Deep Learning for Matching in Search and Recommendation

Jun Xu
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
junxu@ict.ac.cn

Xiangnan He
School of Computing,
National University of Singapore
Singapore
xiangnanhe@gmail.com

Hang Li
Toutiao AI Lab
Beijing, China
hangli65@gmail.com

ABSTRACT

Matching is the key problem in both search and recommendation, that is to measure the relevance of a document to a query or the interest of a user on an item. Previously, machine learning methods have been exploited to address the problem, which learns a matching function from labeled data, also referred to as "learning to match" [21]. In recent years, deep learning has been successfully applied to matching and significant progresses have been made. Deep semantic matching models for search [25] and neural collaborative filtering models for recommendation [12] are becoming the state-of-the-art technologies. The key to the success of the deep learning approach is its strong ability in learning of representations and generalization of matching patterns from raw data (e.g., queries, documents, users, and items, particularly in their raw forms). In this tutorial, we aim to give a comprehensive survey on recent progress in deep learning for matching in search and recommendation. Our tutorial is unique in that we try to give a unified view on search and recommendation. In this way, we expect researchers from the two fields can get deep understanding and accurate insight on the spaces, stimulate more ideas and discussions, and promote developments of technologies.

The tutorial mainly consists of three parts. Firstly, we introduce the general problem of matching, which is fundamental in both search and recommendation. Secondly, we explain how traditional machine learning techniques are utilized to address the matching problem in search and recommendation. Lastly, we elaborate how deep learning can be effectively used to solve the matching problems in both tasks.

CCS CONCEPTS

• **Information systems → Web search engines**; **Recommender systems**; • **Computing methodologies → Neural networks**;

KEYWORDS

Learning to match; Deep learning; Web search; Recommender system

ACM Reference Format:
Jun Xu, Xiangnan He, and Hang Li. 2018. Deep Learning for Matching in Search and Recommendation. In *Proceedings of The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210181

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210181

1 INTRODUCTION

The explosive growth of various information on the Web has resulted in information overload, which greatly hinders users to accurately and timely obtain information of interest. Search and recommendation are two major approaches to address the challenge, which are two modes of information access: pull and push [8]. With information pull, the search engine accepts a query submitted by the user, and then returns relevant results to satisfy the user's information need. On the other hand, with information push the recommendation engine provides information that may be of interest to the user. The fundamental problem in both search and recommendation is how to conduct **matching** between heterogenous objects, which are query and document in search, user and item in recommendation, respectively.

The main technical difficulty in solving the matching problem lies in the so-called **semantic gap**. In search, traditional approaches perform query-document matching at the term level. However, a high degree of matching at the term level does not necessarily represent high relevance, and vice versa. For example, if the query is "ny times" and the document only contains "New York Times", then the matching degree of the query and the document is low, although they are relevant. Semantic gap is pervasive due to the ambiguous and variable nature of human language, since the same term can represent different meanings and the same meaning can be represented by different terms. While in recommendation, the problem of semantic gap is even severe, because the matching is performed between user attributes and item attributes, and there may not be any overlap between the features. For example, in the collaborative filtering setting, users and items are represented as ID features, and it is challenging to perform matching on superficial features of users and items.

To address the problem, researchers in the areas of search and recommendation have been taking similar approaches to perform matching at the semantic level, referred to as **semantic matching** [9]. In search, people try to perform more query and document understanding to represent the meanings of them (e.g, using topic models) and conduct better matching between the enriched query and document representations. Machine learning models have been developed for semantic matching and significant progress has been made, referred to as "learning to match" [21]. These methods conduct the matching through either mapping the query and document into a new semantic space [31], or conducting translation between the document and the query [2, 7]. In recommendation, people try to represent the user and the item as real-value vectors that encode rich semantics (e.g., semantically relevant objects should

have large similarities), and then perform matching at the semantic level [13, 17].

While these traditional approaches work well to some extent, their performance can still be limited by the insufficient representation ability of the models and simple matching functions.

Inspired by the recent renaissance of deep neural networks in computer vision and natural language processing, a number of deep learning methods have been developed for addressing the matching problem in search and recommendation. They have shown promising results and demonstrated great potentials for further improvements [5, 10, 14, 24]. Generally speaking, the success of deep learning (DL) in semantic matching mainly comes from two aspects: 1) representation learning, and 2) matching function learning.

