Deep Search Relevance Ranking in Practice

Linsey Pang
Salesforce, CA, USA
xpang@salesforce.com

Wei Liu
University of Technology Sydney, NSW, Australia
Wei.Liu@uts.edu.au

Keng-Hao Chang
Microsoft, Mountain View, CA, USA
kenchan@microsoft.com

Xue Li
Microsoft, Mountain View, CA, USA
xeli@microsoft.com

Moumita Bhattacharya
Netflix, Los Gatos, CA, USA
mbhattacharya@netflix.com

Xianjing Liu
Twitter, San Jose, CA, USA
xjliu.pku@gmail.com

Stephen Guo
Walmart Global Tech, Sunnyvale CA, USA
Stephen.Guo@walmart.com

ABSTRACT

Machine learning techniques for developing industry-scale search engines have long been a prominent part of most domains and their online products. Search relevance algorithms are key components of products across different fields, including e-commerce, streaming services, and social networks. In this tutorial, we give an introduction to such large-scale search ranking systems, specifically focusing on deep learning techniques in this area. The topics we cover are the following: (1) Overview of search ranking systems in practice, including classical and machine learning techniques; (2) Introduction to sequential and language models in the context of search ranking; and (3) Knowledge distillation approaches for this area. For each of the aforementioned sessions, we first give an introductory talk and then go over an hands-on tutorial to really hone in on the concepts. We cover fundamental concepts using demos, case studies, and hands-on examples, including the latest Deep Learning methods that have achieved state-of-the-art results in generating the most relevant search results. Moreover, we show example implementations of these methods in python, leveraging a variety of open-source machine-learning/deep-learning libraries as well as real industrial data or open-source data.

Key Words: Deep Learning, Search Relevance, Sequential Models, Attention, Transformer, Knowledge Distillation

ACM Reference Format:
Linsey Pang, Wei Liu, Keng-Hao Chang, Xue Li, Moumita Bhattacharya, Xianjing Liu, and Stephen Guo. 2022. Deep Search Relevance Ranking in Practice. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542632

Relevance ranking is a core problem of Information Retrieval (IR) which plays a fundamental role in many applications [4], such as web search engines. Given a query and a set of candidate text documents, a ranking function is used to determine the relevance degree of a document with respect to the query by generating a score. Early rank approaches focus on text matching between queries and web documents such as BM25 [7], vector space model [1], etc. However, with the tremendous growth of web information, an increasing number of queries in the format of natural language and more dimensional features including both temporal and spatial dimensions presents challenges to existing ranking solutions. In recent years, deep learning approach has shown great success in many machine learning ranking applications, including DSSM [3], CDSSM [8], DeepRank [6] etc.

In this tutorial, we provide an overview of search ranking in practice and demonstrate the various classical and popular ranking algorithms to help audience understand search relevance algorithms and their applications to the real-world. The tutorial outline is as follows:

Introduction to Search Relevance Ranking: In this session, we provide an overview of ranking problem in information retrieval [9]. A few early work of ranking functions are reviewed and the history of various models are briefly introduced. We choose some key algorithms to explain and demonstrate their ranking performance using real-world data. We cover several key performance offline metrics to evaluate ranking and online metrics are introduced. Our hands-on session cover classical ranking functions' implementation.

Attention based Models for Search Relevance: In this session, we give an overview of the evolution of sequence models and then covering the attention mechanism. We also give an introduction to the transformer architecture and how some of these can be leveraged in the context of search ranking systems. The details are: (1) we cover what Sequence models (such as RNN and LSTM) are, what are the assumptions made when training them, what kind of datasets are they more suitable for a search ranking systems. (2) Attention/self-Attention: we explain the general attention mechanism. (3) Transformer: similar to the above two points, we explain the transformer architecture and motivate it with in the context of real search ranking and natural language processing task. (4) Hand-on session cover training attention/transformer models.

Knowledge Distillation for Search Relevance: In this session, we provide an introduction on Deep Structured Semantic Models (DSSM) [3] , which has been widely adopted in industry by its quality and efficient architecture. We also cover recent NLP breakthrough with BERT [2] significantly beats the DSSM and its variant in scoring query-document pairs. However, we show that its transformer crossing layer is expensive in the meantime so that it doesn't allow precomputing documents offline. To bridge the

All authors contributed equally.

KDD '22, August 14–18, 2022, Washington, DC, USA.

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542632

two, we share our proposed knowledge distillation [5] from teacher BERT model to student model with two tower frameworks. The new learning approach significantly wins over traditional DSMM models that were learned from clicks. During hands-on session, the audience get training on knowledge distillation for search relevance on an open source dataset. The code examples are provided to train a two tower student model and test dataset are used for audience to experience the metric difference between the teacher and student models.

Future Trend and Conclusion: In this session, we briefly summarize the pros/cons of current ranking models. We also have an open discussion with the audience about the future trend of AI-based search relevance algorithms.

