Group Non-negative Matrix Factorization with Natural Categories for Question Retrieval in Community Question Answer Archives

Guangyou Zhou, Yubo Chen, Daojian Zeng, and Jun Zhao
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{gyzhou,yubo.chen,djzeng,jzhao}@nlpr.ia.ac.cn

Abstract

Community question answering (CQA) has become an important service due to the popularity of CQA archives on the web. A distinctive feature is that CQA services usually organize questions into a hierarchy of natural categories. In this paper, we focus on the problem of question retrieval and propose a novel approach, called group non-negative matrix factorization with natural categories (GNMFNC). This is achieved by learning the category-specific topics for each category as well as shared topics across all categories via a group non-negative matrix factorization framework. We derive an efficient algorithm for learning the factorization, analyze its complexity, and provide proof of convergence. Experiments are carried out on a real world CQA data set from Yahoo! Answers. The results show that our proposed approach significantly outperforms various baseline methods and achieves the state-of-the-art performance for question retrieval.

1 Introduction

Community question answering (CQA) such as Yahoo! Answers[1] and Quora[2], has become an important service due to the popularity of CQA archives on the web. To make use of the large-scale questions and their answers, it is critical to have functionality of helping users to retrieve previous answers (Duan et al., 2008). Typically, such functionality is achieved by first retrieving the historical questions that best match a user's queried question, and then using answers of these returned questions to answer the queried question. This is what we called *question retrieval* in this paper.

The major challenge for question retrieval, as for most information retrieval tasks, is the *lexical gap* between the queried questions and the historical questions in the archives. For example, if a queried question contains the word "company" but a relevant historical question instead contains the word "firm", then there is a mismatch and the historical question may not be easily distinguished from an irrelevant one. To solve the *lexical gap* problem, most researchers focused on translation-based approaches since the relationships between words (or phrases) can be explicitly modeled through word-to-word (or phrases) translation probabilities (Jeon et al., 2005; Riezler et al., 2007; Xue et al., 2008; Lee et al., 2008; Bernhard and Gurevych, 2009; Zhou et al., 2011; Singh, 2012). However, these existing methods model the relevance ranking without considering the category-specific and shared topics with natural categories, it is not clear whether this information is useful for question retrieval.

A distinctive feature of question-answer pairs in CQA is that CQA services usually organize questions into a hierarchy of natural categories. For example, Yahoo! Answers contains a hierarchy of 26 categories at the first level and more than 1262 subcategories at the leaf level. When a user asks a question, the user is typically required to choose a category label for the question from a predefined hierarchy. Questions in the predefined hierarchy usually share certain generic topics while questions in different categories have their specific topics. For example, questions in categories "Arts & Humanities" and "Beauty & Style" may share the generic topic of "dance" but they also have the category-specific topics of "poem" and "wearing", respectively.

[1] http://answers.yahoo.com/

[2] http://www.quora.com/

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 89–98, Dublin, Ireland, August 23-29 2014.

Inspired by the above observation, we propose a novel approach, called group non-negative matrix factorization with natural categories (GNMFNC). GNMFNC assumes that there exists a set of category-specific topics for each of the category, and there also exists a set of shared topics for all of the categories. Each question in CQA is specified by its category label, category-specific topics, as well as shared topics. In this way, the large-scale question retrieval problem can be decomposed into small-scale subproblems.

In GNMFNC, questions in each category are represented as a term-question matrix. The term-question matrix is then approximated as the product of two matrices: one matrix represents the category-specific topics as well as the shared topics, and the other matrix denotes the question representation based on topics. An objective function is defined to measure the goodness of prediction of the data with the model. Optimization of the objective function leads to the automatic discovery of topics as well as the topic representation of questions. Finally, we calculate the relevance ranking between the queried questions and the historical questions in the latent topic space.

Past studies by (Cao et al., 2009; Cao et al., 2010; Ming et al., 2010; Cai et al., 2011; Ji et al., 2012; Zhou et al., 2013) confirmed a significant retrieval improvement by adding the natural categories into various existing retrieval models. However, all these previous work regarded natural categories individually without considering the relationships among them. On the contrary, this paper can effectively capture the relationships between the shared aspects and the category-specific individual aspects with natural categories via a group non-negative matrix factorization framework. Also, our work models the relevance ranking in the latent topic space rather than using the existing retrieval models. To date, no attempts have been made regarding group non-negative matrix factorization in studies of question retrieval, which remains an under-explored area.

The remainder of this paper is organized as follows. Section 2 describes our proposed group non-negative matrix factorization with natural categories for question retrieval. Section 3 presents the experimental results. In Section 4, we conclude with ideas for future research.

2 Group Non-negative Matrix Factorization with Natural Categories

2.1 Problem Formulation

In CQA, all questions are usually organized into a hierarchy of categories. When a user asks a question, the user is typically required to choose a category label for the question from a predefined hierarchy of categories. Hence, each question in CQA has a category label. Suppose that we are given a question collection $\mathcal{D}$ in CQA archive with size N, containing terms from a vocabulary $\mathcal{V}$ with size M. A question d is represented as a vector $\mathbf{d} \in \mathbb{R}^M$ where each entry denotes the weight of the corresponding term, for example tf-idf is used in this paper. Let $C = \{c_1, c_2, \cdots, c_P\}$ denote the set of categories (subcategories) of question collection $\mathcal{D}$, where P is the number of categories (subcategories). The question collection $\mathcal{D}$ is organized into P groups according to their category labels and can be represented as $\mathcal{D} = \{\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_P\}$. $\mathbf{D}_p = \{\mathbf{d}_1^{(p)}, \cdots, \mathbf{d}_{N_p}^{(p)}\} \in \mathbb{R}^{M \times N_p}$ is the term-question matrix corresponding to category c_p, in which each row stands for a term and each column stands for a question. N_p is the number of questions in category c_p such that $\sum_{p=1}^{P} N_p = N$.

