Affinity Regularized Non-Negative Matrix Factorization for Lifelong Topic Modeling

Yong Chen, Junjie Wu, Jianying Lin, Rui Liu, Hui Zhang, and Zhiwen Ye

Abstract—Lifelong topic model (LTM), an emerging paradigm for never-ending topic learning, aims to yield higher-quality topics as time passes through knowledge accumulated from the past yet learned for the future. In this paper, we propose a novel lifelong topic model based on non-negative matrix factorization (NMF), called Affinity Regularized NMF for LTM (NMF-LTM), which to our best knowledge is distinctive from the popular LDA-based LTMs. NMF-LTM achieves lifelong learning by introducing word-word graph Laplacian as semantic affinity regularization. Other priors such as sparsity, diversity, and between-class affinity are incorporated as well for better performance, and a theoretical guarantee is provided for the algorithmic convergence to a local minimum. Extensive experiments on various public corpora demonstrate the effectiveness of NMF-LTM, particularly its human-like behaviors in two carefully designed learning tasks and the ability in topic modeling of big data. A further exploration of semantic relatedness in knowledge graphs and a case study on a large-scale real-world corpus exhibit the strength of NMF-LTM in discovering high-quality topics in an efficient and robust way.

Index Terms—Lifelong topic model (LTM), non-negative matrix factorization (NMF), semantic affinity, knowledge graph

1 Introduction

Lifelong machine learning (LML) [1], [2], [3] has recently attracted considerable attention from researchers of various domains. In general, there are three key characteristics in LML, i.e., continuous tasks, knowledge accumulation and maintenance, and a human-like knowledge-based learner that can leverage the past knowledge to help future learning in a never-ending manner [4], [5]. LML has been widely adopted to explore topic models [6], [7], entity recognition and information extraction [3], [8], text categorization [9], sentiment analysis [10], and so forth.

Lifelong topic model (LTM), as a typical example of LML, is gaining more and more research interests than traditional one-shot deal that conducts PLSA [11] or LDA [12] on collected documents just for once without the guidance of any prior knowledge. Up till now, there have been several articles on LTM [6], [7], [13], but they all take a probabilistic perspective on this problem and often employ specially designed sampling schemes. For instance, Generalized Polya Urn (GPU) is induced in LDA hierarchical graphical models to leverage prior knowledge via more related word sampling [14], [15]. Since GPU encourages semantic related words to appear in the same topic, it could contribute to the final high-quality topics.

Despite of the success of probabilistic LTMs, they have some disadvantages in essence. For instance, to set an appropriate promotion scale parameter for the GPU-based Gibbs sampler is often difficult in practice, which might result in low-quality prior knowledge. Also, the nonexchangeable property of the GPU model might make the joint probability of words in any given topic be variant to their different orders, and make the model inference computationally expensive. Additionally, GPU-based LDA learning scheme mainly focuses on the most related limited knowledge, which might fail when the task is targeted at topic mining of short texts. These indeed motivates our study in this paper, which suggests using a matrix factorization framework with prior knowledge stored and accumulated as a word-word affinity matrix for lifelong topic modeling. Specifically, our research contributions can be summarized as follows:

- First, we present a novel NMF framework called NMF-LTM for lifelong topic modeling, which to our best knowledge is the first study distinctive from the existing LDA-based ones. The theoretical proof to the algorithmic convergence of NMF-LTM is also provided with solid math. NMF-LTM also exhibits human-like learning behaviors and great potentials for big data modeling.
- Second, a word-word affinity matrix with three types of knowledge representations, i.e., $0 \sim 1$ naive ratio, $\{0, 1\}$ binary values, and smoothed $0 \sim 1$ ratio, is adopted to link the past to the future in a simple yet effective way.
- Third, NFM-LTM also incorporates other priors such as sparsity, diversity and class-class affinities for seeking potential effective factors for preferable lifelong

- *Y. Chen, J. Lin, R. Liu, H. Zhang, and Z. Ye are with the Department of Computer Science and Engineering, Beihang University, Beijing 100191, China. E-mail: {chenyong, hzhang}@nlsde.buaa.edu.cn, {jianying.lin, lr, yezhiwen}@buaa.edu.cn.*
- *J. Wu is with the School of Economics and Management, Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beihang University, Beijing 100191, China. E-mail: wujj@buaa.edu.cn.*

Manuscript received 26 Nov. 2017; revised 10 Feb. 2019; accepted 3 Mar. 2019. Date of publication 12 Mar. 2019; date of current version 3 June 2020. (Corresponding author: Junjie Wu.)
Recommended for acceptance by J. Ye.
Digital Object Identifier no. 10.1109/TKDE.2019.2904687

topic models, which according to the experimental validation is critically important to the success of NMF-LTM.

Extensive experiments are conducted on several public data sets to validate the superiority of NMF-LTM to some state-of-the-art baselines. In particular, we design two human-like learning tasks for NMF-LTM, i.e., learning with different strategies and learning with multi-run reviews, which disclose the human-like behaviors of NMF-LTM in lifelong learning. In addition, we further explore the semantic quality of a knowledge graph, which provides a clue to explain why NMF-LTM could yield high-quality topics. It is also interesting that NMF-LTM exhibits a natural fit to topic modeling of big data in terms of both runtime efficiency and topic coherence on simulated as well as real-world large-scale corpora.

The remainder of this paper is organized as follows. In Section 2, we introduce some work related to lifelong topic models. In Section 3, we define our problem and give necessary notations. In Sections 4 and 5, we propose our method NMF-LTM and its optimization algorithm with a theoretical convergence guarantee. We conduct experiments in Section 6, explore the word-word semantics of the learned knowledge graph in Section 7, and perform a case study on a real-world large-scale scientific data set in Section 8. We finally draw conclusions in Section 9.

2 Related Work

In this section, we briefly review existing lifelong topic models, and differentiate them from some well-known models in different domains including constrained topic models and dynamic topic models.

2.1 Lifelong Topic Modeling

Motivated by the human learning paradigm, lifelong machine learning (LML) [2], [4], [16] is proposed to design computational systems and algorithms that can learn as humans do, e.g., retaining the results learned in the past, abstracting knowledge from them, and using the knowledge to help future learning. It is noteworthy that lifelong learning does not fall into the category of online learning [17], [18]. The latter is mainly concerned with designing resoluble models so that they can be updated incrementally with incoming data streams. In other words, online learning is more efficiency-driven while lifelong learning is more quality-driven, with the help from knowledge accumulation.

As a specific case of LML, a lifelong topic model (LTM) [6], [7], [10] shares the same learning philosophy as LML while has its unique components such as knowledge-based topic models as the central learners and topic-related knowledge mining and accumulation. Currently, the popular LTMs [6], [10] are almost all based on hierarchical graphical models, which utilize LDA [12] as the learner, and leverage GPU for more related word sampling and inferences with the guidance of self-learned knowledge. It is reported by [6] that these LDA-based LTMs, or *LDA-LTM* for short, are superior to many state-of-the-art topic models [19], [20], [21], [22].

However, LDA-LTM with GPU sampling mainly considers the few most related topic words, and therefore suffers a potential loss in knowledge enhancements. Besides, LDA-based learner might perform poorly for lacking of statistical sufficiency when dealing with short texts or with little knowledge supervision. This indeed motivates us to design NMF-based LTMs, or *NMF-LTM* for short, for high-quality lifelong topic modeling. NMF-LTM is different from LDA-LTM in leveraging a parts-based nonnegative matrix factorization [23] learner, regularized by self-learned word-word semantics for topic mining. Moreover, some priors such as data sparsity, smoothness and supervised information can be integrated into NMF-LTM easily for higher-quality topics learning. More comparative performances would be elaborated in the experimental sections.

2.2 Constrained Topic Modeling

In recent years, topic modeling with external knowledge obtained from such as Probase[1] and Wikipedia[2] emerge as knowledge-constrained methods that imitate human learning for high-quality topics. For instance, DF-LDA [19] utilizes user-provided domain priors in the form as must-links and cannot-links to supervise LDA with Dirichlet Forest. GK-LDA [20] leverages a probabilistic value under each topic to reduce the effect of wrong knowledge with just must-link type knowledge. MC-LDA [21] also exploits must-link and cannot-link knowledge but relates each term with only one tightly associated must-link so as to avoid the wrong knowledge. AKL [22] is capable of handling faulted knowledge and can produce more coherent aspects via Generalized Plya Urn (GPU).

The above models mostly assume knowledge is given in advance and takes the form as must-links and/or cannot-links constraints. In lifelong topic modeling, however, the knowledge is accumulated continuously and can take flexible forms, for example as a word-word affinity matrix in our paper, according to specific learning tasks.

2.3 Dynamic Topic Modeling

From the continuous learning perspective, our work is also related to dynamic topic modeling (DTM). In general, DTM aims to discover latent topics from sequentially arrived data blocks so as to capture their evolutionary trends. For instance, Blei et al. [24], [25] proposed the dynamic probabilistic topic models with the assumption that all the sequential organized documents are sharing the same focused topics with smooth variations. They were then scaled to big data by Arnab et al. [26] using a kind of Gibbs Sampling with Stochastic Gradient Langevin Dynamics and a Metropolis-Hastings based $O(1)$ sampler. To further capture the newly emerging and fading topics, Ankan and Vikas [27], [28] proposed a dynamic NMF approach with temporal regularization to learn evolving and emerging topics in social media. Chen et al. [29] put forward a comprehensive solution to model the emerging, evolving and fading topics in a more flexible NMF framework.

The above research indeed highlights the characteristics of DTM: to mine topics without the guidance of knowledge and concentrate on finding the evolutionary patterns of topics. This is in sharp contrast to LTM, which tries to learn

1. http://probase.msra.cn/
2. https://www.wikipedia.org/

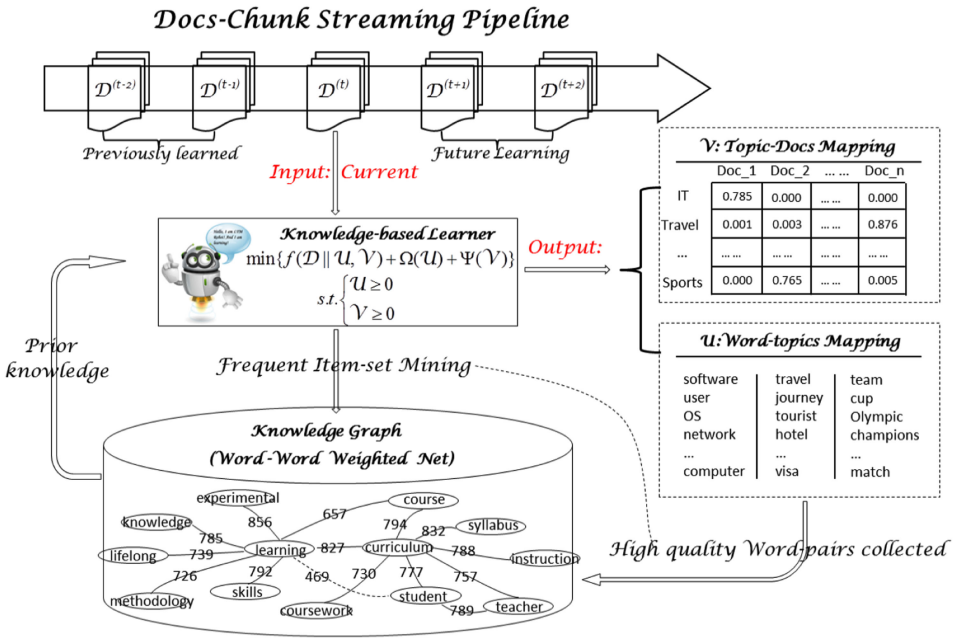


Fig. 1. The framework of NMF-LTM.

higher quality topics through accumulated knowledge in a never-ending mode.

