we can have the following interpretation about (1). Given q and C_i, under the constraints of $\sum_{T_i} p(T_i|q) = 1$ and $\sum_{T_i} p(T_i|C_i) = 1$, it's easy to prove that $p(C_i|q)$ tends to be larger when q and C_i tend to show similar distribution over the intermediate categories.

The prior distribution of C_i represents the popularity of the target category. With no more knowledge about C_i, $p(C_i)$ is assumed as even distribution in this paper. Furthermore, by viewing intermediate categories as topics, and query/target category as document, it is naturally to estimate the probabilities $p(T_i)$, $p(T_i|C_i)$ and $p(T_i|q)$ under a topic model. Specifically, we can collect a large collection of data to learn a topic model based on the predefined topics (intermediate categories), and obtain the probability of $p(T_i)$ in this process. For the given query/ target category, which is enriched as a document, the learned topic model can then be applied to predict the probabilities $p(T_i|C_i)$ and $p(T_i|q)$. The special challenge here is that the topics are predefined (intermediate categories). Therefore, we will introduce a new regularized topic model to address this problem in the next section.

Now a query q can be classified according to (2):

$$C^* = \arg\max_{C_i} p(C_i|q) \quad (2)$$

2.2. Regularized Topic Model

In this section, we will introduce RCTM used in our QC framework, which is an extension of the conventional CTM. As we mentioned before, the special challenge in topic modeling is that hidden topics are predefined by the intermediate taxonomy. However, the conventional CTM is a latent-topic model, which does not guarantee the alignment between the result topics and the predefined categories. Each document in the intermediate taxonomy is associated with annotations (i.e., categories). This knowledge is a good guidance for learning. Therefore, RCTM is proposed to learn a topic model by leveraging weak supervision from existing annotated data rather than in an unsupervised fashion, and thus it can effectively address the problem in topic modeling while the topics are predefined.

2.2.1. CTM. We start by reviewing the conventional CTM. CTM builds on the earlier topic model LDA. In contrast to LDA, CTM is able to capture the correlation between different topics, which gives a better fit of corpus than LDA[2].

Suppose there is a collection of M documents $D = \{w_1, w_2, ..., w_M\}$ sharing the same set of K topics, and each document is a sequence N of words denoted by $w = \{w_1, w_2, ..., w_n\}$. CTM assumed the following generative process of each document w in a corpus D:

1) Draw $\eta \sim N_K(\mu, \Sigma)$
2) For each of the N words w_n

   a) Draw topic assignment $z_n \sim Multinomial(f(\eta))$ $p(z|\eta) = \exp\{\eta^T z - \log(\sum_{i=1}^{K} e^{\eta_i})\}$

   b) Draw word $w_n \sim Multinomial(\beta_{z_n})$, β is a $K \times V$ matrix

Given the parameters μ, Σ and β, the generative probability of the document is

$$p(w|\mu,\Sigma,\beta)=\int p(\eta|\mu,\Sigma)(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\eta)p(w_n|z_n,\beta))\mathrm{d}\eta$$

Finally, taking the product of the probabilities of documents, we obtain the generative probability of corpus:

$$p(D|\mu,\Sigma,\beta)=\prod_{d=1}^{M}\int p(\eta_d|\mu,\Sigma)(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\eta)p(w_{dn}|z_{dn},\beta))\mathrm{d}\eta_d$$

2.2.2. RCTM. Each document in the intermediate taxonomy is associated with annotations (i.e., categories). This knowledge is a good guidance for learning. RCTM tries to align the result topics to the intermediate categories by incorporating the annotations information into the objective function as a regularizer in a weak fashion.

Here we use vector $y = \{y_1, y_2, ..., y_K\}$ to represent the annotations of each document, where K denotes the topic number and y_i takes a value of 1(0) when the i-th topic(category) is(not) assigned to a document. We propose RCTM to model corpus with predefined class labels (categories) by regularizing the conventional CTM with a regularizer over the topics. Formally, we define the new objective function as

$$O(w, y) = L(w) + \rho R(y) \tag{3}$$

where ρ is a regularization parameter, $L(w) = \log p(w|\mu, \Sigma, \beta)$ is the log likelihood of the document, and $R(y) = \sum_{i=1}^{K} y_i p(c_i|w)$ is the regularizer defined over labels of the document, where $p(c_i|w)$ denotes probability of the i-th topic in document learned by the model. $p(c_i|w) = \frac{1}{N}\sum_{n=1}^{N} z_n^i$ where z_n^i is a value of 1(0) when i-th topic is(not) assigned to n-th word.

Taking sum over all the documents, we obtain the following total objective function:

$$\begin{aligned} O(D,Y) &= L(D) + \rho R(Y) \\ &= \sum_{d=1}^{M} \log p(w_d|\mu,\Sigma,\beta) + \rho \sum_{d=1}^{M}\sum_{i=1}^{K} y_{di} p(c_{di}|w_d) \end{aligned} \tag{4}$$

In (4), we give the objective function of RCTM by regularizing CTM in an explicit mathematical form. In this objective function, $L(D)$ measures how likely the data is generated from the topic model, and $R(Y)$ constraints the topic distribution of each document to concentrate on the existing annotated topics. The regularization parameter ρ indicates how much we want to follow the annotation of the documents. When ρ equals 0, RCTM degenerates into CTM.