When delivering the tutorial, we create discussions on related topics including but not limited to: what is search relevance, how to measure search relevance, why it is important, and what are those classical and advanced algorithms, etcs. At the same time, we interact with participants by providing guided answers with given case studies or sample data sets. The goal is to help the participants to understand search relevance algorithm and computation to apply to real-world. After attending the tutorial, we expect audience can get a deep understanding and have a more grounded perspective on the search relevance domain and also have first-hand experience building a search relevance system using deep learning techniques.

The target audience is anyone who works on or is interested in learning more about search relevance and large-scale search ranking systems, as well as gaining more practical experience of implementing such algorithms for solving real-world problems. We expect our tutorial is beneficial for researchers and engineer, in areas of information retrieval, data science, machine learning, recommender systems, etc., either working in the industry or in universities such as students and researchers.

SPEAKER BIBLIOGRAPHIES

Linsey Pang is currently a Principal Applied Scientist at Salesforce, California. She has co-authored papers in several conferences including KDD, NIPS, ICDM and published patents on time-series forecasting and semantic learning etc. She got her PhD degree from the School of Information Technologies at University of Sydney in 2015 and her research interests include data mining, machine learning, deep learning, etc.

Wei Liu is an Associate Professor in Machine Learning, and the Director of Future Intelligence Research Lab, in the School of Computer Science, the University of Technology Sydney. He obtained his PhD degree in Machine Learning research at the University of Sydney. His current research focuses are adversarial machine learning, cybersecurity, game theory, multimodal machine learning, natural language processing, and intrusion detection. Wei's research papers are constantly published in CORE A*/A and Q1 (i.e., top-prestigious) journals and conferences. He has received 3 Best Paper Awards. In addition, one of his first-authored papers received the Most Influential Paper Award in PAKDD 2021. Wei served as a tutorial co-chair at ICDM 2021.

Keng-hao Chang is a Principal Applied Scientist Manager at Bing Ads at Microsoft. Keng-hao is leading the Product Ads Selection and Relevance team. Keng-hao has co-authored several papers on attention models, chat bot, neural ad generation, and multi-modality retrieval in premier conferences including KDD and CVPR. Keng-hao graduated with a computer science PhD from UC Berkeley in 2012.

Xue Li is a senior SDE at Bing Ads Microsoft. She has co-authored several papers on improving search ads quality by leveraging techniques such as knowledge distillation, graphical neutral network, etc. Xue got her PhD. degree in Tsinghua University in 2016, and her PhD research focused mainly on Computer Vision and Machine Learning.

Moumita Bhattacharya is a Senior Research Scientist at Netflix research. Her research interest includes search ranking models and recommender systems. Moumita has a PhD in Computer Science from University of Delaware. Moumita has co-authored several papers on top-tier Journals and Conference proceedings, including American Journal of Cardiology and RecSys.

Xianjing Liu now is a Senior Machine Learning Engineer at Twitter. She is leading the machine learning modeling of the Twitter contextual ads. She has co-authored several papers and published patents on semantic search, semi-supervised learning, and text classification. She got her PhD degree from the University of Colorado at Boulder.

Stephen Guo is currently a Director of Machine Learning, Advertising Technology at Walmart Global Tech. He is leading machine learning teams for display advertising (targeting, measurement) and sponsored products. Stephen has published and presented his work in leading conferences or journals such as WWW, SIMOD, ICDE, EC, SDM, NAACL, etc. He studied Computer Science at Stanford University

Acknowledgement: Thanks for Connor (Bing Huan) Lee and Elizabeth Wang working on tutorial website design and maintenance.

REFERENCES

[1] Christopher D. Manning, Hinrich Schutze, and Prabhakar Raghavan. 2008. Introduction to Information Retrieval. *Cambridge University Press* (2008).

[2] Jacob Devlin, Ming-Wei Chang, Jianjin Zhang, Weihao Han, Kenton Lee, Kristina Toutanova, and Qi Zhang. 2019. Pre-training of Deep Bidirectional Transformers for Language Understanding:BERT. In *NAACL-HLT (1)*. 4171–4186.

[3] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[4] João Lages and Joao Paulo Carvalho. 2020. Relevance Ranking for Web Search. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–8. https://doi.org/10.1109/FUZZ48607.2020.9177802

[5] Xue Li, Zhipeng Luo, Jianjin Zhang, Weihao Han, Xianqi Chu, Liangjie Zhang, and Qi Zhang. 2019. Learning Fast Matching Models from Weak Annotation. In *WWW '19: The World Wide Web Conference*. 2985—-29914.

[6] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. 2017. Deeprank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 257–266.

[7] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond.* Now Publishers Inc.

[8] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*. 373–374.

[9] Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning clicks into purchases: Revenue optimization for product search in e-commerce. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 365–374.