Incremental Topic Modeling for Scientific Trend Topics Extraction

Anonymous ACL submission

Abstract

Caused by the exponential growth of scientific research, the number of scientific publications and reports, one of the most urgent and challenging tasks now is the early detection of trending topics. In this paper, we investigate recent topic modeling approaches to accurately extract trending topics at an early stage. The incremental training technique is suggested so that the model can operate on data in real-time. For validation, we propose a novel dataset that contains a collection of early-stage articles and a set of key collocations for each trend. The proposed metric estimates the delay in days when determining the trend, and the developed matching method suffices to calculate it automatically. The conducted experiments demonstrate that the topic model with regularization, namely ARTM, is superior to the base PLSA model. Apart from that, the best ARTM-based model is able to extract most of the labeled trends during the first year of their evolution.

1 Introduction

The rapid growth of the number of scientific publications, journals, and conferences makes it effortful to reconstruct a full view of specific subject areas. Nowadays, people have to keep track of numerous emerging approaches, for which the global scientific importance is not always explicit at the first moment. Triggered by this fact, more attention is paid to the methods that solve the research trend identification task (Ho et al., 2014; Rotolo et al., 2015; Prabhakaran et al., 2016; Färber and Jatowt, 2019; Uban et al., 2021).

In this study, we consider the task of trend-like topic detection in real-time. The resulting topics should satisfy the following conditions. Firstly, they should contain as many trending topics as possible. We utilize the definition of trend proposed by Kontostathis et al. (2004), where the emerging trend is a topic, interest to which was strongly increased in a particular time interval. Secondly, trend-like topics should be identified as early as possible by the time they appear. Finally, each topic should be semantically homogeneous and impartible.

In our experiments we extract trend topics from Artificial Intelligent (AI) field, but the proposed approach can be as well applied to other scientific fields in the future. The systems solving the defined tasks can be used as a base for various applications: early trend detection, visualization and analysis of topic emergence, etc.

In order for the final model to operate in real-time, we suggest incremental training. At each timestamp, we aim to generate new topics as distant as possible from existing ones, which is not implied a priori in some topic models. Apart from that, most current topic modeling approaches have issues associated with the dilution of topics and terms, and the decorrelation of terms. To overcome these and other similar problems, we apply a topic model with additive regularization, namely ARTM (Vorontsov and Potapenko, 2015).

Despite active research in the field, there is no single quality metric for comparing trend detection models. Thus, we propose our intuitional metric in accordance with the assigned task. Moreover, we propose an expertly assembled dataset for comparison, which we publish in the public domain.

Our contributions can be summarized as the following:

- We propose the usage of the topic model with additive regularization to overcome the disadvantages of current trend detection approaches.

- We suggest the incremental mechanism of ARTM training utilizing trend keywords to detect trend topics in real-time.

- We propose a novel public dataset to validate trend topic detection approaches.

2 Related Work

Trend detection systems generally can be divided into two groups: semi-auto and auto approaches. We investigate only approaches that do not require human interaction.

Generally, automatic detection of trends involves two stages: topic detection (or identification) and topic evolution (with emerging trend classification).

The first stage is needed to construct the set of topics from which the trends will be selected. The following types of approaches can be distinguished: statistical, knowledge-based and hybrid. Statistical approaches operate only when provided with text context. Various models have been already investigated in this direction: topic modeling (Prabhakaran et al., 2016; Uban et al., 2021), clustering approaches (Mei and Zhai, 2005; Behpour et al., 2021), etc. Apart from that, some models utilize information from knowledge bases like the web (Roy et al., 2002) or citation graphs (Erten et al., 2004; Chang and Blei, 2010). Hybrid approaches (Jo et al., 2007; He et al., 2009) combine term-based topic detection and co-citation/co-authorship graph analysis.

Due to the specifics of our collection, namely the length of the full texts of articles, some of the neural approaches to the topic modeling (like BERTopic) are not directly applicable.

Topic evolution is utilized to consider topic emergence in time. Here, some approaches use custom metrics based on the topic characteristics (Ho et al., 2014; Prabhakaran et al., 2016; Färber and Jatowt, 2019; Behpour et al., 2021). Another category of approaches considers citations-based metrics. In this way, Le et al. (2006) proposed to use various temporal citation-based features to evaluate the growth in interest and utility of topics over time.