3 Problem Statement

Given a set of sequential document chunks $\{\mathcal{D}^{(t)}\}_{t=1}^{T}$ accumulated in an endless manner ($T = 1, 2, \ldots, +\infty$) within a steady docs-stream, lifelong topic modeling is to seek high-quality topics by means of a knowledge-based learner that could conduct topic models like human with prior accumulated knowledge learned from the past docs-chunks for a much better future learning. Obviously, there are three key components in such learning paradigm: (1) learning continuously without stops, (2) shared knowledge-based topic models, and (3) knowledge maintenance and utilization.

Fig. 1 shows our proposed framework of the NMF-based Lifelong Topic Model (NMF-LTM). In NMF-LTM, the documents are collected in chunks through a steaming pipeline that can provide never-ending data sources to the learning robot. In terms of such knowledge-based learner, we mainly design a non-negative matrix factorization framework with priors as follows:

$$\begin{aligned} &\min\{f(\mathcal{D}||\mathcal{U},\mathcal{V}) + \Omega(\mathcal{U}) + \Psi(\mathcal{V})\} \\ &s.t.\ \mathcal{U} \geq 0,\ \mathcal{V} \geq 0, \end{aligned} \tag{1}$$

where $\mathcal{D}$ represents a document chunk in the term-document form, and $\mathcal{U}$, $\mathcal{V}$ correspond to word-topic and topic-document mappings illustrated in Fig. 1. With respect to the specific forms of loss function $f(\mathcal{D}||\mathcal{U},\mathcal{V})$ and priors $\Omega(\mathcal{U})$ and $\Psi(\mathcal{V})$, more details would be given in the sections to follow.

As far as we know, a NMF-based framework has seldom been touched in solving lifelong topic models—almost all previous works [6], [7], [10], [13] adopt a LDA-based framework for lifelong topic mining. In fact, NMF has been regarded as an effective tool in mining abstract semantics in both images and texts [23], [29]. Its flexibility in transforming various types of prior knowledge into regulations makes it a promising candidate for lifelong topic modeling.

Regarding to the shared knowledge, we choose to accumulate the word-word pairs from the learned word-topic mappings to construct a word-word weighted net (WWWNet), also called knowledge graph ($\mathcal{KG}$ for short). Just as shown in Fig. 1, these principal components form NMF-LTM, a closed cycle for human-like lifelong learning.

4 The Model

In this section, we elaborate our proposed Affinity Regularized NMF for Human-like Lifelong Topic Model (NMF-LTM). First, we would depict the classic non-negative matrix factorization for topic model, and then discuss some potential priors on word-topic and topic-document matrices. Next, the maintenance on the linker ($\mathcal{KG}$) is presented at length. Finally, the overall framework for lifelong topic learning is summarized.

4.1 Topic Mining with NMF

What if we take an algebra perspective on topics rather than the above-mentioned probabilistic view? Undoubtedly, each topic $\mathcal{U}_k^{(t)}$ corresponds to a coordinate axis that is the linear combination of words, and then each document could be expressed by the given K topics, i.e., $\mathcal{D}_n^{(t)} = \sum_{k=1}^{K} \mathcal{V}_{kn}^{(t)} \mathcal{U}_k^{(t)}$, where t represents the index of the current docs-chunk, K denotes the pre-set total number of latent topics, $\mathcal{D}_n^{(t)}$ marks the nth document in the original term-space under the current task, and $\mathcal{V}_{kn}^{(t)}$ means the weight of the kth topic within the nth document under the current document set. Moreover, $\mathcal{D}^{(t)} \in R^{M\times N}$, $\mathcal{U}^{(t)} \in R^{M\times K}$ and $\mathcal{V}^{(t)} \in R^{K\times N}$, where M, K and N correspond to the number of words, topics and documents, respectively, under the current topic learning task. The goodness of the fit between the text data and latent/representation factors could be formulated in the following optimization problem:

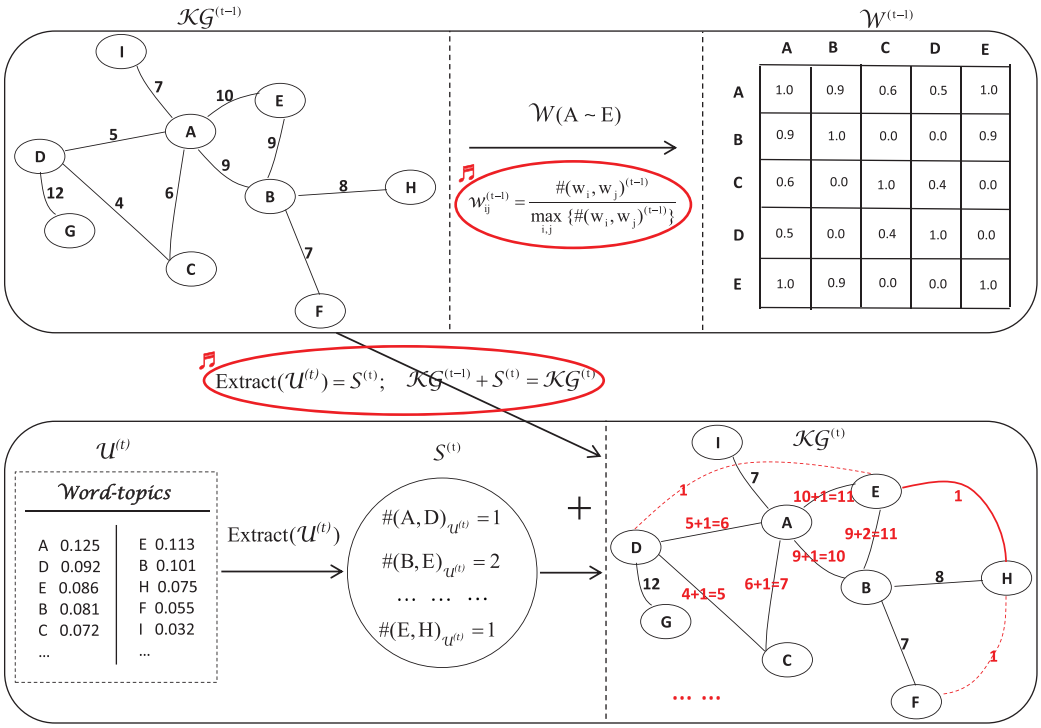


Fig. 2. The operations for knowledge graph. (top: $\mathcal{KG}^{(t-1)} \rightarrow \mathcal{W}^{(t-1)}$; bottom: $Extract(\mathcal{U}^{(t)}) \rightarrow \mathcal{S}^{(t)}$ and $\mathcal{KG}^{(t-1)} + \mathcal{S}^{(t)} \rightarrow \mathcal{KG}^{(t)}$).

$$\min f(\mathcal{D}^{(t)}||\mathcal{U}^{(t)}, \mathcal{V}^{(t)}) = ||\mathcal{D}^{(t)} - \mathcal{U}^{(t)}\mathcal{V}^{(t)}||_F^2$$
$$s.t.\ \mathcal{U}^{(t)} \geq 0,\ \mathcal{V}^{(t)} \geq 0, \tag{2}$$

where the non-negative constraints on word-topic $\mathcal{U}^{(t)}$ and topic-document $\mathcal{V}^{(t)}$ matrices would guide this Non-negative Matrix Factorization (NMF) to yield additive topics like human recognition [23]; therefore, the explanation and analysis on documents and topics are quite natural and practical for real-world scenarios.

4.2 Priors on Matrices $\mathcal{U}$ and $\mathcal{V}$

The general framework of machine learning has two parts: (1) data-based learning and (2) priors-guided instructions. The first part has already been discussed in the previous section. We then turn to the priors on $\mathcal{U}^{(t)}$ and $\mathcal{V}^{(t)}$ for higher quality topics. In general, there are several popular regularizations for characterizing the priors such as sparsity [30], [31], smoothness [32], diversity [29], [33] and affinity [34]. Given the streaming (labeled or not) documents in an endless pipeline, we would make use of the diversity and word-word affinity regularizations on the word-topic matrix $\mathcal{U}^{(t)}$, and sparsity and class-class affinity priors on the topic-document matrix $\mathcal{V}^{(t)}$ as semi-supervised guidance.

More specifically, the diversity regularization aims to reduce the overlapping of the learned topics $\mathcal{U}^{(t)}$ as much as possible, which could contribute to richer semantics given any fixed number of topics. The word-word affinity regularization is thus exerted on $\mathcal{U}^{(t)}$, with the purpose of capturing the word-word semantics from the shared knowledge graph. The sparsity prior on $\mathcal{V}^{(t)}$ aims to clearly depict that each document is just talking about limited topics, and the class-class affinity regularization for $\mathcal{V}^{(t)}$ holds that documents within a same class would share more topics than those distributed in different classes. Formally, priors $\Omega(\mathcal{U}^{(t)})$ and $\Psi(\mathcal{V}^{(t)})$ in Eq. (1) are modeled as Eqs. (3) and (5):

$$\Omega(\mathcal{U}^{(t)}) = \alpha \cdot \mathcal{K}^{(t-1)} + \beta \cdot ||(\mathcal{U}^{(t)})^\top \mathcal{U}^{(t)} - \mathcal{I}_K||_F^2, \tag{3}$$

where $\mathcal{K}^{(t-1)}$ denotes the word-word semantic affinity regularization, which is constructed from the shared knowledge and thus empowered to bridge the past and the present during lifelong topic learning. With respect to the details of $\mathcal{K}^{(t-1)}$, we have:

$$\mathcal{K}^{(t-1)} = g(\mathcal{KG}^{(t-1)}, \mathcal{U}^{(t)}) = tr((\mathcal{U}^{(t)})^\top \mathcal{H}^{(t-1)} \mathcal{U}^{(t)}), \tag{4}$$

where $\mathcal{H}^{(t-1)} = diag(\mathcal{W}^{(t-1)} \cdot 1) - \mathcal{W}^{(t-1)}$ marks the graph Laplacian of WWWNet (see Fig. 1). Here, $\mathcal{W}^{(t-1)}$ can adopt three popular forms as follows:

a) $$w_{ij}^{(t-1)} = \begin{cases} 1, & i = j \\ \frac{\#(w_i, w_j)^{(t-1)}}{\max_{i,j}\{\#(w_i, w_j)^{(t-1)}\}}, & \text{otherwise,} \end{cases}$$

b) $$w_{ij}^{(t-1)} = \begin{cases} 1, & i = j \\ \mathrm{I}\{\#(w_i, w_j)^{(t-1)} \neq 0\}, & \text{otherwise,} \end{cases}$$

c) $$w_{ij}^{(t-1)} = \begin{cases} 1, & i = j \\ \frac{1}{1+\exp\{-2w_{ij}^{(t-1)}\}}, & \#(w_i, w_j)^{(t-1)} \neq 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\#(w_i, w_j)^{(t-1)}$ (illustrated in Fig. 2) marks the number of co-occurrences of the ith and jth words in the $(t-1)$-th prior knowledge graph. $\mathrm{I}\{\cdot\}$ is an indicator function, which equals to 1 when the condition in the brace is satisfied, or 0 otherwise. Formula (a) in naive ratio means that the relatedness is consistent with the co-occurrence; the more related, the tighter in word-word semantics. Formula (b) adopts a (0,1) binary calculation assuming that w_i and w_j are definitely correlated as long as they appear in the WWWNet.

