An Online Topic Modeling Framework with Topics Automatically Labeled

Fenglei Jin [1] **Cuiyun Gao** [1] **Michael R. Lyu** [1]

Abstract

In this paper, we propose a novel online topic tracking framework, named IEDL, for tracking the topic changes related to deep learning techniques on Stack Exchange and automatically interpreting each identified topic. The proposed framework combines the prior topic distributions in a time window during inferring the topics in current time slice, and introduces a new ranking scheme to select most representative phrases and sentences for the inferred topics in each time slice. Experiments on 7,076 Stack Exchange posts show the effectiveness of IEDL in tracking topic changes and labeling topics.

1. Introduction

Recent advances in deep learning promote the innovation of many intelligent systems and applications such as autonomous driving and image recognition. Tracking the changes of focus for deep learning engineers and researchers is helpful to identify current emerging deep learning-related topics. In this work, we choose Stack Exchange to collect experimental dataset due to its popularity among the developers and researchers (Huang et al., 2018).

Previous topic tracking approaches (Blei & Lafferty, 2006; AlSumait et al., 2008; He et al., 2013) are mainly based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003). For example, the work (AlSumait et al., 2008) proposes an Online Latent Dirichlet Allocation (OLDA) model to capture the evolution of topics, where only the topic distribution of documents in the prior one time slice is considered for inferring the topics in current time slice. In (He et al., 2013), the authors focus on modeling sentiment and topic changes synchronously, and the topics in all the prior time slices are involved during inferring the current topic distribution. In (Espinoza et al., 2018), the proposed approach also selects sentiment words for each topic based on Dynamic Topic Model (DTM) (Blei & Lafferty, 2006).

Inspired by the recent work (Gao et al., 2018), where an adaptively online latent Dirichlet allocation approach, named IDEA, is introduced to track user opinions in user feedback, and outperforms the OLDA approach (AlSumait et al., 2008), we propose a new framework **IEDL** for **I**dentifying **E**merging **D**eep **L**earning-related topics. The difference between IDEA and our approach lies in the combination styles of the prior topic distributions. In IDEA, the similarities between the topics in previous time slices and those in the previous one time slice are taken into account for inferring the topics in current time slice, while we introduce an exponential decay function in a time window. Besides, we propose a novel topic labeling approach based on the unique characteristics of Stack Exchange posts.

The experimental results on 7,076 Stack Exchange posts verify the effectiveness of IEDL in detecting topic changes and topic labeling.

The contributions of our paper are elaborated as below.

- We propose a framework called IEDL to automatically track topic changes and identify emerging topics from deep learning-related posts in Q&A forum effectively.
- We propose a novel topic interpretation method, which improve the topic coherence dramatically.
- We visualize the variations of the captured (emerging) topics along with time slices, with the emerging ones highlighted.

2. Methodology of IEDL

IEDL mainly contains two parts: Emerging topic detection and automatic topic interpretation.

2.1. Emerging Topic Detection

In this section, we aim to detect the emerging topics of current time slice by considering the topics in previous time

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China. Correspondence to: Fenglei Jin <fengleijin@gmail.com>, Cuiyun Gao <cygao@cse.cuhk.edu.hk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

slices. We first introduce how we use online topic modeling to capture the topic evolutions with time going by. Then we present how we discover the emerging topics (anomaly topics).

2.1.1. Online Topic Modeling

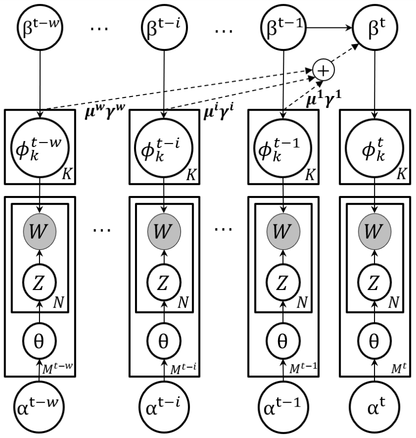


Figure 1. Our online topic modeling approach.