Let $\mathbf{U}_p' = [\mathbf{U}_s, \mathbf{U}_p] \in \mathbb{R}^{M \times (K_s + K_p)}$ be the term-topic matrix corresponding to category c_p, where K_s is the number of shared topics, K_p is the number of category-specific topics corresponding to category c_p, and $p \in [1, P]$. Term-topic matrix $\mathbf{U}_s$ can be represented as $\mathbf{U}_s = [\mathbf{u}_1^{(s)}, \cdots, \mathbf{u}_{K_s}^{(s)}] \in \mathbb{R}^{M \times K_s}$, in which each column corresponds to a shared topic. While the term-topic matrix $\mathbf{U}_p$ can be represented as $\mathbf{U}_p = [\mathbf{u}_1^{(p)}, \cdots, \mathbf{u}_{K_p}^{(p)}] \in \mathbb{R}^{M \times K_p}$. The total number of topics in the question collection $\mathcal{D}$ is $K = K_s + PK_p$. Let $\mathbf{V}_p = [\mathbf{v}_1^{(p)}, \cdots, \mathbf{v}_{N_p}^{(p)}] \in \mathbb{R}^{(K_s + K_p) \times N_p}$ be the topic-question matrix corresponding to category c_p, in which each column denotes the question representation in the topic space. We also denote $\mathbf{V}_p^T = [\mathbf{H}_p^T, \mathbf{W}_p^T]$, where $\mathbf{H}_p \in \mathbb{R}^{K_s \times N_p}$ and $\mathbf{W}_p \in \mathbb{R}^{K_p \times N_p}$ correspond to the coefficients of shared topics $\mathbf{U}_s$ and category-specific topics $\mathbf{U}_p$, respectively.

Thus, given a question collection $\mathcal{D} = \{\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_P\}$ together with the category labels $C = \{c_1, c_2, \cdots, c_P\}$, our proposed GNMFNC amounts to modeling the question collection $\mathcal{D}$ with P group

simultaneously, arriving at the following objective function:

$$\mathcal{O} = \sum_{p=1}^{P} \left\{ \lambda_p \left\| \mathbf{D}_p - [\mathbf{U}_s, \mathbf{U}_p]\mathbf{V}_p \right\|_F^2 + R(\mathbf{U}_s, \mathbf{U}_p) \right\} \tag{1}$$

where $\lambda_p \triangleq \|\mathbf{D}_p\|_F^{-2}$. $R(\mathbf{U}_s, \mathbf{U}_p)$ is a regularization term used to penalize the "similarity" between the shared topics and category-specific topics through $\mathbf{U}_s$ and $\mathbf{U}_p$.

In this paper, we aim to ensure that matrix $\mathbf{U}_s$ captures only shared topics and matrix $\mathbf{U}_p$ captures only the category-specific topics. For example, if matrices $\mathbf{U}_s$ and $\mathbf{U}_p$ are mutually orthogonal, we have $\mathbf{U}_s^T\mathbf{U}_p = \mathbf{0}$. To impose this constraint, we attempt to minimize the sum-of-squares of entries of the matrix $\mathbf{U}_s^T\mathbf{U}_p$ (e.g., $\|\mathbf{U}_s^T\mathbf{U}_p\|_F^2$ which uniformly optimizes each entry of $\mathbf{U}_s^T\mathbf{U}_p$). With this choice, the regularization term of $R(\mathbf{U}_s, \mathbf{U}_p)$ is given by

$$R(\mathbf{U}_s, \mathbf{U}_p) = \sum_{p=1}^{P} \alpha_p \left\| \mathbf{U}_s^T\mathbf{U}_p \right\|_F^2 + \sum_{l=1, l\neq p}^{P} \beta_l \left\| \mathbf{U}_p^T\mathbf{U}_l \right\|_F^2 \tag{2}$$

where α_p and β_l are the regularization parameters, $\forall p \in [1, P]$, $\forall l \in [1, P]$.

Learning the objective function in equation (1) involves the following optimization problem:

$$\min_{\mathbf{U}_s, \mathbf{U}_p, \mathbf{V}_p \geq 0} \mathcal{L} = \mathcal{O} + \sigma_1 \left\| \mathbf{U}_s^T\mathbf{1}_M - \mathbf{1}_{K_s} \right\|_F^2 + \sigma_2 \left\| \mathbf{U}_p^T\mathbf{1}_M - \mathbf{1}_{K_p} \right\|_F^2 + \sigma_3 \left\| \mathbf{V}_p\mathbf{1}_{N_p} - \mathbf{1}_{K_s+K_p} \right\|_F^2 \tag{3}$$

where σ_1, σ_2 and σ_3 are the shrinkage regularization parameters. Based on the shrinkage methodology, we can approximately satisfy the normalization constraints for each column of $[\mathbf{U}_s, \mathbf{U}_p]$ and $\mathbf{V}_p^T$ by guaranteeing the optimization converges to a stationary point.