2.2.3. Inference and Estimate. The parameter estimation under RCTM is to maximize the objective function, which is similar to that under CTM. we use variational expectation-maximization (EM) to estimate the parameters.
E-step:

$$\widehat{\phi}_{n,i} \propto \exp\{\lambda_i + \frac{\rho}{N} y_i\}\beta_{i,w_n} \tag{5}$$

$$\widehat{\zeta} = \sum_{i}^{K} \exp(\lambda_i + \nu_i^2) \tag{6}$$

$$\mathrm{d}L/\mathrm{d}\lambda = -\Sigma^{-1}(\lambda - \mu) + \sum_{n=1}^{N} \phi_{n1:K} - (N/\zeta)\exp(\lambda + \nu^2/2) \tag{7}$$

$$\mathrm{d}L/\mathrm{d}\nu_i^2 = -\Sigma_{ii}^{-1}/2 - N/2\zeta \exp(\lambda_i + \nu_i^2/2) + 1/2\nu_i^2 \tag{8}$$

M-step:

$$\widehat{\beta}_i \propto \sum_{d} \phi_{d,i} n_d \tag{9}$$

$$\widehat{\mu} = \frac{1}{D}\sum_{d} \lambda_d \tag{10}$$

$$\widehat{\Sigma} = \frac{1}{D}\sum_{d} I\nu_d^2 + (\lambda_d - \widehat{\mu})(\lambda_d - \widehat{\mu})^T \tag{11}$$

The learned RCTM can be applied to predict the topic distribution for new documents. The prediction procedure is the same as the conventional CTM [2].

3. Experiment

In this section, we provide empirical evidence to demonstrate the effectiveness of our QC method based on RCTM through comparison with other baseline methods. In the experiments the regularization parameter ρ was set to 1 by default.

3.1. Data Sets and the Performance Measure

The Open Directory Project (ODP, http://dmoz.org/) data set is used as our intermediate taxonomy. The ODP data set consists of approximately 4,590,000 Web pages from 590,000 categories. In this paper we use all the leaf categories since they can be viewed as the finest categories. At most one hundred Web pages are randomly selected for each category.

We conduct the QC performance evaluation using the data set from KDDCUP 2005 competition. In this data set, the target taxonomy consists of 67 categories. 800,000

queries are provided in this data set, in which 800 queries are randomly chosen for evaluating. The 800 queries are annotated by three human query-labelers L1, L2 and L3, respectively. In this paper the 800 queries are used for the QC performance evaluation.

To evaluate the QC approaches, we use the standard measures to evaluate the performance of query classification, that is precision, recall and F1- measure.

$$Precision = \frac{\sum_i \#\text{of queries are correctly tagged as } C_i}{\sum_i \#\text{of queries are tagged as } C_i}$$

$$Recall = \frac{\sum_i \#\text{of queries are correctly tagged as } C_i}{\sum_i \#\text{of queries labeled by human experts as } C_i}$$

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

3.2. Results and Analysis

As described in Section 3, the top N titles and snippets from the returned Web pages are assembled together as the description documents for queries and target categories. Here we conduct experiments to verify the effect of the parameter N. We chose N from 20 to 100, with the step size of 20. Table 1 shows the experimental results for our method. From Table 1, we can see that when N increases, the performance of our method gets better. And it achieves the best when N takes value of 80. But when N further increases, the performance will drop as some noisy results will be included by search engines.

Moreover, we compare our method based on RCTM against the base classifiers of the KDDCUP 2005 winning solution (SVM-E and Exact-E) [7] and the state-of-the-art method (Bridging classifier) [8] in Table 2. Table 2 shows that the best precision and F1 of our method is higher than the best reported results of the base classifiers of the KDDCUP 2005 winning solution and the state-of-the-art method. The relative improvement for the base classifiers of the KDDCUP 2005 winning solution is more than 16.66% and 9.62% in terms of precision and F1 respectively. And the relative improvement for the bridging method is more than 3.35% and 4.66% in terms of precision and F1 respectively. The experiments show that our QC method based on RCTM outperforms other baseline methods on the KDDCUP 2005 data set.

Table 1. Performance of Our Method with Different Number of Web Search Results

N	Precision	Recall	F_1
20	0.4046	0.4111	0.3926
40	0.4433	0.4449	0.4273
60	0.4497	0.4484	0.4329
80	**0.4620**	**0.4601**	**0.4438**
100	0.4608	0.4587	0.4409

Table 2. Performance of Our Method, Bridging classifier, SVM-E and Exact-E

	Precision	F_1
Our Method	**0.462**	**0.444**
Bridging classifier	0.447	0.424
SVM-E	0.383	0.335
Exact-E	0.396	0.405

4. Conclusion

In this paper, we address the problem of QC, which aims to classify Web search queries into one or more predefined categories. We built a novel bridging classifier, in which a probabilistic topic model is proposed to capture the relationship from queries and target categories to the intermediate taxonomy. Experiments on the KDDCUP 2005 data set demonstrated the effectiveness of our method. The best precision and F1 of our method is higher than those of the base classifiers of the KDDCUP 2005 winning solution and the state-of-the-art method. The relative improvement for the base classifiers of the KDDCUP 2005 winning solution is more than 16.66% and 9.62% in terms of precision and F1 respectively and the relative improvement for the bridging method is more than 3.35% and 4.66% in terms of precision and F1 respectively.

References

[1] S. M. Beitzel, E. C. Jensen, O. Frieder, D. Grossman, D. D. Lewis, A. Chowdhury, and A. Kolcz. Automatic web query classification using labeled and unlabeled training data. In *SIGIR'05*, pages 581–582, 2005.

[2] D. M. Blei and J. D. Lafferty. Correlated topic models. In *ICML'06*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *SIGIR'07*, pages 231–238, 2007.

[6] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *CIKM '03*, pages 325–333, 2003.

[7] D. Shen, R. Pan, J. tao Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Q2c@ust: our winning solution to query classification in kddcup 2005. *SIGKDD Explor. Newsl*, 7:100–110, 2005.

[8] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Building bridges for web query classification. In *SIGIR'06*, pages 131–138, 2006.