The preprocessed posts are divided by time, denoted as $M = \{M^1, M^1, ..., M^t...\}$ (where t indicates t-th month in our experiment), and each post is treated as one document. The prior distributions over document-topic (α) and topic-word distributions (β) are defined initially. K represents the number of the topics, while ϕ_k^t is the probability distribution vector for the k-th topic over all the input posts. We also introduce a predefined parameter - window size w, which refers to the number of previous time slices to be considered for analyzing the topic distributions of the current time slice. The overview of the model is shown in Figure1.

We adaptively integrate the topic distributions of the previous w time slices, denoted as $\{\phi^{t-1}, ..., \phi^{t-i}, ..., \phi^{t-w}\}$, for generating the prior β^t of the t-th time slice. Since the popularity of a topic always lasts for a time period, compared to IDEA (Gao et al., 2018) which only considers the similarity between topic distributions, we think the topics discussed last month are more related to current topics compared to those mentioned several months before. Therefore, an exponential decay factor μ is added, multiplying the similarity between topics γ to determine the influence. And now the adaptive integration refers to sum of the topic distributions of different time slices with different weights γ^i and μ^i:

$$\beta_k^t = \sum_{i=1}^{w} \mu^i \gamma_k^i \phi_k^{t-i} \tag{1}$$

where i denotes the i-th previous time slice $(1 \leq i \leq w)$. We denote the weight γ^i as the similarity of topic distributions between the $(t-i)$-th time slice and the $(t-1)$-th time slice, which is calculated by the softmax function:

$$\gamma_k^i = \frac{\exp(\phi_k^{t-i} \cdot \beta_k^{t-1})}{\sum_{j=1}^{w} \phi_k^{t-j} \cdot \beta_k^{t-1}}, \tag{2}$$

where the dot product $(\phi_k^{t-i} \cdot \beta_k^{t-1})$ computes the similarity between the topic distribution ϕ_k^{t-i} and the prior of the $(t-1)$-th time slice β_k^{t-1}. μ^i is calculated by a simple exponential decay function:

$$\mu^i = \exp(-\lambda i), \tag{3}$$

where i means the i-th time slice before the current, and λ is a predefined exponential decay coefficient.

2.1.2. Anomaly Discovery

Based on the topic distribution captured by online topic model, we regard anomaly topics, which present obvious distinctions compared to those of the previous time slices, as emerging topics. To calculate the distinction of the k-th topics between two successive time slices, we implement the classic Jensen-Shannon (JS) divergence[1]. If we take ϕ_k^t and ϕ_k^{t-1} as an example:

$$\begin{aligned} D_{JS}(\phi_k^t || \phi_k^{t-1}) = \frac{1}{2} D_{KL}(\phi_k^t || M) + \\ \frac{1}{2} D_{KL}(\phi_k^{t-1} || M), \end{aligned} \tag{4}$$

where $M = \frac{1}{2}(\phi_k^t + \phi_k^{t-1})$.And the Kullback-Leibler (KL) divergence D_{KL} is utilized to measure the difference from one probability distribution P to another Q:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}, \tag{5}$$

where $P(i)$ is the i-th item in P. The higher the JS divergence is, the larger distinction the two topic distributions have. To find anomaly topics, we set a threshold by leveraging a typical outlier detection method (Rousseeuw & Hubert, 2011). For each time slice, the topics with divergences higher than the threshold are regarded as anomaly topics.

2.2. Automatic Topic Interpretation

The dataset we use are questions asked in Stack Exchange, which have two significant attributes "votes" and "views". Users can manually click the "like" button or the "dislike" button to show their preference, while high "votes" represents this is a valuable question. And "views" shows how many users or tourists have visited this page, which refers to the popularity of this post. To make good use of these

[1] https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence

two attributes, we develop a novel method for deep learning related posts interpretation. We denote it as Quality Score:

$$SCORE_{qua}(l) = \exp(-\frac{1}{\ln(v_l+1)\ln(r_l+1)} - \eta \cdot \frac{1}{\ln(h_l+1)}) \quad (6)$$

where l is the post, and v_l, r_l, h_l are the votes, views, and length of the post respectively. The attributes are adding 1 in case of they are 0. If a post has both high votes and views, it is more likely to be a good post. Also, length slightly influence the score by the predefined factor η, since long posts may contain more information. Therefore, the motivation of this Quality Score is to select questions with both high votes and views with longer length.