2.2 Learning Algorithm

We present the solution to the GNMFNC optimization problem in equation (3) as the following theorem. The theoretical aspects of the optimization are presented in the next subsection.

Theorem 2.1. *Updating $\mathbf{U}_s$, $\mathbf{U}_p$ and $\mathbf{V}_p$ using equations (4)∼(6) corresponds to category c_p will monotonically decrease the objective function in equation (3) until convergence.*

$$\mathbf{U}_s \leftarrow \mathbf{U}_s \circ \frac{\left[\sum_{p=1}^{P} \lambda_p \mathbf{D}_p\mathbf{H}_p^T \right]}{\left[\sum_{p=1}^{P} \lambda_p [\mathbf{U}_s, \mathbf{U}_p]\mathbf{V}_p\mathbf{H}_p^T + \alpha_p \mathbf{U}_p\mathbf{U}_p^T\mathbf{U}_s \right]} \tag{4}$$

$$\mathbf{U}_p \leftarrow \mathbf{U}_p \circ \frac{\left[\lambda_p \mathbf{D}_p\mathbf{W}_p^T \right]}{\left[\lambda_p [\mathbf{U}_s, \mathbf{U}_p]\mathbf{V}_p\mathbf{W}_p^T + \alpha_p \mathbf{U}_s\mathbf{U}_s^T\mathbf{U}_p + \sum_{l=1, l\neq p}^{P} \beta_l \mathbf{U}_l\mathbf{U}_l^T\mathbf{U}_p \right]} \tag{5}$$

$$\mathbf{V}_p \leftarrow \mathbf{V}_p \circ \frac{\left[\lambda_p \mathbf{D}_p^T [\mathbf{U}_s, \mathbf{U}_p] \right]}{\left[\lambda_p \mathbf{V}_p^T [\mathbf{U}_s, \mathbf{U}_p]^T [\mathbf{U}_s, \mathbf{U}_p] \right]} \tag{6}$$

where operator $\circ$ is element-wise product and $\frac{[\cdot]}{[\cdot]}$ is element-wise division.

Based on Theorem 2.1, we note that multiplicative update rules given by equations (4)∼(6) are obtained by extending the updates of standard NMF (Lee and Seung, 2001). A number of techniques can be used here to optimize the objective function in equation (3), such as alternating least squares (Kim and Park, 2008), the active set method (Kim and Park, 2008), and the projected gradients approach (Lin, 2007). Nonetheless, the multiplicative updates derived in this paper have reasonably fast convergence behavior as shown empirically in the experiments.

2.3 Theoretical Analysis

In this subsection, we give the theoretical analysis of the optimization, convergence and computational complexity.

Without loss of generality, we only show the optimization of $\mathbf{U}_s$ and formulate the Lagrange function with constraints as follows:

$$\mathcal{L}(\mathbf{U}_s) = \mathcal{O} + \sigma_1 \left\| \mathbf{U}_s^T \mathbf{1}_M - \mathbf{1}_{K_s} \right\|_F^2 + \mathrm{Tr}(\Psi_s \mathbf{U}_s^T) \tag{7}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, $\Psi_s \in \mathbb{R}^{K_s \times K_s}$ is the Lagrange multiplier for the nonnegative constraint $\mathbf{U}_s \geq \mathbf{0}$.

The partial derivative of $\mathcal{L}(\mathbf{U}_s)$ w.r.t. $\mathbf{U}_s$ is

$$\begin{aligned} \bigtriangledown_{\mathbf{U}_s} \mathcal{L}(\mathbf{U}_s) = & -2 \sum_{p=1}^{P} \lambda_p \mathbf{D}_p \mathbf{H}_p^T + 2 \sum_{p=1}^{P} \lambda_p [\mathbf{U}_s, \mathbf{U}_p] \mathbf{V}_p \mathbf{H}_p^T \\ & + 2 \sum_{p=1}^{P} \alpha_p \mathbf{U}_p \mathbf{U}_p^T \mathbf{U}_s + 2\sigma_1 \mathbf{U}_s - 2\sigma_1 + \Psi_s \end{aligned} \tag{8}$$

Using the Karush-Kuhn-Tucker (KKT) (Boyd and Vandenberghe, 2004) condition $\Psi_s \circ \mathbf{U}_s = \mathbf{0}$, we obtain

$$\bigtriangledown_{\mathbf{U}_s} \mathcal{L}(\mathbf{U}_s) \circ \mathbf{U}_s = \left\{ \begin{array}{l} -\sum_{p=1}^{P} \lambda_p \mathbf{D}_p \mathbf{H}_p^T + \sum_{p=1}^{P} \lambda_p [\mathbf{U}_s, \mathbf{U}_p] \mathbf{V}_p \mathbf{H}_p^T \\ +\sum_{p=1}^{P} \alpha_p \mathbf{U}_p \mathbf{U}_p^T \mathbf{U}_s + \sigma_1 \mathbf{U}_s - \sigma_1 \end{array} \right\} \circ \mathbf{U}_s = \mathbf{0} \tag{9}$$

After normalization of $\mathbf{U}_s$, the terms $\sigma_1 \mathbf{U}_s$ and σ_1 are in fact equal. They can be safely ignored from the above formula without influencing convergence. This leads to the updating rule for $\mathbf{U}_s$ in equation (4). Following the similar derivations as shown above, we can obtain the updating rules for the rest variables $\mathbf{U}_p$ and $\mathbf{V}_p$ in GNMFNC optimization, as shown in equations (5) and (6).