- For representation learning, DL methods can learn abstract representations for data objects (or features), specially tailored for the matching task. For example, in search, Feedforward Neural Networks (FNNs) have been used to learn representations for queries and documents [16], and Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have also been used, for taking the ordering information of words into consideration [15, 23, 27, 28]. Similarly in recommendation, FNNs like Stacked Denoising Auto-Encoder (SDAE) have been used to enrich item representation learning from texts and images [34], and RNNs have been used to learn representation for sessions [22] and multimedia content [3].
- For matching function learning, DL methods utilize multi-level neural networks as the ranking function, which can effectively aggregate obscure low-level signals to the matching score. For example, in search, CNNs and RNNs have been used as the matching function to aggregate the term-level interaction signals [6, 15, 24, 29, 33]. Attention mechanism has also been used for the purpose [26]. In recommendation, CNNs [11], FNNs like Factorization Machine (FM) [10] and Multi-Layer Perceptron (MLP) [4, 12], and attention networks [32] have been integrated into the matching function to learn second-order and higher-order feature interactions. Several recent efforts have combined embedding-based and tree-based models to learn the matching function for recommendation [30, 35].

In this tutorial, we focus on the matching perspective of search and recommendation, aiming to deliver a systematic review on conventional machine learning as well as deep learning methods for addressing the problems. As pointed out by [1, 8], search (information retrieval) and recommendation (information filtering) are the two sides of the same coin, having strong connections and similarities. By unifying the two tasks under the same view of matching and comparably reviewing existing techniques, we can provide more insights into solving the semantic matching problems. We expect this tutorial to be useful for researchers and practitioners working on both tasks, since the innovations and experiences derived from one task might be transferable to the other. This will facilitate researchers from the search and recommendation communities having fruitful idea exchanges, promoting the technical developments of both search and recommender systems.

In addition to search and recommendation, matching also plays a central role in online advertising, question answering, image annotation, and drug design, and other applications. The solutions for semantic matching can be generalized to solve the matching problems between any two types of objects. As such, we believe the developments of matching techniques for search and recommendation can not only benefit each other, but also facilitate a wide range of other applications.

2 CONTENT AND SCHEDULE

The outline of the proposed tutorial is as follows.

1. Introduction (20 minutes)
   - 1.1 Search and recommendation
   - 1.2 The matching problem
   - 1.3 Organization of the tutorial

Part I: Traditional approaches to matching

2. Traditional matching models for search (20 minutes)
   - 2.1 Query-document matching
   - 2.2 Matching with translation model
   - 2.3 Matching with latent space model
3. Traditional matching models for recommendation (20 minutes)
   - 3.1 Collaborative filtering
   - 3.2 Matching with neighbor-based model
   - 3.3 Matching with latent factor model

Part II: Deep learning approaches to matching

4. A unified view for search and recommendation (30 minutes)
   - 4.1 Feature representation learning
   - 4.2 Matching function learning
5. Deep matching models for search (30 minutes)
   - 5.1 Methods of representation learning
   - 5.2 Methods of matching function learning
6. Deep matching models for recommendation (30 minutes)
   - 6.1 Methods of representation learning
   - 6.2 Methods of matching function learning
7. Conclusion and open discussions (10 minutes)

After briefly introducing the semantic matching problem, in Part I, we will recapitulate traditional learning methods for query-document matching in search and user-item matching in recommendation. In Part II, we will introduce deep learning approaches to matching. Specifically, we will first abstract a unified framework for deep learning solutions for search and recommendation, namely, feature representation learning and matching function learning. We then review previous works based on the unified view. Lastly, we will summarize the tutorial and discuss the future directions.

The tutorial materials to be supplied to the attendees include

Slides: tutorial slides will be made publicly available on the lecturers' personal homepages.

Bibliography: a list of references will cover all the work discussed in the tutorial and provide a good resource for further study.

The lecturers gave a tutorial entitled "semantic matching in search" at WWW 2012 [20], WSDM 2012 [19], and SIGIR 2012 [18]. In those tutorials, the traditional machine learning approaches to the semantic matching problem were introduced under the web search scenario. This tutorial is completely new with rich content of the recent technologies, including 1) the newly developed deep learning methods for matching, and 2) the methods for matching in recommendation.