To track topic emergence in real time, we investigate incremental topic models. Some researchers (Canini et al., 2009; Hoffman et al., 2010) suggested online techniques for LDA. Nevertheless, due to the qualitative limitations of LDA-based approaches, we use the ARTM model (Vorontsov and Potapenko, 2015) and propose a method of its incremental training. Our incremental mechanism is based on trend keywords detection. Similar to our approach, Färber and Jatowt (2019) proposed a method to estimate the impact index of keywords but did not integrate it in the trend detection pipeline.

3 Trend Topic Detection

3.1 Approach

In this paper we consider the task of trending topic detection in real-time. In order to be able to experiment not only with models based on matrix factorization but also other popular approaches (e.g. clustering-based), we suggest to reduce the topic detection task to a search problem. Broadly speaking, we have a query for each topic (a topic name), and the goal is to get relevant lists of terms and documents associated with it. In our case, the queries are hidden. Thus, the system should return ranked lists of per-topic documents and words for each predefined timestamp.

To obtain real-time predictions, we suggest incremental training of the topic model, so we can only fine-tune the current model for each update timestamp, not retrain it from scratch.

Let D be a collection of documents and W be a dictionary of words (terms). After the last model update, a new collection of documents D' appears. The model should analyze a set of emerging words W' and update current topics T by adding new topics T' to it.

The incremental model solves two subtasks: choosing the number of new topics $|T'|$ and initializing new topics and adjusting them later.

Generally, topic modeling approaches operate with matrices Φ and Θ representing word-topic and topic-document distributions respectively. We suggest an incremental update for each of them. We consider block matrices Φ and Θ where each block is associated with a timestamp and a collection of documents at that time. At each update, we add new rows and columns to these matrices. Figure 1 shows their features in terms of sparsity. We have only zero values for the blocks showing the distributions of W' in T, as well as T in D. Then, we try to update the matrices so that new topics consist mainly of new words. Thus, the block showing the distribution of W in T' will be relatively sparse. The same is true for T and D'.

In order to determine the number of new topics for updating, we propose to use a criterion based on the size of the emerging trend vocabulary V. This vocabulary consists of terms that have become much more commonly used compared to the moment of the last update of the topic model.

Let $w \in W \cup W'$ be a word from the current corpus. At the current timestamp, this word is added to V if it appears in at least mindf documents

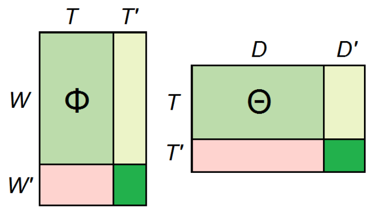

Figure 1: Incremental topic modeling. Zeros are marked by red, a more sparse matrix is marked with a lighter green.

and it satisfies the trend condition:

$$\frac{\text{tf}_{\text{new}} - \text{tf}_{\text{old}}}{\text{tf}_{\text{old}}} > \alpha \qquad (1)$$

Here, tf_{old} is the count of the occurrence of w in documents D, and tf_{new} is the count of the occurrence of w in $D \cup D'$. α is a regulation hyperparameter that sets the degree of increase in the occurrence of the words to classify them as trending.

Further, the number of topics is defined as

$$|T'| = |T_{\text{start}}| + \left\lfloor \frac{|V|}{\beta} \right\rfloor \qquad (2)$$

In (2), T_{start} determines the number of topics at the initial timestamp, $\beta \in \mathbb{N}$ limits the number of added topics, and $\lfloor \cdot \rfloor$ denotes an integer part.

To solve the recommendation task, we leverage scores from the Φ and Θ matrices to rank documents and words most appropriate for each topic.

3.2 Evaluation

At each iteration of the additional training of the incremental model, the search for the best topic for each trend is performed as the following.

Let D_{trend} and W_{trend} be the labeled sets of documents and words associated with the given trend respectively. At the output of the model, each topic is represented by two ranked lists denoted as D_{topic} and W_{topic}.

To perform matching, we calculate scores based on the Recall@k metric:

$$\text{XRecall@k} = \frac{|X_{\text{topic}}[\text{:k}] \cap X_{\text{trend}}|}{k} \qquad (3)$$

Here, $X[\text{:k}]$ denotes first k elements of the list X. X is replaced with W or D. We use two different values of k for documents and words, which are denoted as k_D and k_W.