2.3.1 Convergence Analysis

In this subsection, we prove the convergence of multiplicative updates given by equations (4)$\sim$(6). We first introduce the definition of auxiliary function as follows.

Definition 2.1. *$\mathcal{F}(\mathbf{X}, \mathbf{X}')$ is an auxiliary function for $\mathcal{L}(\mathbf{X})$ if $\mathcal{L}(\mathbf{X}) \leq \mathcal{F}(\mathbf{X}, \mathbf{X}')$ and equality holds if and only if $\mathcal{L}(\mathbf{X}) = \mathcal{F}(\mathbf{X}, \mathbf{X})$.*

Lemma 2.1. *(Lee and Seung, 2001) If $\mathcal{F}$ is an auxiliary function for $\mathcal{L}$, $\mathcal{L}$ is non-increasing under the update*

$$\mathbf{X}^{(t+1)} = \arg\min_{\mathbf{X}} \mathcal{F}(\mathbf{X}, \mathbf{X}^{(t)})$$

Proof. By Definition 2.1, $\mathcal{L}(\mathbf{X}^{(t+1)}) \leq \mathcal{F}(\mathbf{X}^{(t+1)}, \mathbf{X}^{(t)}) \leq \mathcal{F}(\mathbf{X}^{(t)}, \mathbf{X}^{(t)}) = \mathcal{L}(\mathbf{X}^{(t)})$ □

Theorem 2.2. *Let $\mathcal{L}(\mathbf{U}_s^{(t+1)})$ denote the sum of all terms in $\mathcal{L}$ that contain $\mathbf{U}_s^{(t+1)}$, the following function is an auxiliary function for $\mathcal{L}(\mathbf{U}_s^{(t+1)})$*

$$\mathcal{F}(\mathbf{U}_s^{(t+1)}, \mathbf{U}_s^{(t)}) = \mathcal{L}(\mathbf{U}_s^{(t)}) + (\mathbf{U}_s^{(t+1)} - \mathbf{U}_s^{(t)}) \bigtriangledown_{\mathbf{U}_s^{(t)}} \mathcal{L}(\mathbf{U}_s^{(t)}) + \frac{1}{2} (\mathbf{U}_s^{(t+1)} - \mathbf{U}_s^{(t)})^2 \mathcal{P}(\mathbf{U}_s^{(t)}) \tag{10}$$

$$\mathcal{P}(\mathbf{U}_s^{(t)}) = \frac{\sum_{ij} \left[\sum_{p=1}^{P} \lambda_p [\mathbf{U}_s^{(t)}, \mathbf{U}_p] \mathbf{V}_p \mathbf{W}_p^T + \alpha_p \mathbf{U}_p \mathbf{U}_p^T \mathbf{U}_s^{(t)} + \sigma_1 \mathbf{U}_s^{(t)} \right]_{ij}}{\sum_{ij} [\mathbf{U}_s^{(t)}]_{ij}}$$

where $\bigtriangledown_{\mathbf{U}_s^{(t)}} \mathcal{L}(\mathbf{U}_s^{(t)})$ is the first-order derivative of $\mathcal{L}(\mathbf{U}_s^{(t)})$ with respect to $\mathbf{U}_s^{(t)}$. Theorem 2.2 can be proved similarly to (Lee and Seung, 2001) by validating $\mathcal{L}(\mathbf{U}_s^{(t+1)}) \leq \mathcal{F}(\mathbf{U}_s^{(t+1)}, \mathbf{U}_s^{(t)})$, $\mathcal{L}(\mathbf{U}_s^{(t+1)}) = \mathcal{F}(\mathbf{U}_s^{(t+1)}, \mathbf{U}_s^{(t+1)})$, and the Hessian matrix $\bigtriangledown \bigtriangledown_{\mathbf{U}_s^{(t+1)}} \mathcal{F}(\mathbf{U}_s^{(t+1)}, \mathbf{U}_s^{(t)}) \succeq 0$. Due to limited space, we omit the details of the validation.

	addition	multiplication	division	overall
GNMFNC: $\mathbf{U}_s$	$P(3MN_pK_s + MN_pK_p + MK_s^2)$	$P(3MN_pK_s + MN_pK_p + MK_s^2)$	MK_s	$O(PMN_pK_{max})$
GNMFNC: $\mathbf{U}_p$	$3MN_pK_p + MN_pK_s + PM^2K'$	$3MN_pK_p + MN_pK_s + PM^2K'$	MK_p	$O(PMRK')$
GNMFNC: $\mathbf{V}_p$	$3MN_pK'$	$3MN_pK'$	N_pK'	$O(MN_pK')$

Table 1: Computational operation counts for each iteration in GNMFNC.