Formula (c) smooths (a) on the belief that w_i and w_j should be trusted with high probability because they are extracted from the previous topic matrix $\mathcal{U}^{(t-1)}$.

Similarly, considering the sparse encoding for each document under the large numbers of latent topics and the labeled documents, we draw the following priors:

$$\Psi(\mathcal{V}^{(t)}) = \lambda \cdot ||\mathcal{V}^{(t)}||_1^1 + \gamma \cdot tr(\mathcal{V}^{(t)}\mathcal{M}^{(t)}(\mathcal{V}^{(t)})^\top), \tag{5}$$

where $\mathcal{M}^{(t)} = diag(\mathcal{C}^{(t)} \cdot 1) - \mathcal{C}^{(t)}$ denotes the graph Laplacian of document's class-class graph, and $\mathcal{C}^{(t)}$ adopts the (0,1) binary weighted scheme. The above hyper-parameters α, β, λ and γ are preset through a series of experiments.

4.3 Knowledge Graph Maintenance

Knowledge graph is accumulated chunk by chunk with more and more reliable word-word pairs. Hence, just like a human, its brain ($\mathcal{KG}$) becomes smarter and smarter that would supervise the learning robot to perform better on the future topic mining tasks. In this paper, we extract word-word pairs in the top-10 words representing each topic, with the consideration that topic model could be treated as term clustering. Specifically, if w_i and w_j appear simultaneously in the top-10 words under the same topic, then add 1 to $\#(w_i, w_j)$; for the whole word-topic mapping $\mathcal{U}^{(t)}$, we finally accumulated all the possible pairs within the current task, marked by $\#(w_i, w_j)_{\mathcal{U}^{(t)}}$. Formally,

$$\mathcal{S}^{(t)} = Extract(\mathcal{U}^{(t)}) = \{\#(w_i, w_j)_{\mathcal{U}^{(t)}}\}, \tag{6}$$

where $\mathcal{S}^{(t)}$ denotes the set of all the mined word pairs from the current topic models. With such new knowledge, the previous WWWNet could be updated as:

$$\mathcal{KG}^{(t)} = \mathcal{KG}^{(t-1)} + \mathcal{S}^{(t)}, \tag{7}$$

which essentially proceeds on operations like

$$\#(w_i, w_j)^{(t)} = \#(w_i, w_j)^{(t-1)} + \#(w_i, w_j)_{\mathcal{U}^{(t)}}, \tag{8}$$

as illustrated in Fig. 2. In a word, the knowledge graph operations would guarantee the machine brain ($\mathcal{KG}$) becomes more and more knowledgeable and professional in semantics, which would constantly help the NMF-LTM learner to perform better and better.

4.4 The Overall Model

The overall objective of our lifelong topic model combined with the data-based NMF topic model in Eq. (2), the priors on $\mathcal{U}^{(t)}$ and $\mathcal{V}^{(t)}$ for instructions in Eqs. (3) and (5), and the knowledge graph maintenance for future topic learning in Eqs. (6) and (7), is written as follows:

$$\begin{aligned} \min \mathcal{J} &\triangleq f(\mathcal{D}^{(t)}||\mathcal{U}^{(t)}, \mathcal{V}^{(t)}) + \Omega(\mathcal{U}^{(t)}) + \Psi(\mathcal{V}^{(t)}) \\ s.t. &\begin{cases} \mathcal{U}^{(t)} \geq 0,\ \mathcal{V}^{(t)} \geq 0 \\ \mathcal{S}^{(t)} = Extract(\mathcal{U}^{(t)}) \\ \mathcal{KG}^{(t)} = \mathcal{KG}^{(t-1)} + \mathcal{S}^{(t)}. \end{cases} \end{aligned} \tag{9}$$

5 The Algorithm

In this section, we design the algorithm named NMF-LTM for the optimization of the objective function in Eq. (9). The convergence proof is also provided for theoretical understanding.

5.1 Algorithm Derivation and Design

Using the properties $||\mathcal{X}||_F^2 = tr(\mathcal{X}^\top\mathcal{X})$ and $tr(\mathcal{XY}) = tr(\mathcal{YX})$, where $\mathcal{X}$ and $\mathcal{Y}$ are two matrices with size $M \times N$ and $N \times M$, respectively, the whole objective function Eq. (9) is transformed as follows:

$$\begin{aligned} \mathcal{J} &\triangleq f(\mathcal{D}^{(t)}||\mathcal{U}^{(t)}, \mathcal{V}^{(t)}) + \Omega(\mathcal{U}^{(t)}) + \Psi(\mathcal{V}^{(t)}) \\ &= tr((\mathcal{V}^{(t)})^\top(\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)}\mathcal{V}^{(t)}) - 2tr((\mathcal{D}^{(t)})^\top\mathcal{U}^{(t)}\mathcal{V}^{(t)}) \\ &\quad + \alpha \cdot tr((\mathcal{U}^{(t)})^\top diag(\mathcal{W}^{(t-1)}1)\mathcal{U}^{(t)}) \\ &\quad - \alpha \cdot tr((\mathcal{U}^{(t)})^\top\mathcal{W}^{(t-1)}\mathcal{U}^{(t)}) \\ &\quad + \beta \cdot tr((\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)}(\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)}) - 2\beta \cdot tr((\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)}) \\ &\quad + \lambda \cdot tr(1^\top\mathcal{V}^{(t)}1) - \gamma \cdot tr(\mathcal{V}^{(t)}\mathcal{C}^{(t)}(\mathcal{V}^{(t)})^\top) \\ &\quad + \gamma \cdot tr(\mathcal{V}^{(t)}diag(\mathcal{C}^{(t)}1)(\mathcal{V}^{(t)})^\top) + const, \end{aligned} \tag{10}$$

where $const$ is a constant with respect to $\mathcal{U}^{(t)}$ and $\mathcal{V}^{(t)}$.

Let ρ_{mk} and θ_{kn} be the Lagrange multipliers for constraints $\mathcal{U}_{mk}^{(t)} \geq 0$ and $\mathcal{V}_{kn}^{(t)} \geq 0$, respectively, i.e., $\mathrm{P} = (\rho_{mk})$, and $\Theta = (\theta_{kn})$. The Lagrangian function is given as:

$$\mathcal{L} = \mathcal{J} + tr(\mathcal{U}^{(t)}\mathrm{P}^\top) + tr(\mathcal{V}^{(t)}\Theta^\top). \tag{11}$$

The derivatives of $\mathcal{L}$ with respect to $\mathcal{U}^{(t)}$ and $\mathcal{V}^{(t)}$ are listed as follows:

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\mathcal{U}^{(t)}} &= 2\mathcal{U}^{(t)}\mathcal{V}^{(t)}(\mathcal{V}^{(t)})^\top + 2\alpha \cdot diag(\mathcal{W}^{(t-1)}1)\mathcal{U}^{(t)} \\ &\quad - 2\mathcal{D}^{(t)}(\mathcal{V}^{(t)})^\top - 2\alpha \cdot \mathcal{W}^{(t-1)}\mathcal{U}^{(t)} \\ &\quad + 4\beta \cdot \mathcal{U}^{(t)}(\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)} - 4\beta \cdot \mathcal{U}^{(t)} + \mathrm{P}, \end{aligned} \tag{12}$$

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\mathcal{V}^{(t)}} &= 2(\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)}\mathcal{V}^{(t)} - 2(\mathcal{U}^{(t)})^\top\mathcal{D}^{(t)} + \lambda \cdot 1 \cdot 1^\top \\ &\quad + 2\gamma \cdot \mathcal{V}^{(t)}diag(\mathcal{C}^{(t)}1) - 2\gamma \cdot \mathcal{V}^{(t)}\mathcal{C}^{(t)} + \Theta. \end{aligned} \tag{13}$$

By the KKT conditions: $\rho_{mk}\mathcal{U}_{mk}^{(t)} = 0$ and $\theta_{kn}\mathcal{V}_{kn}^{(t)} = 0$, we have the following equations for $\mathcal{U}_{mk}^{(t)}$ and $\mathcal{V}_{kn}^{(t)}$:

$$\begin{aligned} &\{2\mathcal{D}^{(t)}(\mathcal{V}^{(t)})^\top + 2\alpha \cdot \mathcal{W}^{(t-1)}\mathcal{U}^{(t)} + 4\beta \cdot \mathcal{U}^{(t)}\}_{mk}\mathcal{U}_{mk}^{(t)} \\ &- \{2\mathcal{U}^{(t)}\mathcal{V}^{(t)}(\mathcal{V}^{(t)})^\top + 2\alpha \cdot diag(\mathcal{W}^{(t-1)}1)\mathcal{U}^{(t)} \\ &+ 4\beta \cdot \mathcal{U}^{(t)}(\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)}\}_{mk}\mathcal{U}_{mk}^{(t)} = 0, \end{aligned} \tag{14}$$

$$\begin{aligned} &\{2(\mathcal{U}^{(t)})^\top\mathcal{D}^{(t)} + 2\gamma \cdot \mathcal{V}^{(t)}\mathcal{C}^{(t)}\}_{kn}\mathcal{V}_{kn}^{(t)} \\ &- \{2(\mathcal{U}^{(t)})^\top\mathcal{U}^{(t)}\mathcal{V}^{(t)} + \lambda \cdot 1 \cdot 1^\top \\ &+ 2\gamma \cdot \mathcal{V}^{(t)}diag(\mathcal{C}^{(t)}1)\}_{kn}\mathcal{V}_{kn}^{(t)} = 0. \end{aligned} \tag{15}$$

Therefore, the multiplicative update rules for $\mathcal{U}^{(t)}$ and $\mathcal{V}^{(t)}$ can be inferred as below:

$$\begin{aligned}\mathcal{U}_{(i+1)}^{(t)} = \mathcal{U}_{(i)}^{(t)} \odot \{\mathcal{D}^{(t)}(\mathcal{V}_{(i)}^{(t)})^\top + \alpha \cdot \mathcal{W}^{(t-1)} \cdot \mathcal{U}_{(i)}^{(t)} \\ + 2\beta \cdot \mathcal{U}_{(i)}^{(t)}\} \oslash \{\mathcal{U}_{(i)}^{(t)}\mathcal{V}_{(i)}^{(t)}(\mathcal{V}_{(i)}^{(t)})^\top \\ + \alpha \cdot diag(\mathcal{W}^{(t-1)}1) \cdot \mathcal{U}_{(i)}^{(t)} + 2\beta \cdot \mathcal{U}_{(i)}^{(t)}(\mathcal{U}_{(i)}^{(t)})^\top \mathcal{U}_{(i)}^{(t)}\},\end{aligned} \tag{16}$$

$$\begin{aligned}\mathcal{V}_{(i+1)}^{(t)} = \mathcal{V}_{(i)}^{(t)} \odot \{\mathcal{U}_{(i+1)}^{(t)}\mathcal{D}^{(t)} + \gamma \cdot \mathcal{V}_{(i)}^{(t)}\mathcal{C}^{(t)}\} \oslash \\ \{(\mathcal{U}_{(i+1)}^{(t)})^\top \mathcal{U}_{(i+1)}^{(t)}\mathcal{V}_{(i)}^{(t)} + \gamma \cdot \mathcal{V}_{(i)}^{(t)} \cdot diag(\mathcal{C}^{(t)}1) + \frac{\lambda}{2}1 \cdot 1^\top\},\end{aligned} \tag{17}$$

where $\odot$ represents the element-wise product, and $\oslash$ denotes the element-wise division. The corresponding algorithm is then designed as Algorithm 1.