We combine DRecall@k and WRecall@k scores to estimate the relevance of the selected topic to the selected trend. We consider the trend to be detected once it has been matched to some extracted topic.

Since our goal is to minimize time delay for the trend detection, the final quality metric is the average number of days that elapsed from the inception of a trend to its selection by the model. In our case, the inception date is the date of the earliest labeled publication referred to the trend.

4 Dataset

4.1 Data Sources

We used the part of Semantic Scholar Open Research Corpus as the main source of scientific publications. We considered only publications from 11 conferences that were selected based on data of top venues of Google Scholar (AI, Computational Linguistic, Computer Vision & Pattern Recognition sections were chosen) and h-index exceeding 100.

We enriched our dataset by adding information from the arXiv dataset[1], and updated years for some publications. We exclude the cases when the article was first published on the arXiv site, became available to the scientific community and only after some time appeared in the proceedings of some conference.

Eventually, our dataset contains the following attributes for each publication: the paper id on Semantic Scholar, the title, authors' ids, venue, ids of publications it refers to, ids of papers that refer to it, the date of publication on the arXiv and the date of the conference.

4.2 Markup

We used the reference graph from the Semantic Scholar dataset for markup. Initially, we generated 87 trends for "model", "method" and "task" types. Further, for each trend we expertly selected at least 10 relevant publications and more than 5 key collocations.

Thus, we collected a dataset with the following structure: trend name, trend type, a subset of papers related to the trend and trend keywords. The dataset is publicly available at the link http://....

5 Experiments

5.1 Implementation Details

We extracted collocations based on noun and verb phrases and used it as additional input in the topic

[1] https://www.kaggle.com/Cornell-University/arxiv

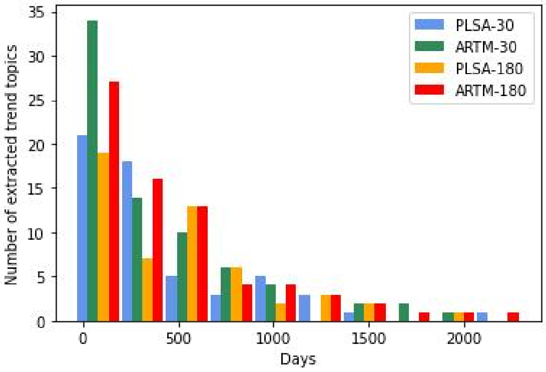

Figure 2: Histogram represents the number of extracted trend topics depending on delay from its start.

modeling approaches.

The open-source BigARTM library (Vorontsov et al., 2015) was used to train topic modeling approaches based on ARTM and PLSA. In the case of the ARTM model, we used the regularizer named Decorrelator Phi that contributes to the decorrelation of columns in the Φ matrix. The regularization coefficient τ was set to 0.2.

We conducted our experiments for two sequences of timestamps, updating every 30 days or 180 days. When updating the model, only those timestamps were used for which the emerging trend vocabulary was updated. In both cases, we started with $|T_{\text{start}}| = 50$ initial topics, mindf was set to 10 and α was set to 0.5. For the sequence with updating every 180 days we used $\beta = 150$ and for the 30-day-based sequence we used $\beta = 100$.

5.2 Topic Detection

We fitted two types of topic models, namely PLSA and ARTM, for two variants of sequences of timestamps (30 and 180 days between updates). Thus, we analyze results of four models denoted as PLSA-30, ARTM-30, PLSA-180 and ARTM-180.

We matched the extracted topics with the labeled trend topics using scores described in Section 3.2. We used thresholds DRecall@k > 0.1 and WRecall@k > 0.3 at each timestamp. Based on the matched trends, we calculated the delay between inception and extraction dates of each trend. Distributions of delays are demonstrated in Figure 2. The histogram illustrates that ARTM-180 is superior to PLSA-180 and even to PLSA-30 in the early detection task.

Table 1 shows the calculated statistics for the day delay metric. It can be seen that ARTM-30 achieved the highest scores in the trend topics extraction task. Moreover, both ARTM models extracted more than 70 trend topics while the PLSA-based models detected only about 50 of 87. Although ARTM-180 is not much superior PLSA-180 in terms of delays, it has extracted 20 more trends due to the regularization.