Based on Theorem 2.2, we can fix $\mathbf{U}_s^{(t)}$ and minimize $\mathcal{F}(\mathbf{U}_s^{(t+1)}, \mathbf{U}_s^{(t)})$ with respect to $\mathbf{U}_s^{(t+1)}$. When setting $\bigtriangledown_{\mathbf{U}_s^{(t+1)}} \mathcal{F}(\mathbf{U}_s^{(t+1)}, \mathbf{U}_s^{(t)}) = \mathbf{0}$, we get the following updating rule

$$\mathbf{U}_s^{(t+1)} \leftarrow \mathbf{U}_s^{(t)} \circ \frac{\left[\sum_{p=1}^{P} \lambda_p \mathbf{D}_p \mathbf{H}_p^T + \sigma_1\right]}{\left[\sum_{p=1}^{P} \lambda_p [\mathbf{U}_s^{(t)}, \mathbf{U}_p] \mathbf{V}_p \mathbf{W}_p^T + \alpha_p \mathbf{U}_p \mathbf{U}_p^T \mathbf{U}_s^{(t)} + \sigma_1 \mathbf{U}_s^{(t)}\right]} \tag{11}$$

which is consistent with the updating rule derived from the KKT conditions aforementioned.

By Lemma 2.1 and Theorem 2.2, we have $\mathcal{L}(\mathbf{U}_s^{(0)}) = \mathcal{F}(\mathbf{U}_s^{(0)}, \mathbf{U}_s^{(0)}) \geq \mathcal{F}(\mathbf{U}_s^{(1)}, \mathbf{U}_s^{(0)}) \geq \mathcal{F}(\mathbf{U}_s^{(1)}, \mathbf{U}_s^{(1)}) = \mathcal{L}(\mathbf{U}_s^{(1)}) \geq \cdots \geq \mathcal{L}(\mathbf{U}_s^{(Iter)})$, where $Iter$ is the number of iterations. Therefore, $\mathbf{U}_s$ is monotonically decreasing. Since the objective function $\mathcal{L}$ is lower bounded by 0, the correctness and convergence of Theorem 2.1 is validated.

2.3.2 Computational Complexity

In this subsection, we discuss the time computational complexity of the proposed algorithm GNMFNC. Besides expressing the complexity of the algorithm using big O notation, we also count the number of arithmetic operations to provide more details about running time. We show the results in Table 1, where $K_{max} = \max\{K_s, K_p\}$, $K' = K_s + K_p$ and $R = \max\{M, N_p\}$.

Suppose the multiplicative updates stop after $Iter$ iterations, the time cost of multiplicative updates then becomes $O(Iter \times PMRK')$. We set $Iter = 100$ empirically in rest of the paper. Therefore, the overall running time of GNMFNC is linear with respect to the size of word vocabulary, the number of questions and categories.

2.4 Relevance Ranking

The motivation of incorporating matrix factorization into relevance ranking is to learn the word relationships and reduce the "lexical gap" (Zhou et al., 2013a). To do so, given a queried question q with category label c_p from Yahoo! Answers, we first represent it in the latent topic space as $\mathbf{v}_q$,

$$\mathbf{v}_q = \arg\min_{\mathbf{v} \geq 0} \|\mathbf{q} - [\mathbf{U}_s, \mathbf{U}_p]\mathbf{v}\|_2^2 \tag{12}$$

where vector $\mathbf{q}$ is the tf-idf representation of queried question q in the term space.

For each historical question d (indexed by r) in question collection $\mathcal{D}$, with representation $\mathbf{v}_d = r$-th column of $\mathbf{V}$, we compute its similarity with queried question $\mathbf{v}_q$ as following

$$s_{topic}(q, d) = \frac{< \mathbf{v}_q, \mathbf{v}_d >}{\|\mathbf{v}_q\|_2 \cdot \|\mathbf{v}_d\|_2} \tag{13}$$

The latent topic space score $s_{topic}(q, d)$ is combined with the conventional term matching score $s_{term}(q, d)$ for final relevance ranking. There are several ways to conduct the combination. Linear combination is a simple and effective way. The final relevance ranking score $s(q, d)$ is:

$$s(q, d) = \gamma s_{topic}(q, d) + (1 - \gamma) s_{term}(q, d) \tag{14}$$

where $\gamma \in [0, 1]$ is the parameter which controls the relative importance of the latent topic space score and term matching score. $s_{term}(q, d)$ can be calculated with any of the conventional relevance models such as BM25 (Robertson et al., 1994) and LM (Zhai and Lafferty, 2001).

3 Experiments

3.1 Data Set and Evaluation Metrics

We collect the data set from Yahoo! Answers and use the *getByCategory* function provided in Yahoo! Answers API[3] to obtain CQA threads from the Yahoo! site. More specifically, we utilize the *resolved* questions and the resulting question repository that we use for question retrieval contains 2,288,607 questions. Each resolved question consists of four parts: "question title", "question description", "question answers" and "question category". We only use the "question title" and "question category" parts, which have been widely used in the literature for question retrieval (Cao et al., 2009; Cao et al., 2010). There are 26 first-level categories in the predefined natural hierarchy, i.e., each historical question is categorized into one of the 26 categories. The categories include "Arts & Humanities", "Beauty & Style", "Business & Finance", etc.