3. Experiment and Result

In this section, we introduce how we preprocess the dataset, and the performance of our IEDL model measured by topic distribution classification precision and topic coherence.

3.1. Data Analysis

The 7,076 deep learning-related posts we used are publicly released by Stack Exchange[2]. To evaluate the topics inferred by our proposed topic model, we also manually labeled 507 posts into six categories for classification: Image, NLP, Game-ai, Self-driving, Programming-languages, and Reinforcement-learning. The labels are determined based on the tags provided by Stack Exchange and to maximize their distinguishability.

3.1.1. Word Formatting

We first convert all words into lowercase, and then perform lemmatization to change each word into its original form. We then replace some segments with general symbols, like converting websites to "<url>" and so on.

3.1.2. Phrase Extraction

Since some words have specific meanings only in phrases and we need them to interpreting topics, phrases (mainly referring to two consecutive words in our paper, and the words in each phrase are connected with "_") are extracted in the preprocessing step and trained along with all the other words. We want the topic labels in phrases to be meaningful and comprehensible, therefore, a typical phrase extraction method based on PMI (Pointwise Mutual Information)[3], which is effective in identifying meaningful phrases based on co-occurrence frequencies, is used:

$$PMI(w_i, w_j) = \log \frac{p(w_i w_j)}{p(w_i)p(w_j)} \quad (7)$$

where $p(w_i w_j)$ refers to the co-occurrence probability of the phrase $w_i w_j$ and $p(w_i)$ and $p(w_j)$ indicates the probability of the word w_i and w_j in the whole post documents. High PMI values indicate that it is more likely for the combination of the two words to be a meaningful phrase. We experimentally set a threshold for PMI, and phrases with higher PMIs are extracted.

3.1.3. Filtering

This step aims to eliminate non-meaningful words, such as emotional words (e.g., "nice" and "bad"), abbreviations (e.g., "btw"), and useless words (e.g., "something"). We use the predefined stop words provided by NLTK[4], and all words in the stop word list are filtered out. Finally, all remaining words and extracted phrases are fed into the model.

3.2. Classification Accuracy

To test the quality of the extracted topics, we use the topic distribution of each post as features, and classify the 507 labeled posts by SVM. The results show that our proposed model outperforms the baseline model IDEA (Gao et al., 2018) by 5% for average precision.

Table 1. Classification result

Category	Model	Precision	Recall	F1
Image	IDEA	0.89	**0.73**	**0.80**
	IEDL	**1.00**	0.64	0.78
NLP	IDEA	0.68	0.76	0.72
	IEDL	**0.73**	**0.94**	**0.82**
Game-ai	IDEA	0.83	0.94	0.88
	IEDL	**0.83**	**0.97**	**0.90**
Self-driving	IDEA	0.94	0.89	0.91
	IEDL	**1.00**	**0.94**	**0.97**
Programming -language	IDEA	**0.92**	0.73	0.81
	IEDL	0.86	**0.86**	**0.86**
Reinforcement -learning	IDEA	0.86	**0.86**	**0.86**
	IEDL	**1.00**	0.62	0.76

3.3. Topic Coherence

Table 2. Topic coherence of different approaches.