Algorithm 1. NMF-LTM

Input: Docs-Chunk $\mathcal{D} = \{\mathcal{D}^{(t)}\}_{t=1}^{T}$ in sequence, T can be infinite; Hyper-parameters: α, β, λ, and γ.
Output: $(\mathcal{U}, \mathcal{V}) = \{(\mathcal{U}^{(t)}, \mathcal{V}^{(t)})\}_{t=1}^{T}$ and $\mathcal{KG}$.

begin
  $\mathcal{KG}^{(0)} = \emptyset$;
  for $t = 1, 2, \ldots, T$ **do**
    Randomly initialize nonnegative $\mathcal{U}_{(0)}^{(t)}$ and $\mathcal{V}_{(0)}^{(t)}$;
    Construct $\mathcal{W}^{(t-1)}$ from $\mathcal{KG}^{(t-1)}$;
    Construct $\mathcal{C}^{(t)}$ from $\mathcal{D}^{(t)}$;
    for $i = 0, 1, \cdots, I$ **do**
      Update $\mathcal{U}_{(i+1)}^{(t)}$ according to Eq. (16);
      Update $\mathcal{V}_{(i+1)}^{(t)}$ according to Eq. (17);
      if *convergent precision is satisfied* **then**
        break;
      end
    end
    $\mathcal{S}^{(t)} = Extract(\mathcal{U}^{(t)})$;
    $\mathcal{KG}^{(t)} = \mathcal{KG}^{(t-1)} + S^{(t)}$;
  end
end

5.2 Convergence Analysis

Generally, it is quite challenging to find out the global solutions for the objective due to its non-convexity. Fortunately, it could at least be shown to converge to a local optimum with the multiplicative update rules. In what follows, before proposing the theoretical guarantee, a definition and a lemma are borrowed from Ref. [35] first.

Definition 1. *$G(h, h^*)$ is an auxiliary function for $\mathcal{J}(h)$ if the conditions:*

$$\begin{cases} G(h, h^*) \geq \mathcal{J}(h) \\ G(h, h) = \mathcal{J}(h), \end{cases} \tag{18}$$

are satisfied.

Lemma 1. *If G is an auxiliary function, then $\mathcal{J}$ is non-increasing under update rule:*

$$h^{(i+1)} = \arg\min_{h} G(h, h^{(i)}). \tag{19}$$

Proof. $\mathcal{J}(h^{(i)}) = G(h^{(i)}, h^{(i)}) \geq G(h^{(i+1)}, h^{(i)}) \geq \mathcal{J}(h^{(i+1)})$. □

Based on the above definition and the lemma, we have:

Lemma 2. *Function*

$$\begin{aligned}G(\mathcal{V}_{kn}, \mathcal{V}_{kn}^{(i)}) = \mathcal{J}(\mathcal{V}_{kn}^{(i)}) + \nabla\mathcal{J}(\mathcal{V}_{kn}^{(i)})(\mathcal{V}_{kn} - \mathcal{V}_{kn}^{(i)}) \\ + \frac{(\mathcal{U}^\top\mathcal{U}\mathcal{V}^{(i)} + \gamma\mathcal{V}^{(i)} \cdot diag(\mathcal{C} \cdot 1) + \frac{\lambda}{2}1 \cdot 1^\top)_{kn}}{\mathcal{V}_{kn}^{(i)}}(\mathcal{V}_{kn} - \mathcal{V}_{kn}^{(i)})^2\end{aligned} \tag{20}$$

is an auxiliary function for $\mathcal{J}(\mathcal{V}_{kn})$.

Proof. Obviously, $G(\mathcal{V}_{kn}, \mathcal{V}_{kn}) = \mathcal{J}(\mathcal{V}_{kn})$ when there is $\mathcal{V}_{kn}^{(i)} = \mathcal{V}_{kn}$. Then we only need to prove that $G(\mathcal{V}_{kn}, \mathcal{V}_{kn}^{(i)}) \geq \mathcal{J}(\mathcal{V}_{kn})$. We expand $\mathcal{J}(\mathcal{V}_{kn})$ using Taylor series as below:

$$\begin{aligned}\mathcal{J}(\mathcal{V}_{kn}) = \mathcal{J}(\mathcal{V}_{kn}^{(i)}) + \nabla\mathcal{J}(\mathcal{V}_{kn}^{(i)})(\mathcal{V}_{kn} - \mathcal{V}_{kn}^{(i)}) \\ + ((\mathcal{U}^\top\mathcal{U})_{kk} + \gamma \cdot diag(\mathcal{C} \cdot 1)_{nn} - \gamma \cdot \mathcal{C}_{nn})(\mathcal{V}_{kn} - \mathcal{V}_{kn}^{(i)})^2.\end{aligned} \tag{21}$$

Meanwhile, there exists:

$$(\mathcal{U}^\top\mathcal{U}\mathcal{V}^{(i)})_{kn} = \sum_p (\mathcal{U}^\top\mathcal{U})_{kp}\mathcal{V}_{pn}^{(i)} \geq (\mathcal{U}^\top\mathcal{U})_{kk}\mathcal{V}_{kn}^{(i)}, \tag{22}$$

and

$$\begin{aligned}(\mathcal{V}^{(i)} \cdot diag(\mathcal{C} \cdot 1))_{kn} = \sum_p \mathcal{V}_{kp}^{(i)} diag(\mathcal{C} \cdot 1)_{pn} \\ \geq \mathcal{V}_{kn}^{(i)} diag(\mathcal{C} \cdot 1)_{nn}.\end{aligned} \tag{23}$$

By comparing Eq. (20) with Eq. (21), we have $G(\mathcal{V}_{kn}, \mathcal{V}_{kn}^{(i)}) \geq \mathcal{J}(\mathcal{V}_{kn})$, and Lemma 2 thus follows. □

Lemma 3. *Function*

$$\begin{aligned}G(\mathcal{U}_{mk}, \mathcal{U}_{mk}^{(i)}) = \mathcal{J}(\mathcal{U}_{mk}^{(i)}) + \nabla\mathcal{J}(\mathcal{U}_{mk}^{(i)})(\mathcal{U}_{mk} - \mathcal{U}_{mk}^{(i)}) \\ + \frac{(\mathcal{U}^{(i)}\mathcal{V}\mathcal{V}^\top + \alpha \cdot diag(\mathcal{W}^{(t-1)}1) \cdot \mathcal{U}^{(i)} + 2\beta \cdot \mathcal{U}^{(i)}(\mathcal{U}^{(i)})^\top\mathcal{U}^{(i)})_{mk}}{\mathcal{U}_{mk}^{(i)}} \\ \times (\mathcal{U}_{mk} - \mathcal{U}_{mk}^{(i)})^2,\end{aligned} \tag{24}$$

is an auxiliary function for $\mathcal{J}(\mathcal{U}_{mk})$.

Proof. We omit the proof here for concision. Readers with interests can refer to Refs. [35], [36] for similar proofs. □

Theorem 1. *The objective function $\mathcal{J}(\mathcal{U}, \mathcal{V})$ is convergent to a local minimum given the iterative rules in Eqs. (16) and (17).*

Proof. First, with Lemmas 1, 2 and 3, we could achieve:

$$\mathcal{J}(\mathcal{U}^{(i)}, \mathcal{V}^{(i)}) \geq \mathcal{J}(\mathcal{U}^{(i+1)}, \mathcal{V}^{(i)}), \tag{25}$$

and

$$\mathcal{J}(\mathcal{U}^{(i+1)}, \mathcal{V}^{(i)}) \geq \mathcal{J}(\mathcal{U}^{(i+1)}, \mathcal{V}^{(i+1)}), \tag{26}$$

where i denotes the iteration index. Note that the specific proof details of the above two results in Eqs. (25) and (26) could be found in Ref. [35].

Then there holds:

$$\mathcal{J}(\mathcal{U}^{(i)}, \mathcal{V}^{(i)}) \geq \mathcal{J}(\mathcal{U}^{(i+1)}, \mathcal{V}^{(i)}) \geq \mathcal{J}(\mathcal{U}^{(i+1)}, \mathcal{V}^{(i+1)}), \tag{27}$$

TABLE 1
Statistics of Data Sets

Dataset	#Label	#Docs	#Words	#Vocabulary
Tweet	1	20000	8.24	13621
Snippet	8	12283	15.32	4067
Reviews	50	337559	3.39	5574

(#Label: the number of ground-truth labels or categories; #Docs: the total number of documents; #Words: the average number of words per document; #Vocabulary: the total number of distinctive terms in the corpus).

which indicates the non-increasing properties with the updating rules in Eqs. (16) and (17). Moreover, the whole objective function $\mathcal{J}(\mathcal{U},\mathcal{V})$ is nonnegative and therefore lower-bounded by zero. These two conditions together guarantee that $\mathcal{J}(\mathcal{U},\mathcal{V})$ is convergent to a local minimum. The theorem thus follows. □

The detailed procedures in the NMF-LTM (Algorithm 1) exhibit that any two adjacent timeslots are independent of each other after constructing word-word affinity regularizations and class-class affinity priors, corresponding to line 5 and line 6 in Algorithm 1, respectively. Therefore, Theorem 1 indicates the convergence for any timeslot t in the lifelong learning process, which naturally reveals the convergence of the whole NMF-LTM algorithm.

6 Experimental Results

In this section, we conduct extensive comparative experiments to evaluate the performance of NMF-LTM.