In order to evaluate our approach, we randomly select 2,000 questions as queried questions from the above data collection to construct the validation/test sets, and the remaining data collection as training set. Note that we select the queried questions in proportion to the number of questions and categories against the whole distribution to have a better control over a possible imbalance. To obtain the ground-truth, we employ the Vector Space Model (VSM) (Salton et al., 1975) to retrieve the top 10 results and obtain manual judgements. The top 10 results don't include the queried question itself. Given a returned result by VSM, an annotator is asked to label it with "relevant" or "irrelevant". If a returned result is considered semantically equivalent to the queried question, the annotator will label it as "relevant"; otherwise, the annotator will label it as "irrelevant". Two annotators are involved in the annotation process. If a conflict happens, a third person will make judgement for the final result. In the process of manually judging questions, the annotators are presented only the questions. As a result, there are in total 20,000 judged question pairs. We randomly split the 2,000 queried questions into validation/test sets, each has 1,000/1,000 queried questions. We use the validation set for parameter tuning and the test set for evaluation.

Evaluation Metrics: We evaluate the performance of question retrieval using the following metrics: Mean Average Precision (MAP) and Precision@N (P@N). MAP rewards methods that return relevant questions early and also rewards correct ranking of the results. P@N reports the fraction of the top-N questions retrieved that are relevant. We perform a significant test, i.e., a t-test with a default significant level of 0.05.

There are several parameters used in the paper, we tune these parameters on the validation set. Specifically, we set the number of category-specific topics per category and the number of shared topics in GNMFNC as $(K_s, K_p) = \{(5,2),(10,4),(20,8),(40,16),(80,32)\}$, resulting in $K = \{57, 114, 228, 456, 912\}$ total number of topics. (Note that the total number of topics in GNMFNC is $K_s + 26 \times K_p$, where 26 is the number of categories in the first-level predefined natural hierarchy[4]). Finally, we set $(K_s, K_p) = (20, 8)$ and $K = 228$ empirically as this setting yields the best performance.

For regularization parameters α_p and β_l, it is difficult to directly tune on the validation set, we present an alternative way by adding a common factor a to look at the objective function of optimization problem in equation (3) on the training data. In other words, we set $\alpha_p = \frac{a}{K_s \times K_p}$ and $\beta_l = \frac{a}{K_p \times K_l}$. Therefore, we tune the parameters α_p and β_l by alternatively adjusting the common factor a via grid search. As a result, we set $a = 100$, resulting in $\alpha_p = \beta_l = 0.625$ in the following experiments. The trade-off parameter γ in the linear combination is set from 0 to 1 in steps of 0.1 for all methods. We set $\gamma = 0.6$ empirically. For shrinkage regularization parameters, we empirically set $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

3.2 Question Retrieval Results

In this experiment, we present the experimental results for question retrieval on the test data set. Specifically, for our proposed GNMFNC, we combine the latent topic matching scores with the term matching scores given by BM25 and LM, denoted as "BM25+GNMFNC" and "LM+GNMFNC". Table 2 shows

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

[4] Here we do not use the leaf categories because we find that it is not possible to run GNMFNC with such large number of topics on the current machines, and we will leave it for future work.

Table 2: Comparison with different methods for question retrieval.

#	Methods	MAP	P@10
1	BM25	0.243	0.225
2	LM	0.286	0.232
3	(Jeon et al., 2005)	0.327	0.235
4	(Xue et al., 2008)	0.341	0.238
5	(Zhou et al., 2011)	0.365	0.243
6	(Singh, 2012)	0.354	0.240
7	(Cao et al., 2010)	0.358	0.242
8	(Cai et al., 2011)	0.331	0.236
9	**BM25+GNMFNC**	**0.369**	**0.248**
10	**LM+GNMFNC**	**0.374**	**0.251**

Table 3: Comparison of matrix factorizations for question retrieval.

#	Methods	MAP	P@10
1	BM25	0.243	0.225
2	BM25+NMF	0.325	0.235
3	BM25+CNMF	0.344	0.239
4	BM25+GNMF	0.361	0.242
5	BM25+GNMFNC	0.369	0.248
6	LM	0.286	0.232
7	LM+NMF	0.337	0.237
8	LM+CNMF	0.352	0.240
9	LM+GNMF	0.365	0.243
10	LM+GNMFNC	0.374	0.251

the main retrieval performances under the evaluation metrics MAP, P@1 and P@10. Row 1 and row 2 are the baseline systems, which model the relevance ranking using BM25 (Robertson et al., 1994) and language model (LM) (Zhai and Lafferty, 2001) in the term space. Row 3 is word-based translation model (Jeon et al., 2005), and row 4 is word-based translation language model (TRLM) (Xue et al., 2008). Row 5 is phrase-based translation model (Zhou et al., 2011), and row 6 is the entity-based translation model (Singh, 2012). Row 7 to row 11 explore the natural categories for question retrieval. In row 7, Cao et al. (2010) employed the natural categories to compute the local and global relevance with different model combination, here we use the combination VSM + TRLM for comparison because this combination obtains the superior performance than others. In row 8, Cai et al. (2011) proposed a category-enhanced TRLM for question retrieval. There are some clear trends in the results of Table 2:

(1) BM25+GNMFNC and LM+GNMFNC perform *significantly* better than BM25 and LM respectively (t-test, p-value < 0.05, row 1 vs. row 9; row 2 vs. row 10), indicating the effective of GNMFNC.

(2) BM25+GNMFNC and LM+GNMFNC perform better than translation methods, some improvements are statistical significant (t-test, p-value < 0.05, row 3 and row 4 vs. row 9 and row 10). The reason may be that GNMFNC models the relevance ranking in the latent topic space, which can also effectively solve the the lexical gap problem.

