Identify Emerging Trends In Machine Learning Using Published Literature

1st Yunwei Hu, 2nd Cheng Zhan, 3rd Jielin Xu, 4th Yu Zeng, 5th Junjie Yu, 6th Guo Yu, 7th Licheng Zhang

1. General Electric
2. Microsoft

Short Abstract:

We have developed a Machine Learning (ML) based system that analyzes the published literature to identify trends to help ML practitioners to understand emerging topics that will have a long-term impact.

The workflow automatically uncovers important research trends by periodically analyzing the arXiv publications and other open academic literature search engines. Our system retrieves papers published in the Machine Learning field on arXiv, and then crawls the SemanticScholar engine to acquire the citations information. Citations are widely used as the input to analyze the impact. An impact rating score was calculated for each paper by analyzing the citations of papers published in a window close to the date of a paper of interest. We chose only papers meeting certain impact score criteria for our next step.

The text of selected papers was pre-processed, whose steps are strongly related to the field where the paper is in, because the vocabulary, abbreviations, equations, and special characters used in the scientific literature are highly specialized.

We fed the pre-processed papers into the Topic Modeling models to identify the topics. The model is a hybrid of both the traditional Latent Dirichlet allocation(LDA)/ Non-negative matrix factorization (NMF) and BERT encoder. The BERT encoder allows us to use the pertained model to understand the context better in addition to the proven bag-of-words representations. The topics identified are grouped as themes that provide us the insight into how they have evolved and predict which may emerge as new breakthroughs.

Knowledge graph and citation network are also used to identify trends, we plan to expand our system to incorporate analysis from those methods in future work.

Index Terms:

Topic Modeling, Machine Learning, natural language processing, Trends

I. Motivation

As the field of Machine Learning (ML) is growing rapidly, it becomes ever more difficult to digest the incoming information and thereby identify emerging topics that will have a long-term impact, even for ML practitioners. We propose a ML based system that analyzes the published literature to identify trends.

II. Hypothesis

Previous work of identifying trends from scientific publications varies in terms of data-sets and methodologies. For example, recently, in [1], the authors manually summarized the topics of selected arXiv machine learning papers, and in [2], the authors used Latent Dirichlet allocation(LDA) to identify “hot topics” for covid-19 research.

Topic Modeling techniques, LDA and others, can substitute the manual summarizing process in [1] and the ideal workflow would require minimal human intervention for its periodical updates.

The arXiv system is an open-access online archive for scholarly manuscripts in the fields of science and technology, and serves as a platform for knowledge sharing. We choose arXiv because of its popularity among ML researchers, and ease of access to the dataset via Kaggle, but the methodology can be expanded to any other corpus/datasets.

III. Methods and Results

The workflow that we developed automatically uncovers important research trends by periodically analyzing the arXiv publications and other open academic literature search engines. Our system retrieves papers published in the Machine Learning field on arXiv, and then crawls the SemanticScholar engine to acquire the citations information. Citations are widely used as the input to analyze the impact. An impact **rating** score was calculated for each paper by analyzing the citations of papers published in a window close to the date of a paper of interest[4]. We chose only papers meeting certain impact score criteria for our next step.

The text of selected papers was **pre-processed**, and the steps are strongly related to the field where the paper is in, because the vocabulary, abbreviations, equations, and special characters used in the scientific literature are highly specialized.

We fed the pre-processed papers into the **Topic Modeling** models to identify the topics. The model is a hybrid of both the traditional Latent Dirichlet allocation(LDA) / Non-negative matrix factorization (NMF) and BERT encoder. The BERT encoder allows us to use the pertained model to understand the context better in addition to the proven bag-of-words representations. Similar approach was discussed in [3].

The topics identified are grouped as themes that provide us the insight into how they have evolved and predict which may emerge as new breakthroughs.

IV. Conclusion

We have developed a system to detect research trends in the ML field. It can be applied to other fields with minor modifications, and the discovered trends could be used to understand future evolution ranging from technological orientation, consumer end products, to labor markets.

In future work, we plan to expand our system to incorporate other techniques such as knowledge graph and citation networks.

Refs

1. S. Eger, C. Li, F. Netzer, I. Gurevych, Predicting Research Trends From Arxiv, arXiv:1903.0283, https://arxiv.org/abs/1903.02831
2. M. Dong, X. Cao, M. Liang, L. Li, G. Liu, H. Liang, Understand Research Hotspots Surrounding COVID-19 and Other Coronavirus Infections Using Topic Modeling, https://doi.org/10.1101/2020.03.26.20044164
3. F. Bianchi, S. Terragni, D. Hovy, Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence, arXiv:2004.03974
4. M. E. J. Newman, Prediction of highly cited papers, Europhys. Lett. 105, 28002 (2014), https://arxiv.org/abs/1310.8220