6.1 Experimental Setup

6.1.1 Datasets and Preprocessing

Table 1 shows the statistics of the three experimental data sets. The *Tweet* data set is a collection of about 360000 labeled tweets within 10 different categories. It was crawled and labeled by ODP,[3] a community-driven open access web page directory. In our experiment, we choose one category, i.e., Arts_ids with 20000 tweets, for lifelong topic modeling of relatively dense concepts. Data preprocessing includes converting letters to lowercase, filtering non-alphabetic characters and stop words, retaining nominal words and further removing terms with document frequency less than 3. We finally obtain a dictionary with 13621 unique words, and the average length of tweets is 8.24. To obtain sequential docs-chunks for LTMs, we then divide *Tweet* into 20, 50, 100 and 200 batches in time order, each containing 1000, 400, 200 and 100 tweets, respectively. Since only one category (Arts_ids) is used, the shared knowledge between sequential learning tasks is deemed to be dense.

The *Snippet* archive[4] was created by Phan et al. [37] and is composed of 12K+ snippets drawn from Google. The text preprocessing is the same as that for *Tweet*, which results in 12283 snippets in 8 categories, with the group sizes varying from 369 to 2654 and a word vocabulary size of 4067. We then randomly split the whole 12K+ documents into 8 data chunks, each containing about 1.5K snippets with an average text length about 15.32. Note that since each data chunk consists of snippets from various categories, the shared knowledge between sequential learning tasks is expected to be looser than that of *Tweet*.

The *Reviews* data set contains consumer reviews to 50 product categories, each of which contains 1000 reviews evenly. This corpus was crawled from Amazon.com and now is publicly available.[5] We preprocess this collection similarly as the above two, and finally obtain 337559 sentences as different comments and 5574 distinctive terms. We then randomly partition the review comments into 50 docs-chunks in a streaming pipeline, each accommodating about 6751 comments with the shortest average text length: 3.39. Note that both *Snippet* and *Reviews* are with multiple classes, which ensure lifelong learning conducted in practical scenarios.

Note that the *Tweet* data set has time tags and therefore is divided into batches in time order; but the other two publicly available data sets *Snippet* and *Reviews* are collected without time labels, so they are randomly partitioned into chunks to simulate the time-sequential tasks in the lifelong learning process.

6.1.2 Baselines and Evaluation

As indicated by Ref. [6], LDA-LTM outperforms many well-designed baselines, including AKL [22], GK-LDA [20], DF-LDA [19], etc. in terms of topic quality. We therefore mainly compare NMF-LTM with LDA-LTM in this paper, but allow for multiple variants with different configurations of parameters (see Table 2 for the variants). Moreover, we also compare NMF-LTM with the classic LDA [12] and NMF [23] models to highlight the benefits from knowledge sharing.

For model evaluation, we adopt the popular *Topic Coherence* [38] measure. Specifically, given a topic $\mathcal{U}_k$ with $topT$ (usually $topT = 20$) indicative words, the topic coherence is computed as

$$\mathcal{C}(\mathcal{U}_k, topT) = \sum_{1 \le i < j \le topT} \log \frac{\mathcal{D}(w_i^k, w_j^k) + 1}{\mathcal{D}(w_i^k)}, \tag{28}$$

where $\mathcal{D}(w_i^k, w_j^k)$ represents the total number of documents containing the words w_i and w_j in topic $\mathcal{U}_k$; likewise, $\mathcal{D}(w_i^k)$ denotes the number of documents containing the word w_i in the kth topic. For the general performance of a topic model, the evaluation is the topic coherence averaged on all the K topics. Moreover, we also explore the effectiveness of NMF-LTM via *Lifelong Machine Learning Test* [39], which measures an agent's or a model's learning better-ness and faster-ness.

All the experiments are conducted on a workstation in C++/Java environment equipped with Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50 GHz, 24 Cores and 128G memory.

6.1.3 Parameter Settings

For LDA-based models, the common hyper-parameters α and β for Dirichlet distribution are set to 1 and 0.1, respectively, after careful sensitivity analysis. The other parameters are set as suggested by Ref. [6]. For NMF-based models,

3. http://dmoz.org
4. http://acube.di.unipi.it/tmn-dataset/
5. https://www.cs.uic.edu/~zchen/downloads/ICML2014-Chen-Dataset.zip

TABLE 2
Lookup Table for All Competitors

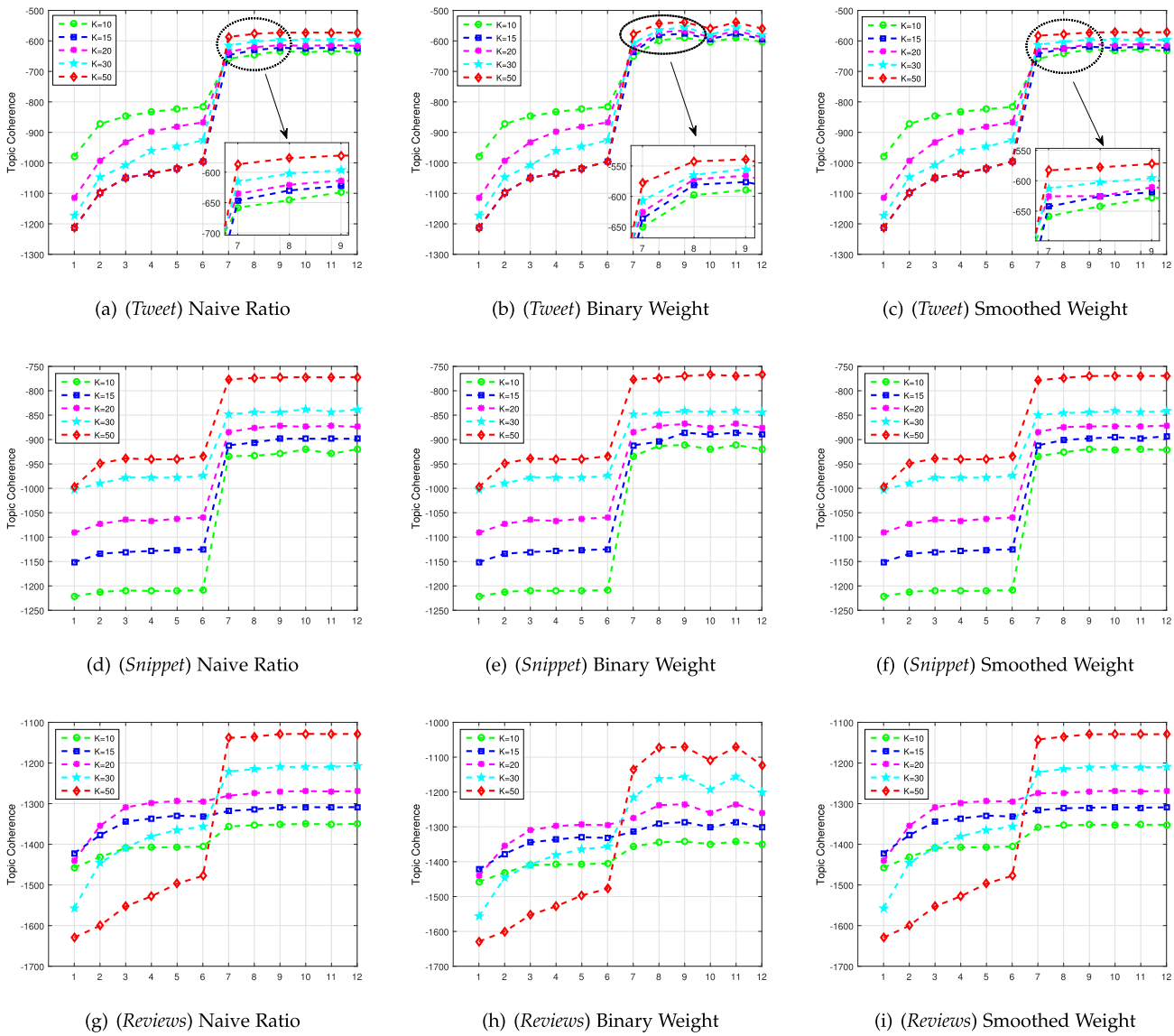
ID	Name	Detail
1	LDA	classic LDA without knowledge sharing
2-5	LDA-LTM-i	LDA-based LTM with i times iterations, i=1,...,4
6	LDA-LTM-5	LDA-based LTM with 5 iterations, the best in [6]
7	NMF	classic NMF without knowledge sharing
8	NMF-LTM-α	NMF-based LTM
9	NMF-LTM-$\alpha\lambda$	NMF-based LTM with sparsity prior
10	NMF-LTM-$\alpha\lambda\beta$	NMF-based LTM with sparsity, diversity priors
11	NMF-LTM-$\alpha\lambda\gamma$	NMF-based LTM with sparsity, supervised priors
12	NMF-LTM-all	NMF-based LTM with all types of priors

there are four parameters: α, β, λ and γ, representing the knowledge graph regularization, topic diversity, sparse coding, and label-supervised regularization, respectively. To set these parameters properly, we take the *Tweet* dataset (20 batches $\times$ 1000 tweets/batch, the number of topics $K = 15$ and 30, three forms of $\mathcal{KG}$ weights), change each parameter with the values ranging from 1e-6, 1e-5 to 1e2 while holding other parameters fixed, and record the final $3 \times 2 \times 9 \times 9 \times 9 \times 9$ results. We finally adopt the configuration $\alpha = 10$, $\beta = 0.5$, $\lambda = 0.001$ and $\gamma = 0.001$ for NMF-based models after careful sensitivity analysis.

6.2 Comparison among Lifelong Topic Models

Here, we compare our NMF-LTM with various baselines including LDA-LTM [6], the classic LDA [12] and NMF [23], and some variants. Table 2 lists all the competitors, where the methods with ID numbers from 2 to 6 are the variants of LDA-LTM with different numbers of knowledge-learning iterations, and the methods with ID numbers from 8 to 12 are the variants of NMF-LTM with different combinations of priors.

Fig. 3 reports the average topic coherence values returned by the competitors conducted on different data chunks. The topic number is set to 10, 15, 20, 30 and 50, respectively, for

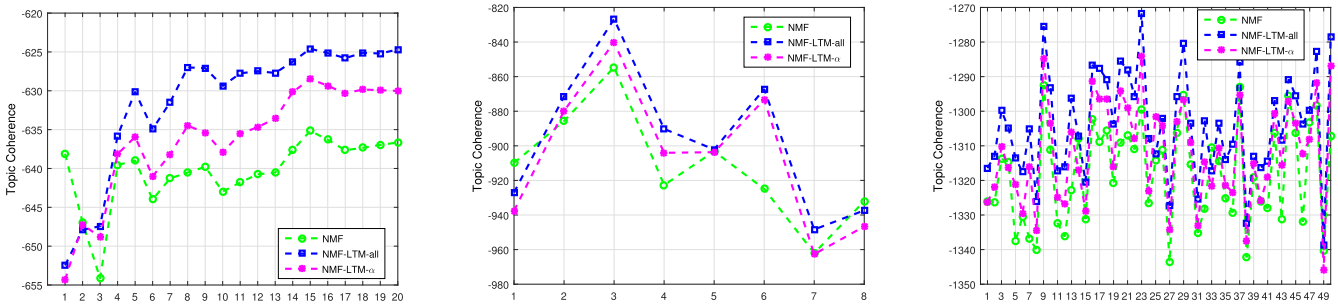


(a) (*Tweet*) Naive Ratio (b) (*Tweet*) Binary Weight (c) (*Tweet*) Smoothed Weight

(d) (*Snippet*) Naive Ratio (e) (*Snippet*) Binary Weight (f) (*Snippet*) Smoothed Weight

(g) (*Reviews*) Naive Ratio (h) (*Reviews*) Binary Weight (i) (*Reviews*) Smoothed Weight

Fig. 3. Performance of 12 competitors on three data sets. (1. Three types of knowledge formulations are tried for NMF-based methods. 2. K is the topic number. 3. The numbers on the x-axis denote different competitors, i.e., 1: LDA, 2: LDA-LTM-1, 3: LDA-LTM-2, 4: LDA-LTM-3, 5: LDA-LTM-4, 6: LDA-LTM-5, 7: NMF, 8: NMF-LTM-α, 9: NMF-LTM-$\alpha\lambda$, 10: NMF-LTM-$\alpha\lambda\beta$, 11: NMF-LTM-$\alpha\lambda\gamma$, and 12: NMF-LTM-all. More details are provided in Table 2.).