(3) Capturing the shared aspects and the category-specific individual aspects with natural categories in the group modeling framework can *significantly* improve the performance of question retrieval (t-test, p-value < 0.05, row 7 and row 8 vs. row 9 and row 10).

(4) Natural categories are useful and effectiveness for question retrieval, no matter in the group modeling framework or existing retrieval models (row 3$\sim$ row 6 vs. row 7$\sim$row 10).

3.3 Comparison of Matrix Factorizations

We note that our proposed GNMFNC is related to non-negative matrix factorization (NMF) (Lee and Seung, 2001) and its variants, we introduce three baselines. The first baseline is NMF, which is trained on the whole training data. The second baseline is CNMF, which is trained on each category without considering the shared topics. The third baseline is GNMF (Lee and Choi, 2009; Wang et al., 2012), which is similar to our GNMFNC but there are no constraints on the category-specific topics to prevent them from capturing the information from the shared topics.

NMF and GNMF are trained on the training data with the same parameter settings in section 4.1 for fair comparison. For CNMF, we also train the model on the training data with the same parameter settings in section 4.1, except parameter K_s, as there exists no shared topics in CNMF.

Table 3 shows the question retrieval performance of NMF families on the test set, obtained with the best parameter settings determined by the validation set. From the results, we draw the following observations:

(1) All of these methods can *significantly* improve the performance in comparison to the baseline BM25 and LM (t-test, p-value < 0.05).

(2) GNMF and GNMFNC perform *significantly* better than NMF and CNMF respectively (t-test, p-value < 0.05), indicating the effectiveness of group matrix factorization framework, especially the use of shared topics.

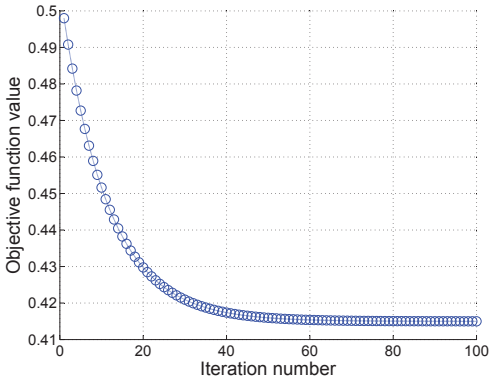

Figure 1: Convergence curve of GNMFNC.

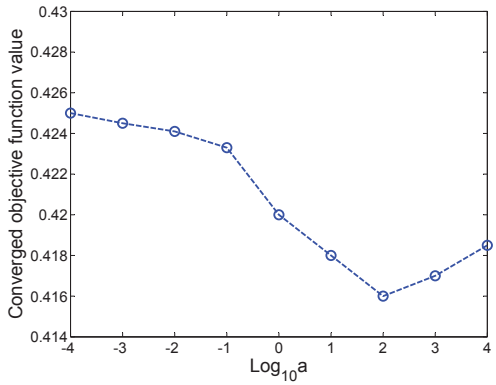

Figure 2: Objective function value vs. factor a.

(3) GNMFNC performs *significantly* better than GNMF (t-test, p-value < 0.05, row 4 vs. row 5; row 9 vs. row 10), indicating the effectiveness of the regularization term on the category-specific topics to prevent them from capturing the information from the shared topics.

From the experimental results reported above, we can conclude that our proposed GNMFNC is useful for question retrieval with high accuracies. To the best of our knowledge, it is the first time to investigate the group matrix factorization for question retrieval.

3.4 Convergence Behavior

In subsection 2.3.1, we have shown that the multiplicative updates given by equations (4)$\sim$(6) are convergent. Here, we empirically show the convergence behavior of GNMFNC.

Figure 1 shows the convergence curve of GNMFNC on the training data set. From the figure, y-axis is the value of objective function and x-axis denotes the iteration number. We can see that the multiplicative updates for GNMFNC converge very fast, usually within 80 iterations.

3.5 Regularization Parameters Selection

One success of this paper is to use regularized constrains on the category-specific topics to prevent them from capturing the information from the shared topics. It is necessary to give an in-depth analysis of the regularization parameters used in the paper. Consider the regularization term used in equation (2), each element in $\mathbf{U}_s^T\mathbf{U}_p$ and $\mathbf{U}_p^T\mathbf{U}_l$ has a value between 0 and 1 as each column of $\mathbf{U}_s$, $\mathbf{U}_p$ and $\mathbf{U}_l$ is normalized. Therefore, it is appropriate to normalize the term having $\|\mathbf{U}_s^T\mathbf{U}_p\|_F^2$ by K_sK_p since there are $K_s \times K_p$ elements in $\mathbf{U}_s^T\mathbf{U}_p$. Similarly, $\|\mathbf{U}_p^T\mathbf{U}_l\|_F^2$ is normalized by K_lK_p. Note that $K_l = K_p$ and $l \neq p$. As discussed in subsection 4.1, we present an alternative way by adding a common factor a and set $\alpha_p = \frac{a}{K_s \times K_p}$ and $\beta_l = \frac{a}{K_p \times K_l}$. The common factor a is used to adjust a trade-off between the matrix factorization errors and the mutual orthogonality, which cannot directly tune on the validation set. Thus, we look at the objective function of optimization problem in equation (3) on the training data and find the optimum value for a.