Several wonderful tutorials were given at related conferences: Bhaskar Mitra and Nick Craswell, Neural Text Embeddings for Information Retrieval, at WSDM 2017; Kyomin Jungj, Byoung-Tak Zhan, and Prasenjit Mitra, Deep Learning for the Web, at WWW 2015; Tom Kenter et al., Neural Networks for Information Retrieval (NN4IR), at SIGIR 2017; Hang Li and Zhengdong Lu, Deep Learning for Information Retrieval, at SIGIR 2016; Ganesh Venkataraman et al., Deep Learning for Personalized Search and Recommender Systems, at KDD 2017; Alexandros Karatzoglou et al., Deep Learning for Recommender Systems, at Recsys 2017. This tutorial is significantly different from the previous tutorials in the sense that it focuses on the semantic matching problem in search and recommendation.

3 PRESENTERS' BIOGRAPHY

Dr. Jun Xu is a Professor at Institute of Computing Technology, Chinese Academy of Sciences. He received his Ph.D. in Computer Science from Nankai University in 2006. Jun Xu's research interests focus on applying machine learning to information retrieval. He has published about 40 papers and 1 monograph at top international journals and conferences, including TOIS, JMLR, SIGIR, CIKM, ACL etc. His work on information retrieval has received the Best Paper Runner-up of ACM CIKM 2017 and Best Paper Award of AIRS 2010. Jun Xu is very active in the research communities and has served or is serving top international conferences as Senior PC member or PC member, including SIGIR, ACML, KDD, WWW, ACL, NIPS, IJCAI, AAAI, CIKM, and top international journal of JASIST as editorial board member. Jun Xu has also been working on the development of several commercial products (e.g., Microsoft Bing 2010, Microsoft Office 2011, and Huawei GTS search) and is leading the Easy Machine Learning open source project. He has gave tutorials at top conferences like SIGIR, WSDM, WWW on the topic of semantic matching in search.

Dr. Xiangnan He is a senior research fellow with School of Computing, National University of Singapore (NUS). He received his Ph.D. in Computer Science from NUS. His research interests span recommender system, information retrieval, and multi-media processing. He has over 49 publications appeared in several top conferences such as SIGIR, WWW, MM, CIKM, and IJCAI, and journals including TKDE, TOIS, and TMM. His work on recommender system has received the Best Paper Award Honorable Mention of ACM SIGIR 2016 and WWW 2018. Moreover, he has served as the PC member for the prestigious conferences including SIGIR, WWW, KDD, MM, WSDM, CIKM, and ACL, and the regular reviewer for prestigious journals including TKDE, TOIS, TKDD, TMM etc.

Dr. Hang Li is director of Toutiao AI Lab, adjunct professors of Peking University and Nanjing University. He is an IEEE Fellow and an ACM Distinguished Scientist. His research areas include information retrieval, natural language processing, machine learning, and data mining. Hang graduated from Kyoto University in 1988 and earned his PhD from the University of Tokyo in 1998. He worked at NEC Research as researcher from 1990 to 2001, Microsoft Research Asia as senior researcher and research manager from 2001 to 2012, and chief scientist and director of Huawei Noah's Ark from 2012 to 2017. He joined Toutiao in 2017. Hang has published three technical books, and more than 120 technical papers at top international conferences including SIGIR, WWW, WSDM, ACL, EMNLP, ICML, NIPS, SIGKDD, AAAI, IJCAI, and top international journals including CL, NLE, JMLR, TOIS, IRJ, IPM, TKDE, TWEB, TIST. Hang is also very active in the research communities and has served or is serving top international conferences as PC chair, Senior PC member, or PC member, including SIGIR, WWW, WSDM, ACL, NACL, EMNLP, NIPS, SIGKDD, ICDM, IJCAI, ACML, and top international journals as associate editor or editorial board member, including CL, IRJ, TIST, JASIST, JCST. He gave tutorials at top conferences like SIGIR, WSDM, WWW many times on a number of topics with regard to machine learning for information retrieval and natural language processing.

4 ACKNOWLEDGEMENT

This work is funded by the National Key R&D Program of China under Grants No. 2016QY02D0405. This work is also under NExT, which is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

REFERENCES

[1] Nicholas J. Belkin and W. Bruce Croft. 1992. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Commun. ACM* 35, 12 (1992), 29–38.

[2] Adam Berger and John Lafferty. 1999. Information Retrieval As Statistical Translation. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 222–229.

[3] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 335–344. https://doi.org/10.1145/3077136.3080797

[4] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 191–198.