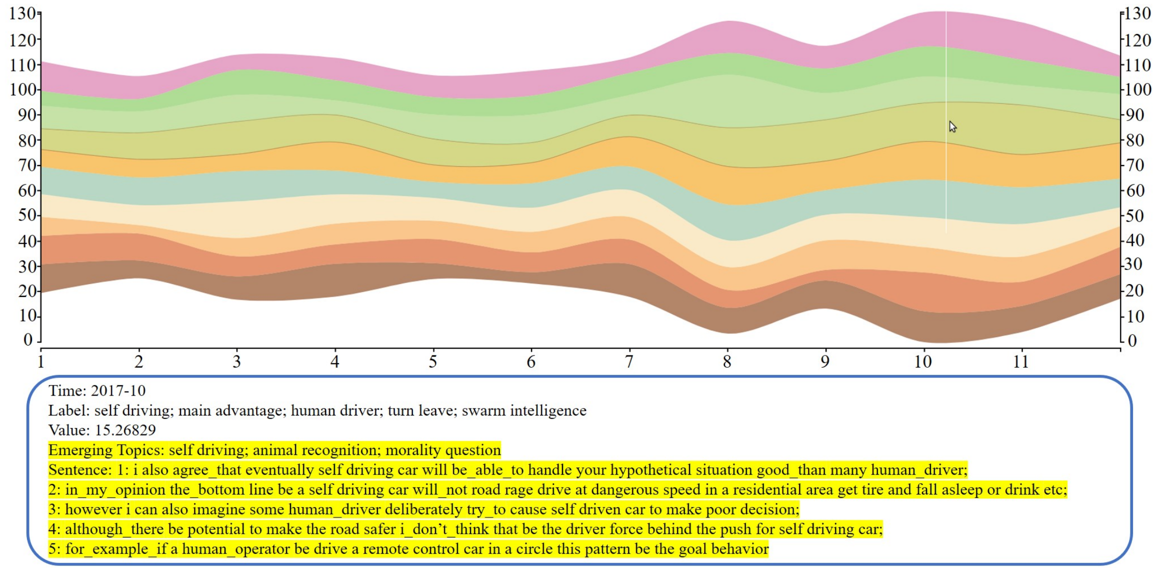
OLDA	IDEA	IDEA+Quality Score	IEDL
0.133	0.166	0.217	**0.222**

Topic coherence score (Lau & Baldwin, 2016) is another way to measure the performance of models by detecting the coherence between extracted words or phrases assigned to each topic. The method we use is an extension of PMI,

[2] https://archive.org/download/stackexchange

[3] https://en.wikipedia.org/wiki/Pointwise_mutual_information

[4] http://www.nltk.org/



Time: 2017-10
Label: self driving; main advantage; human driver; turn leave; swarm intelligence
Value: 15.26829
Emerging Topics: self driving; animal recognition; morality question
Sentence: 1: i also agree_that eventually self driving car will be_able_to handle your hypothetical situation good_than many human_driver;
2: in_my_opinion the_bottom line be a self driving car will_not road rage drive at dangerous speed in a residential area get tire and fall asleep or drink etc;
3: however i can also imagine some human_driver deliberately try_to cause self driven car to make poor decision;
4: although_there be potential to make the road safer i_don't_think that be the driver force behind the push for self driving car;
5: for_example_if a human_operator be drive a remote control car in a circle this pattern be the goal behavior

Figure 3. Visualization of topic changes based on ThemeRiver (Havre et al., 2000). Texts highlighted in yellow are the emerging topics in the corresponding month (Oct. 2017) where the mouse is pointing at.

where t refers to a topic, and N is the number of words:

$$OC\_Auto\_PMI(t) = \sum_{j=2}^{N}\sum_{i=1}^{j-1} \log \frac{p(w_i w_j)}{p(w_i)p(w_j)}, \qquad (8)$$

We feed the whole 7,076 dataset into our IEDL model and compare the topic coherence score to IDEA (Gao et al., 2018). The result shows IEDL improves the topic coherence score by 33.7%. The topic coherence scores with error bars are shown in Figure 2.