(a) (*Tweet*) Naive Ratio (b) (*Snippet*) Naive Ratio (c) (*Reviews*) Naive Ratio

Fig. 4. Performance of three competitors in lifelong machine learning test. (1. Both NMF-LTM-α and NMF-LTM-all employ Naive Ratio mode. 2. The numbers on the x-axis denote different batches. 3. The topic number $K = 15$.).

sensitivity analysis. For NMF-based methods (with ID numbers from 7 to 12), all the three forms of WWWNet graph Laplacian are tried for the purpose of comparison. For the *Tweet* data set, we adopt the partition with 20 data chunks. From the figure, we have some interesting observations as follows:

- Regardless of the topic numbers and the knowledge representations, the NMF-based models are significantly superior to the LDA-based models in obtaining high-quality topics from each data set. This demonstrates the effectiveness of the matrix factorization framework for lifelong topic modeling, especially for short-text collections that lack of semantics from a probabilistic perspective.
- Among all the sub-figures, NMF-LTM-α performs consistently better than NMF. This indicates that the lifelong learning scheme with self-learnt knowledge sharing is indeed beneficial to enhancing the classic NMF. Similar situations can be found for LDA versus LDA-LTM-i ($i = 1, 2, \ldots, 5$), which further demonstrate the effectiveness of LTMs.
- In general, the variants of NMF-LTM with different forms of priors display higher-quality topics than NMF-LTM-α. This indeed suggests the goodness of some priors including data sparsity, topic diversity and label-supervised information in lifelong topic modeling.
- In regard to the adopted three strategies for $\mathcal{KG}$ construction, the difference between Naive Ratio and Smoothed Weight is barely perceptible; Binary Weight shows slightly higher values than the others on *Tweet* and *Reviews*, but it also brings in extra fluctuations. Thus, for a comprehensive consideration, the Naive Ratio is preferred for constructing $\mathcal{KG}$ in NMF-LTMs in subsequent studies.

Fig. 4 further showcases the detailed results for Lifelong Machine Learning Test of the three NMF-based methods, i.e., NMF, NMF-LTM-α and NMF-LTM-*all*, given the Naive Ratio mode as knowledge representation and a topic number $K = 15$. Some interesting observations are given as below:

- First, it is obvious that on nearly all data chunks in the x-axes, NMF with knowledge sharing and/or mixed with other priors performs much better than the standard NMF without any prior knowledge. This highlights the dominant strength of lifelong learning.
- Second, in terms of the learning rate, "NMF-LTM-*all*" and "NMF-LTM-α" generally hold much more fasterness on the dataset *Tweet* with dense knowledge. That is, both the two models exhibit a rising trend as more knowledge accumulated from the first batch to the last one. This means the lifelong learning is very efficient for such an intelligent agent like NMF-LTM.

Since NMF-LTM-*all* and LDA-LTM-5 generally perform better and more robustly in topic learning, in what follows, we agree that *NMF-LTM and LDA-LTM correspond to NMF-LTM-all (with Naive Ratio) and LDA-LTM-5, respectively*, for simplicity.

6.3 NMF-LTM for Human-Like Learning

In order to test the human-like learning ability of NMF-LTM, we elaborately design two tasks to simulate human behaviors: learning with different strategies and learning with multi-run reviews. The first task is to explore the preference of NMF-LTM to two human learning strategies, i.e., frequent learning with small amounts of knowledge per time, or infrequent learning with large amounts of knowledge per time. It is realized by running NMF-LTM on various document streams with a same total size but different numbers of batches. The second task introduces the review scheme into the learning process by having NMF-LTM run in a user-specific multiple times. The *Tweet* data set that has been preprocessed into multiple docs-streams with 20, 50, 100 and 200 batches, respectively, is adopted for experiments, and the topic number is set as $K = 50$.

6.3.1 Different Strategies

Fig. 5 shows the experimental results of running NMF-LTM on four docs-streams generated from the *Tweet* data set, where "x*y" label on the horizontal axis means x batches with y documents per batch. As can be seen, different learning strategies indeed have great impact to the learning outcome of NMF-LTM, and the best performance comes from the smallest batch size $y = 100$. This implies that for knowledge dense data sets such as *Tweet*, the frequent learning strategy is more suitable to NMF-LTM, upon which more useful knowledge can be accumulated effectively as learning guidance.

6.3.2 Multi-Run Reviews

Fig. 6 illustrates the learning performance of NMF-LTM with different runs of review on *Tweet* (20 batches with 1000

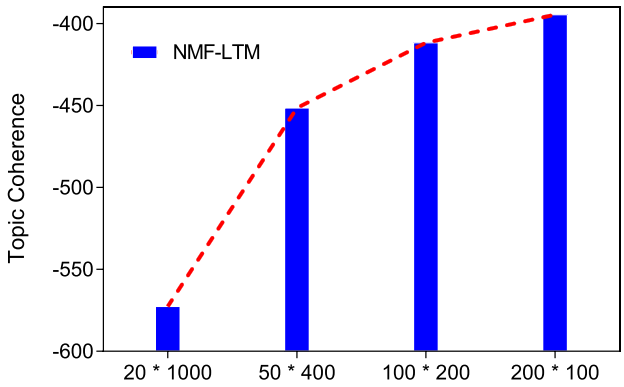


Fig. 5. Learning with different strategies on *Tweet*. ($K = 50$).

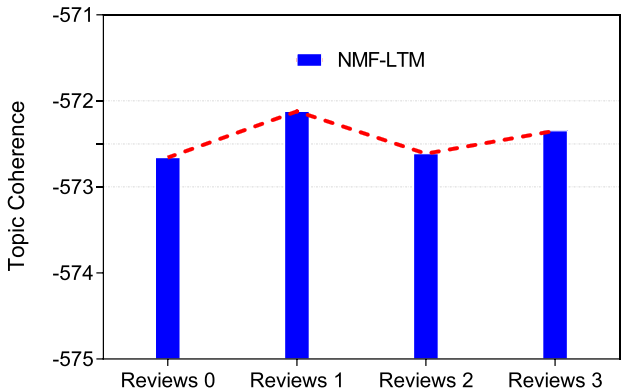


Fig. 6. Learning with multi-run reviews on *Tweet*. ($K = 50$).

tweets/batch). It's interesting that NMF-LTM shows rather stable performances with varied numbers of review process, i.e., the topic coherence value fluctuates mildly between -572 and -573. This implies that the multi-run review scheme might not be suitable to NMF-LTM, which has captured the essential topic information during the knowledge sharing process enabled by the affinity matrix. In other words, NMF-LTM demonstrates its high self-learning efficiency in this experiment, which is in sharp contrast to LDA-LTM that needs multi-run reviews for improvement, as illustrated in Fig. 3.

6.4 NMF-LTM for Big Data Modeling

Topic models like LDA and NMF are often criticized for the scalability issue. A direct solution is to take the divide-and-conquer strategy to partition big data into various smaller blocks for learning. Here, we demonstrate NMF-LTM is a natural fit to this solution and can generate high-quality topics. To this end, we take the *Snippet* corpus with 8 batches as the experimental data set, and employ NMF-LTM, LDA-LTM, LDA-all (classic LDA on all the documents) and NMF-all (standard NMF on all the documents) for topic modeling. The number of topics is set as $K = 30$. Note that *Snippet* here is a simulation of big data such that LDA-all and NMF-all can finish running quickly. A real big data practice will be given as a case study in Section 8.

Fig. 7 shows the comparative results, where NMF-LTM and LDA-LTM return the average topic coherence values, and NMF-all and LDA-all return the one-shot values. As can be seen, both the two lifelong learning models perform much better than the one-shot models, which indicates the excellent suitability of lifelong learning to topic modeling of big data. In particular, NMF-LTM has a dominant advantage in computational efficiency, which will be touched in the later section for case study of big data.

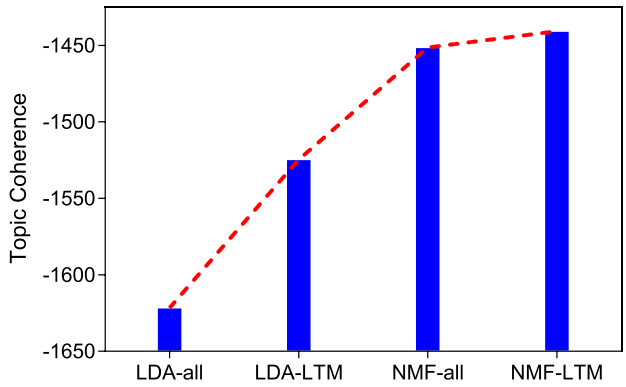


Fig. 7. Topic modeling of big data: A simulation on *Snippet*. ($K = 30$).

7 Exploration of Semantic Relatedness

7.1 Motivation and Background

To explore the semantic quality of the shared knowledge graph in our proposed NMF-LTM, we design an experiment for semantic relatedness computing with different methods, e.g., LDA [12], NMF [23], word2vec_skipgram [40], [41], word2vec_cbow [40], [41] and our NMF-LTM. Note that the former four learning models embed each word with a vector and the relatedness between two words can then be computed as the cosine value, namely:

$$R(w_i, w_j) = \cos(w_i, w_j) = \frac{(w_i)^\top (w_j)}{||w_i|| \times ||w_j||}, \tag{29}$$

where w_i (w_j) represents an embedding vector for the ith (jth) word. With respect to NMF-LTM, the $\mathcal{KG}$ is constructed with weighted word-word network; therefore, the weight between two words just denotes the corresponding semantic relatedness. However, LDA-LTM maintains the $\mathcal{KG}$ in frequent item-sets and does not hold the differentiated weights, and therefore it is not adopted in the competitive study. By default, we set the embedding dimension as 50 in all the competing methods for the purpose of fairness.

7.2 Data and Preprocessing

For semantic relatedness computing, we choose a relatively large data set, dubbed *SogouNews*.[6] It contains tens of thousands text-segments collected from Sohu news website,[7] most of which are edited and classified manually into 18 categories. In this experiment, we randomly select 200,000 pieces of news with each treated as one document. Then we leverage the popular NLP tool Ansj[8] to split the Chinese texts into word sequences with part-of-speech, filter meaningless words such as non-Chinese tokens and non-topic words, and finally obtain 200,000 lines, each corresponding to one piece of news. Further statistics show that the average length of the

6. http://www.sogou.com/labs/resource/list_news.php
7. http://news.sohu.com/
8. https://github.com/NLPchina/ansj_seg

whole corpus is 174.4 (words per news) and the volume of the vocabulary is 53,816 in total.