Statistic	**PLSA**		**ARTM**	
	180	**30**	**180**	**30**
mean	526	450	541	**428**
min	12	1	16	**0**
25%	153	123	160	**105**
50%	484	300	367	**238**
75%	702	**603**	666	637
max	1966	2326	2880	**2002**
# extracted	53	57	73	**74**

Table 1: Statistics of delays over all extracted trend topics for the chosen sequences of timestamps.

The quality is limited by several factors: the sizes of topics and their presence in the validation dataset (for instance, “em algorithm” and “pattern recognition” present quite weakly); the occurrence of keywords in articles (the keyword “gpt” usually appears in a paper only several times); the dataset quality and the quality of internal components of the approach such as the matching procedure.

6 Conclusion and Future Work

In this paper we investigated the topic modeling approaches to the scientific trend topics detection task. To make predictions in real-time, we proposed incremental training which consists of topic updating based on the current vocabulary of trend words. We used the topic model with the regularization, namely ARTM, to detect topics. We described the validation process and proposed a method for matching labeled trends and extracted topics.

The expertly labeled dataset was collected for experiments. It consists of 87 groups of AI articles (one group per trend) with corresponding keywords.

The evaluation demonstrated that ARTM outperforms PLSA almost in all cases. Besides, the best model is able to identify most of the trends during the first year of their existence.

There are possible directions for further research: tuning and improving the components of the current approach, as well as the investigation of the trend identification approaches.

References

Sahar Behpour, Mohammadmahdi Mohammadi, Mark V. Albert, Zinat S. Alam, Lingling Wang, and Ting Xiao. 2021. Automatic trend detection: Time-biased document clustering. *Knowledge-Based Systems*, 220:106907.

Kevin Canini, Lei Shi, and Thomas Griffiths. 2009. Online inference of topics with latent dirichlet allocation. *Journal of Machine Learning Research - Proceedings Track*, 5:65–72.

Jonathan Chang and David M. Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1).

C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee. 2004. Exploring the computing literature using temporal graph visualization. *Proceedings of SPIE - The International Society for Optical Engineering*, 5295:45–56. Visualization and Data Analysis 2004 ; Conference date: 19-01-2004 Through 20-01-2004.

Michael Färber and Adam Jatowt. 2019. Finding temporal trends of scientific concepts. In *BIR@ ECIR*, pages 132–139.

Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. 2009. Detecting topic evolution in scientific literature: How can citations help? In *ACM 18th International Conference on Information and Knowledge Management, CIKM 2009*, International Conference on Information and Knowledge Management, Proceedings, pages 957–966. ACM 18th International Conference on Information and Knowledge Management, CIKM 2009 ; Conference date: 02-11-2009 Through 06-11-2009.

Jonathan C. Ho, Ewe-Chai Saw, Louis Y.Y. Lu, and John S. Liu. 2014. Technological barriers and research trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Change*, 82(C):66–79.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. volume 23, pages 856–864.

Yookyung Jo, Carl Lagoze, and C. Lee Giles. 2007. Detecting research topics via the correlation between graphs and texts. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 370–379, New York, NY, USA. Association for Computing Machinery.

April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. 2004. *A Survey of Emerging Trend Detection in Textual Data Mining*, pages 185–224. Springer New York, New York, NY.

Minh-Hoang Le, Tu Bao Ho, and Yoshiteru Nakamori. 2006. Detecting emerging trends from scientific corpora.

Qiaozhu Mei and ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text - an exploration of temporal text mining. pages 198–207.

Vinodkumar Prabhakaran, William L Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180.

Daniele Rotolo, Diana Hicks, and Ben R. Martin. 2015. What is an emerging technology? *Research Policy*, 44(10):1827–1843.

Soma Roy, David Gevry, and William Pottenger. 2002. Methodologies for trend detection in textual data mining. 2.

Ana Sabina Uban, Cornelia Caragea, and Liviu P Dinu. 2021. Studying the evolution of scientific topics and their relationships. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1908–1922.

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. 2015. Bigartm: Open source library for regularized multimodal topic modeling of large collections. pages 370–381.

Konstantin Vorontsov and Anna Potapenko. 2015. Additive regularization of topic models. *Mach. Learn.*, 101(1–3):303–323.