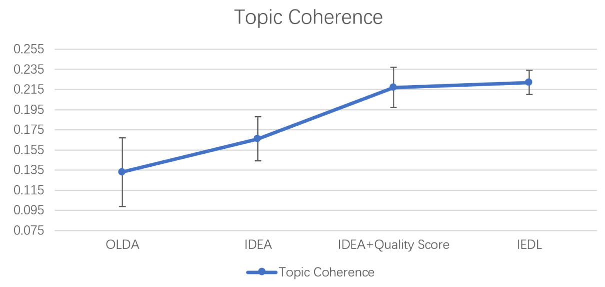


Figure 2. Topic coherence with error bars (standard error).

To further elaborate on the coherence between topics and extracted phrases, we compared the generated phrases for topics "NLP" and "Image" respectively. The result shows IEDL can generate more coherent and meaningful phrases for each topic.

Table 3. Phrases generated by IDEA and IEDL for topics "NLP" and "Image" respectively. Red underlined fonts highlight the phrases that are not closely related to the topic.

NLP		Image	
IDEA	IEDL	IDEA	IEDL
solution space	word vector	cnn model	convolutional network
information science	word embedding	previous layer	pixel value
real environment	feature extraction	specific task	capsule network

4. Visualization

In this part, we visualize the the evolution deep learning topics along with time flow for better understanding. As shown in Figure 3, all the posts constitute one river and each branch of the river indicates one topic. By moving the mouse over one topic, one can track detailed topic changes along with time slices (months in our experiment), where the emerging issues are highlighted.

The topics with wider branches are of greater concern to developers, where the width of the k-th branch in the t-th version is defined as:

$$width_k^t = \sum_a \log Count(a) \times SCORE_{qua}(l_a), \qquad (9)$$

where $Count(a)$ is the count of the phrase label a in the post collection of the t-th version, and $Score_{qua}(l_a)$ denotes the quality score of l_a, which is the post refers to the phrase label a.

We visualize topic changes from January to December 2017.

As shown in Figure 3, our IEDL finds an emerging topic about self-driving, which is not detected by IDEA (Gao et al., 2018). We double check the dataset and find that, compared to only one post from July to September, there are eight posts about self-driving in October (may be caused by a new release of electric semi-truck of Tesla), which further proves the effectiveness of our model in detecting emerging topics.

5. Conclusion and Future Work

Timely and effectively detecting deep learning topics is crucial for developers to capture the trend. We propose IEDL, a novel framework for automatically identifying emerging topics from posts in Q&A forums. The experiment results show IEDL improves the quality of topic distribution and topic coherence greatly. In the future, we will refine IEDL to be capable of defining the topic number automatically, and utilize other information like comments, accepted answers to further improve the performance.

References

AlSumait, L., Barbará, D., and Domeniconi, C. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pp. 3–12, 2008.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, pp. 113–120, 2006.

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

Espinoza, I., Mendoza, M., Ortega, P., Rivera, D., and Weiss, F. Viscovery: Trend tracking in opinion forums based on dynamic topic models. *CoRR*, abs/1805.00457, 2018.

Gao, C., Zeng, J., Lyu, M. R., and King, I. Online app review analysis for identifying emerging issues. In *Proceedings of the 40th International Conference on Software Engineering, ICSE 2018, Gothenburg, Sweden, May 27 - June 03, 2018*, pp. 48–58, 2018.

Havre, S., Hetzler, E. G., and Nowell, L. T. Themeriver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization 2000 (INFOVIS'00), Salt Lake City, Utah, USA, October 9-10, 2000.*, pp. 115–123, 2000.

He, Y., Lin, C., Gao, W., and Wong, K. Dynamic joint sentiment-topic model. *ACM TIST*, 5(1):6:1–6:21, 2013.

Huang, Q., Xia, X., Xing, Z., Lo, D., and Wang, X. API method recommendation without worrying about the task-api knowledge gap. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, Montpellier, France, September 3-7, 2018*, pp. 293–304, 2018.

Lau, J. H. and Baldwin, T. The sensitivity of topic coherence evaluation to topic cardinality. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 483–487, 2016.

Rousseeuw, P. J. and Hubert, M. Robust statistics for outlier detection. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 1(1):73–79, 2011.