For word embedding, 200,000 documents are arranged in one text as a long word sequence to feed word2vecs,[9] with the experimental configurations as follows: {./word2vec -train SogouNews.txt -output vectors.bin -cbow 0 -size 50 -window 5 -negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 1 -cbow 0: Skip-gram; -cbow 1: CBOW}. With respect to the classic LDA, each line represents one document and the hyper-parameters α and β are set the same as above. For the classic NMF, the *SogouNews* corpus is encoded as a TF-IDF matrix. For NMF-LTM, the 200,000 documents are randomly divided into 200 data-chunks in sequence as streaming tasks, each chunk with 1,000 documents.

7.3 Evaluation Measures

To evaluate the semantic effectiveness of generated knowledge by different competitors, we first rank word pairs according to their semantic relatedness, and then employ Words-240[10] for rank evaluation. Words-240 includes 240 Chinese pairs of word relatedness with averaged human-judged marks, which could be used to measure the consistency of two ranked sequences, i.e., the models' semantic orders and the human-judged semantic orders.

NDCG@Υ[11] [42] is a measure widely used for quantitatively assessing two sequences' consistency. The parameter Υ represents the number of word pairs that are randomly selected for evaluation, which is set to 2, 3, 5, 7 and 10, respectively, as in previous studies. Formally,

$$NDCG@\Upsilon = \frac{DCG@\Upsilon}{IDCG@\Upsilon}, \tag{30}$$

where

$$DCG@\Upsilon = \sum_{i=1}^{\Upsilon} \frac{2^{r(i)} - 1}{\log_2(i+1)}, \tag{31}$$

$IDCG@\Upsilon$ is the possible maximum value (ideal) of $DCG@\Upsilon$ for a given set of word-pairs, and $r(i)$ represents the rank of the ith word pair in the ranking list of Υ word pairs.

7.4 Results and Discussions

In the experiment, we randomly select 100 groups of word pairs for each fixed Υ (2, 3, 5, 7 and 10) and average the 100 NDCG@Υ values as the final grade to evaluate the quality of semantic relatedness. Fig. 8 shows the results by different competitors, from which several interesting observations are as follows:

- Compared with LDA, NMF and NMT-LTM show better performances, and the advantage is more obvious as Υ becomes larger. This indicates the strong ability of NMF-based methods for semantics.
- By comparing NMF-LTM and NMF, it is obvious that NMF-LTM further improves the quality of semantic relatedness. This might attribute to the fact that NMF-LTM is a constant learning scheme with knowledge instructions while NMF is just one-shot learning from batch documents. Indeed for NMF-LTM, the NDCG@Υ values under different configurations are all above 0.9, showing the significant high quality of the learned $\mathcal{KG}$ in terms of word pairs' semantics.
- While NMF-LTM achieves high-quality semantic relatedness, there still exists an evident gap to the best performance obtained by word embedding methods. This is not unusual, since word2vec is trained with millions of samples to embed every word for semantics representation purposefully, while NMF-LTM is mainly concerned with topics detection rather than word representation and is trained with only 200 streaming tasks.

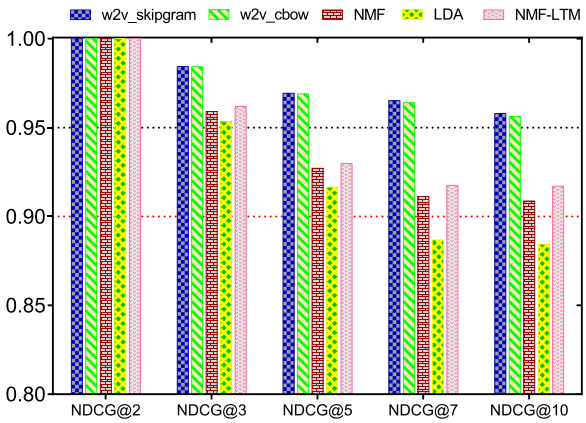


Fig. 8. Semantic relatedness by different learning methods on *SogouNews* corpus. (measured by NDCG@Υ, $\Upsilon = 2, 3, 5, 7, 10$).

In summary, the $\mathcal{KG}$ maintained during the lifelong learning process of NMF-LTM contains rich semantics, which to some extent explains why NMF-LTM can generate high quality topics.

8 Case Study on Real-World Big Data

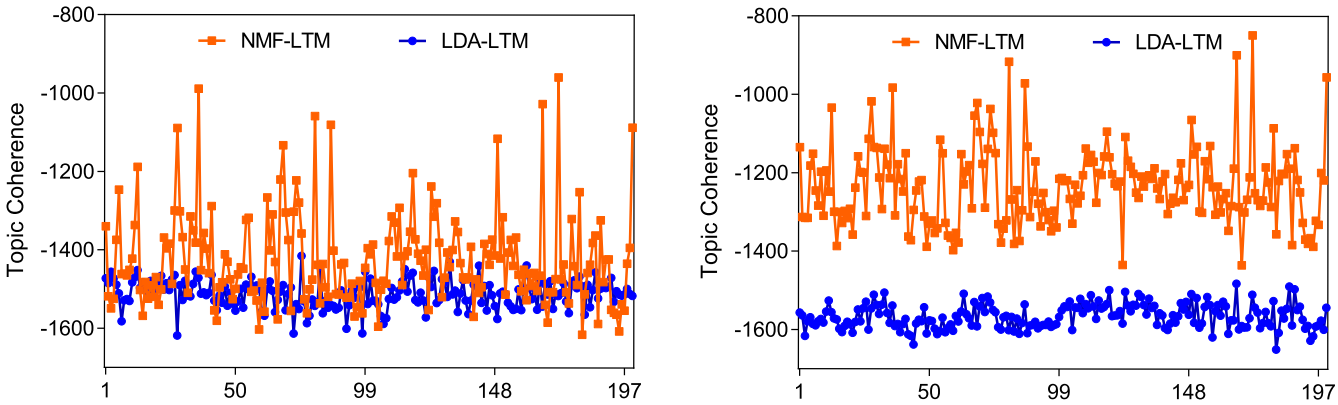
In this section, we demonstrate the ability of NMF-LTM in handling real-world large-scale corpora. Since the classic NMF and LDA models cannot finish running in a reasonable time period, in what follows, we only take NMF-LTM and LDA-LTM for the comparative study.

8.1 Experimental Setup

WanFang[12] and CNKI[13] are two most popular Chinese academic websites storing tremendous resources such as papers, theses, patents, standard documents and so forth. We use a Web-crawler to collect the papers publicly available in the period from 1998 to 2016, with a wide range from medicine, literature, to science and technology. We finally acquire over 12 million papers with titles, authors, abstracts and keywords information.

We treat each title as one document, and thus obtain a short-text corpus of over 12 million samples for topic modeling. We use the NLPIR tool (ANSJ) to split the Chinese titles into sequential terms, retain nominal phrases as candidate

9. https://code.google.com/archive/p/word2vec/
10. http://download.csdn.net/detail/chjshan55/3462335
11. https://en.wikipedia.org/wiki/Discounted_cumulative_gain
12. http://www.wanfangdata.com.cn/
13. http://epub.cnki.net/kns/default.htm



(a) $K = 20$ (b) $K = 50$

Fig. 9. Comparison of topic coherence on 200 tasks of academic papers data.

Term ID /Topics (NMF-LTM v.s. LDA-LTM)	通信技术研究与应用 （R&D on communication technology）		结构生物学 （structural biology）		电力故障检测与维修 （power failure detection and maintenance）	
1	技术(technology)	技术(technology)	蛋白(protein)	载体(carrier)	分析(analysis)	分析(analysis)
2	应用(application)	研究(research)	结构(structure)	病毒(viruses)	故障(fault)	原因(cause)
3	研究(research)	应用(application)	功能(function)	重组(recombinant)	原因(cause)	故障(fault)
4	系统(system)	通信(telecommunication)	激酶(kinase)	基因(gene)	变压器(transformer)	措施(measures)
5	通信(telecommunication)	**铁路(railway)**	重组(recombinant)	蛋白(protein)	线路(circuit)	电力(power)
6	网络(network)	移动(mobile)	基因(gene)	细胞(cell)	维修(maintenance)	运行(function)
7	发展(development)	多媒体(multi-media)	作用(effect)	免疫(immune)	仪器(instrument)	装置(device)
8	信息(information)	**机械(mechianics)**	受体(receptor)	鉴定(appraisal)	井仪(resistivity log)	线路(circuit)
9	分析(analysis)	数据(data)	病毒(viruses)	**关系(relationship)**	对策(measures)	变电站(transformer station)
10	检测(detection)	**工艺(craftwork)**	调节(regulation)	**研究(research)**	性能(performance)	**自动化(automation)**

Fig. 10. Comparison of semantic cohesion of topics generated by two competitors. (1. Top 10 terms are listed in descending order with noisy ones in bold. 2. $K = 20$.).

topic words, and finally reach a vocabulary of over 0.25 million terms. The average length of these titles is as small as 4.28.

To facilitate lifelong modeling, we select the top 2 million samples as a new corpus, and divide it into 200 batches, each with 10000 titles as stream-line tasks. The two competitive models: NMF-LTM and LDA-LTM, are exactly the ones defined in Section 6.2. The numbers of topics are set as $K = 20, 50$ and 100, respectively, and we use the top 20 topic words in each topic for the evaluation of topic coherence. Note that we use a smaller corpus here such that the running time of LDA-LTM can be controlled in a reasonable level.

8.2 Results and Discussions

Fig. 9 shows the detailed topic coherence values by both NMF-LTM and LDA-LTM on every data chunk with different K settings, and Table 3 collects the total time expenditure for them. Note that "—" indicates that LDA-LTM is too time-consuming and is definitely out of the tolerant waiting time.

As can be seen from Fig. 9, NMF-LTM outperforms LDA-LTM on almost all the batches. We can also find from Table 3 that NMF-LTM generally spends much less time than LDA-LTM in the lifelong learning. To conclude, these results demonstrate the advantage of NMF-LTM over LDA-LTM in terms of both topics quality and time efficiency. In what follows, we set $K = 20$ to further explore the topic words, the scalability and the robustness of NMF-LTM.

We map the topics learned by NMF-LTM with the ones by LDA-LTM. That is, we first represent each topic by its top 20 key terms, and then compute the cosine similarity between any two topics from different sides. We finally select three topic pairs randomly from the top 10 pairs, as listed in Fig. 10. It is general that both NMF-LTM and LDA-LTM capture the topics nicely. However, if we take a closer look at the topical words, it is obvious that they are less cohesive in the LDA-LTM topics, and some noisy words have been highlighted in red, such as "railway", "mechanics" and "craftwork" for the topic "R&D in Communication Technology". This demonstrates that NMF-LTM indeed can find semantically more cohesive topics than LDA-LTM.