Figure 2 shows the objective function value vs. common factor a, where y-axis denotes the converged objective function value, and x-axis denotes $\text{Log}_{10}a$. We can see that the optimum value of a is 100. Therefore, the common factor a can be fixed at 100 for our data set used in the paper, resulting in $\alpha_p = \beta_l = 0.625$. Note that the optimum value of (K_s, K_p) are set as $(20, 8)$ in subsection 4.1. Due to limited space, we do not give an in-depth analysis for other parameters.

4 Conclusion and Future Work

In this paper, we propose a novel approach, called group non-negative matrix factorization with natural categories (GNMFNC). The proposed method is achieved by learning the category-specific topics for each category as well as shared topics across all categories via a group non-negative matrix factorization framework. We derive an efficient algorithm for learning the factorization, analyze its complexity, and

provide proof of convergence. Experiments show that our proposed approach significantly outperforms various baseline methods and achieves state-of-the-art performance for question retrieval.

There are some ways in which this research could be continued. First, the optimization of GNMFNC can be decomposed into many sub-optimization problems, a natural avenue for future research is to reduce the running time by executing the optimization in a distributed computing environment (e.g., MapReduce (Dean et al., 2004)). Second, another combination approach will be used to incorporate the latent topic match score as a feature in a learning to rank model, e.g., LambdaRank (Burges et al., 2007). Third, we will try to investigate the use of the proposed approach for other kinds of data sets with larger categories, such as categorized documents from ODP project.[5]

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61333018 and No. 61303180), the Beijing Natural Science Foundation (No. 4144087), CCF Opening Project of Chinese Information Processing, and also Sponsored by CCF-Tencent Open Research Fund. We thank the anonymous reviewers for their insightful comments.

References

D. Bernhard and I. Gurevych. 2009. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of ACL*, pages 728-736.

S. Boyd and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge university press.

C. Boutsidis and E. Gallopoulos. 2008. SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350-1362.

C. Burges, R. Ragno, and Q. Le. 2007. Learning to rank with nonsmooth cost function. In *Proceedings of NIPS*.

L. Cai, G. Zhou, K. Liu, and J. Zhao. 2011. Learning the latent topics for question retrieval in community QA. In *Proceedings of IJCNLP*.

X. Cao, G. Cong, B. Cui, C. Jensen, and C. Zhang. 2009. The use of categorization information in language models for question retrieval. In *Proceedings of CIKM*, pages 265-274.

X. Cao, G. Cong, B. Cui, and C. Jensen. 2010. A generalized framework of exploring category information for question retrieval in community question answer archives. In *Proceedings of WWW*.

J. Dean, S. Ghemanwat, and G. Inc. 2004. Mapreduce: simplified data processing on large clusters. In *Proceedings of OSDI*.

H. Duan, Y. Cao, C. Lin, and Y. Yu. 2008. Searching questions by identifying questions topics and question focus. In *Proceedings of ACL*, pages 156-164.

J. Jeon, W. Croft, and J. Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of CIKM*, pages 84-90.

Z. Ji, F. Xu, and B. Wang. 2012. A category-integrated language model for question retrieval in community question answering. In *Proceedings of AIRS*, pages 14-25.

H. Kim and H. Park. 2008. Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method. *SIAM J Matrix Anal Appl*, 30(2):713-730.

A. Langville, C. Meyer, R. Albright, J. Cox, and D. Duling. 2006. Initializations for the nonnegative matrix factorization. In *Proceedings of KDD*.

J. Lee, S. Kim, Y. Song, and H. Rim. 2008. Bridging lexical gaps between queries and questions on large online Q&A collections with compact translation models. In *Proceedings of EMNLP*, pages 410-418.

D. Lee and H. Seung. 2001. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS*.

[5]http://www.dmoz.org/

H. Lee and S. Choi. 2009. Group nonnegative matrix factorization for eeg classification. In *Proceedings of AISTATS*, pages 320-327.

C. Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural Comput*, 19(10):2756-2779.

Z. Ming, T. Chua, and G. Cong. 2010. Exploring domain-specific term weight in archived question search. In *Proceedings of CIKM*, pages 1605-1608.

S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *Proceedings of ACL*, pages 464-471.

S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at trec-3. In *Proceedings of TREC*, pages 109-126.

G. Salton, A. Wong, and C. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613-620.

A. Singh. 2012. Entity based q&a retrieval. In *Proceedings of EMNLP-CoNLL*, pages 1266-1277.

Q. Wang, Z. Cao, J. Xun, and H. Li. 2012. Group matrix factorizaiton for scalable topic modeling. In *Proceedings of SIGIR*.

X. Xue, J. Jeon, and W. Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of SIGIR*, pages 475-482.

C. Zhai and J. Lafferty. 2001. A study of smooth methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334-342.

G. Zhou, L. Cai, J. Zhao, and K. Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of ACL*, pages 653-662.

G. Zhou, F. Liu, Y. Liu, S. He, and J. Zhao. 2013. Statistical machine translation improves question retrieval in community question answering via matrix factorization. In *Proceedings of ACL*, pages 852-861.

G. Zhou, Y. Chen, D. Zeng, and J. Zhao. 2013. Toward faster and better retrieval models for question search. In *Proceedings of CIKM*, pages 2139-2148.