[5] Nick Craswell, W Bruce Croft, Maarten de Rijke, Jiafeng Guo, and Bhaskar Mitra. 2017. SIGIR 2017 Workshop on Neural Information Retrieval (Neu-IR'17). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 1431–1432.

[6] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 126–134. https://doi.org/10.1145/3159652.3159659

[7] Jianfeng Gao, Xiaodong He, and Jian-Yun Nie. 2010. Clickthrough-based Translation Models for Web Search: From Word Models to Phrase Models. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. ACM, New York, NY, USA, 1139–1148.

[8] Hector Garcia-Molina, Georgia Koutrika, and Aditya Parameswaran. 2011. Information seeking: convergence of search, recommendations, and advertising. *Commun. ACM* 54, 11 (2011), 121–130.

[9] Julio Gonzalo, Hang Li, Alessandro Moschitti, and Jun Xu. 2014. SIGIR 2014 Workshop on Semantic Matching in Information Retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 1296–1296.

[10] Xiangnan He and Tat-Seng Chua. 2017. Neural Factorization Machines for Sparse Predictive Analytics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 355–364.

[11] Xiangnan He, Xiaoyu Du, Xiang Wang, Feng Tian, Jinhui Tang, and Tat-Seng Chua. 2018. Out Product-based Neural Collaborative Filtering. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. AAAI Press.

[12] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 173–182.

[13] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and*

Development in Information Retrieval (SIGIR '16). ACM, New York, NY, USA, 549–558. https://doi.org/10.1145/2911451.2911489

[14] Balázs Hidasi, Alexandros Karatzoglou, Oren Sar-Shalom, Sander Dieleman, Bracha Shapira, and Domonkos Tikk. 2017. DLRS 2017: Second Workshop on Deep Learning for Recommender Systems. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. ACM, New York, NY, USA, 370–371.

[15] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2042–2050.

[16] 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 (CIKM '13)*. ACM, New York, NY, USA, 2333–2338.

[17] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (Aug. 2009), 30–37. https://doi.org/10.1109/MC.2009.263

[18] Hang Li and Jun Xu. 2012. Beyond Bag-of-words: Machine Learning for Query-document Matching in Web Search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 1177–1177.

[19] Hang Li and Jun Xu. 2012. Machine Learning for Query-document Matching in Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 767–768.

[20] Hang Li and Jun Xu. 2012. Semantic Matching in Search. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)*.

[21] Hang Li and Jun Xu. 2014. Semantic Matching in Search. *Foundations and Trends® in Information Retrieval* 7, 5 (2014), 343–469.

[22] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural Attentive Session-based Recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. ACM, New York, NY, USA, 1419–1428.

[23] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep Sentence Embedding Using Long Short-term Memory Networks: Analysis and Application to Information Retrieval. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 24, 4 (2016), 694–707.

[24] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text Matching As Image Recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, 2793–2799.

[25] 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 26th International Conference on Information and Knowledge Mangement (CIKM'17)*.

[26] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 2249–2255. http://aclweb.org/anthology/D/D16/D16-1244.pdf

[27] Xipeng Qiu and Xuanjing Huang. 2015. Convolutional Neural Tensor Network Architecture for Community-based Question Answering. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 1305–1311.

[28] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 101–110. https://doi.org/10.1145/2661829.2661935

[29] Shengxian Wan, Yanyan Lan, Jun Xu, Jiafeng Guo, Liang Pang, and Xueqi Cheng. 2016. Match-SRNN: Modeling the Recursive Matching Structure with Spatial RNN. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2922–2928.

[30] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2018. TEM: Tree-enhanced Embedding Model for Explainable Recommendation. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1543–1552. https://doi.org/10.1145/3178876.3186066

[31] Wei Wu, Hang Li, and Jun Xu. 2013. Learning Query and Document Similarities from Click-through Bipartite Graph with Metadata. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13)*. ACM, New York, NY, USA, 687–696.

[32] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17)*. AAAI Press, 3119–3125. http://dl.acm.org/citation.cfm?id=3172077.3172324

[33] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 55–64. https://doi.org/10.1145/3077136.3080809

[34] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative Knowledge Base Embedding for Recommender Systems. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 353–362.

[35] Qian Zhao, Yue Shi, and Liangjie Hong. 2017. GB-CENT: Gradient Boosted Categorical Embedding and Numerical Trees. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1311–1319. https://doi.org/10.1145/3038912.3052668