We also have interests in the scalability of the two lifelong models. To this end, we make use of the complete corpus of 12.45 million titles, and divide it into 1245 batches (to keep the batch size as 10 thousand) for lifelong learning. Table 4 shows the total running time on the sampled and complete corpora, respectively. As can be seen, NMF-LTM

TABLE 3
Runtime Comparison with Different Topic Numbers

Time Expenditure\Methods	NMF-LTM	LDA-LTM
$K = 20$	0.22 days	2.78 days
$K = 50$	1.29 days	10.09 days
$K = 100$	2.61 days	—

(200 batches).

TABLE 4
Runtime Comparison with Different Scaled Tasks

Time Expenditure \ Methods	NMF-LTM	LDA-LTM
Total time for 200 batches (2 mil.)	0.22 days	2.78 days
Total time for 1245 batches (12.45 mil.)	1.35 days	—

($K = 20$).

TABLE 5
Performance of NMF-LTM on Data Sets with Varied Sizes

NMF-LTM	2 mil.	4 mil.	6 mil.	8 mil.	10 mil.	12.45 mil.
Total time (days)	0.22	0.40	0.62	0.85	1.08	1.35
Average topic coherence	-1424.78	-1433.82	-1430.13	-1416.83	-1419.27	-1425.68
Standard deviation	119.59	112.27	114.34	121.05	120.39	118.51

$(K = 20)$.

only needs 0.22 days for lifelong learning of 2 million titles whereas LDA-LTM needs 2.78 days! For the complete corpus, LDA-LTM even cannot deliver results in tolerable waiting time, while NMF-LTM consumes 1.35 days only! This well demonstrates the superiority of NMF-LTM to LDA-LTM in dealing with scalability issues.

To further verify the robustness of NMF-LTM in big data modeling, we run it on a series of sampled corpora with the sample size increased from 2 million to 12 million. Note that we divide these corpora similarly into various batches such that each batch contains 10 thousand titles. Table 5 gives the learning results. As can be seen, the running time of NMF-LTM is roughly linear to the corpus size, and the rather stable values of the average and standard deviation of topic coherence indicate the robustness of NMF-LTM.

In a nutshell, NMF-LTM exhibits evident advantage over LDA-LTM in terms of both topic quality and time efficiency when facing with large-scale real-world datasets. This makes NMF-LTM a great potential for topic modeling of big data.

9 Conclusions

Lifelong topic model is an emerging research hotspot aiming at imitating human's never-ending learning behavior to gain better performances in future practices. In this paper, we proposed a novel lifelong topic model: NMF-LTM, which differs from LDA based methods by employing a non-negative matrix factorization framework and taking word-affinity knowledge and other important priors as regulations. Extensive experiments on public corpora demonstrated the superior performances of NMF-LTM to some state-of-the-arts in terms of topic coherence. In particular, NMF-LTM shows its human-like learning behaviors and the great potential for topic modeling of big data. In the near future, we plan to investigate the task arrangement strategy in NFM-LTM so as to improve its ability to adapt to new tasks more quickly.

Acknowledgments

Junjie Wu was partially supported by the National Natural Science Foundation of China (NSFC) (71725002, 71531001, U1636210, 71471009, 71490723). Yong Chen was funded by CSC (China Scholarship Council) as a visiting Ph.D. student in Birkbeck and UCL from January 2018 to January 2019. This work was also sponsored in part by the National Key Research and Development Program of China under grant No. 2017YFB1400200. The authors also thank the Network Information Center in Beihang University for providing high-performance servers.

References

[1] D. L. Silver, "Machine lifelong learning: Challenges and benefits for artificial general intelligence," in *Proc. Int. Conf. Artif. General Intell.*, 2011, pp. 370–375.

[2] D. L. Silver, Q. Yang, and L. Li, "Lifelong machine learning systems: Beyond learning algorithms," in *Proc. AAAI Spring Symp.: Lifelong Mach. Learn.*, 2013, pp. 49–55.

[3] T. M. Mitchell, W. W. Cohen, et al., "Never-ending learning," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2302–2310.

[4] Z. Chen and B. Liu, *Lifelong Machine Learning: Synthesis Lectures on Artificial Intelligence and Machine Learning*, San Rafael, CA, USA: Morgan and Claypool Publishers, 2016, pp. 1–127.

[5] B. Liu, "Lifelong machine learning: A paradigm for continuous learning," *Frontiers Comput. Sci.*, vol. 11, pp. 359–361, 2017.

[6] Z. Chen and B. Liu, "Topic modeling using topics from many domains, lifelong learning and big data," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn.*, 2014, pp. 703–711.

[7] Z. Chen, "Lifelong machine learning for topic modeling and beyond," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics: Student Res. Workshop*, 2015, pp. 133–139.

[8] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka Jr., and T. M. Mitchell, "Coupled semi-supervised learning for information extraction," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 101–110.

[9] Z. Chen, N. Ma, and B. Liu, "Lifelong learning for sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics. 7th Int. Joint Conf. Natural Language Process.*, 2015, pp. 750–756.

[10] S. Wang, Z. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 167–176.

[11] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, 1999, pp. 50–57.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, 2003, pp. 993–1022.

[13] M. Khan, M. Durrani, S. Khalid, and F. Aziz, "Lifelong aspect extraction from big data: Knowledge engineering," *Complex Adaptive Syst. Model.*, vol. 4, pp. 1–15, 2016.

[14] G. B. Sorkin, "The power of choice in a generalized polya urn model," in *Proc. Int. Workshop Approximation Algorithms Combinatorial Optimization Int. Workshop Randomization Approximation Techn. Comput. Sci.*, 2008, pp. 571–583.

[15] F. Caron, M. Davy, and A. Doucet, "Generalized polya URN for time-varying dirichlet process mixtures," in *Proc. 23rd Conf. Uncertainty Artif. Intell.*, 2007, pp. 33–40.

[16] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robot. Auton. Syst.*, vol. 15, no. 1-2, pp. 25–46, 1995.

[17] S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations Trends Mach. Learn.*, vol. 4, pp. 107–194, 2012.

[18] C. Tekin and M. Liu, "Online learning methods for networking," *Foundations Trends Netw.*, vol. 8, pp. 281–409, 2015.

[19] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via dirichlet forest priors," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 25–32.

[20] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Discovering coherent topics using general knowledge," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 209–218.

[21] Z. Chen, A. Mukherjee, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting domain knowledge in aspect extraction," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2013, pp. 1655–1667.

[22] Z. Chen, A. Mukherjee, and B. Liu, "Aspect extraction with automated prior knowledge learning," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 347–358.

[23] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[24] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120.

[25] C. Wang, D. M. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Proc. 24th Conf. Uncertainty Artif. Intell.*, 2008, pp. 579–586.

[26] A. Bhadury, J. Chen, J. Zhu, and S. Liu, "Scaling up dynamic topic models," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 381–390.
[27] A. Saha and V. Sindhwani, "Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 693–702.
[28] C. K. Vaca, A. Mantrach, A. Jaimes, and M. Saerens, "A time-based collective factorization for topic discovery and monitoring in news," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 527–538.
[29] Y. Chen, H. Zhang, J. Wu, X. Wang, R. Liu, and M. Lin, "Modeling emerging, evolving and fading topics using dynamic soft orthogonal NMF with sparse representation," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 61–70.
[30] X. Chen, "Learning with sparsity: Structures, optimization and applications," CMU PhD Thesis, Mach. Learn. Dept., Carnegie Mellon Univ., 2013, pp. 1–187.
[31] Z. Xu, X. Chang, and F. Xu, "L-1/2 regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.
[32] N. D. Lawrence and M. A. Girolami, "Fast, exact model selection and permutation testing for l2-regularized logistic regression," in *Proc. Int. Conf. Artif. Intell. Statistics*, 2012, pp. 246–254.
[33] P. Xie, "Learning compact and effective distance metrics with diversity regularization," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2015, pp. 610–624.
[34] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. 14th Int. Conf. Neural Inf. Process. Syst.: Natural Synthetic*, 2001, pp. 585–591.
[35] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Proc. Advances Neural Inf. Process. Syst.*, 2001, pp. 556–562.
[36] Z. Li, X. Wu, and H. Peng, "Nonnegative matrix factorization on orthogonal subspace," *Pattern Recognit. Lett.*, vol. 31, pp. 905–911, 2010.
[37] X. H. Phan, M. L. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text and web with hidden topics from large-scale data collections," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 91–100.
[38] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 262–272.
[39] L. Li and Q. Yang, "Lifelong machine learning test," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 1–2.
[40] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.
[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013, http://arxiv.org/abs/1301.3781
[42] Y. Chen, H. Zhang, Y. Zuo, and D. Wang, "An improved regularized latent semantic indexing with L 1/2 regularization and non-negative constraints," in *Proc. IEEE 16th Int. Conf. Comput. Sci. Eng.*, 2013, pp. 1075–1082.

Yong Chen received the BEng degree in computer science and technology from the Civil Aviation Flight University of China, in 2011, and the MEng degree in computer science and engineering from Beihang University, in 2014. Now, he is currently working toward the PhD degree in the State Key Lab of Software Development Environment, Department of Computer Science and Engineering, Beihang University, Beijing 100191, P.R. China. His research interests include machine learning, data mining, and big data.

Junjie Wu received the PhD degree in management science and engineering from Tsinghua University. He is currently a full professor in the Information Systems Department and the director of the Research Center for Data Intelligence of Beihang University. His general area of research is data mining and complex networks. He is the recipient of the NSFC Distinguished Young Scholars Award and MOE Changjiang Young Scholars Award in China.

Jianying Lin received the BEng degree in computer science and technology from Shantou University, Guangdong Province, in 2015. He is currently working toward the graduate degree in the State Key Lab of Software Development Environment, Department of Computer Science and Engineering, Beihang University, Beijing, P.R. China. His general research interests include data mining, machine learning and natural language processing, especially text summarization, topic models, and information retrieval.

Rui Liu received the MEng and PhD degrees in computer science and engineering from Beihang University, in 1995 and 2011, respectively. He is an associate professor in the School of Computer Science and Engineering, Beihang University, Beijing, P.R. China. He had been working at the University of Pittsburgh, USA, from 2008 to 2009 as a guest researcher. His main research interests include e-Science, data management and data mining, and web information retrieval.

Hui Zhang received the MEng and PhD degrees in computer science and engineering from Beihang University, in 1994 and 2009, respectively. He is a professor in the School of Computer Science and Engineering, Beihang University, Beijing, P.R. China. He worked at the University of Chicago and Argonne National Laboratory, USA, from 2007 to 2008 as a guest researcher. His main research interests include e-Science, data management and data mining, web information retrieval, and cloud computing.

Zhiwen Ye received the BEng degree in computer science and technology from Zhengzhou University, Henan Province, in 2017. Now, he is working toward the master's degree in the State Key Lab of Software Development Environment of the School of Computer Science and Engineering, Beihang University, Beijing, P.R. China. His main research interests include machine learning, data mining, and big data, especially web information retrieval and search engine optimization for large-scale S&T resources (http://www.kejso.com/).