Deep Learning for Matching in Search and Recommendation
Other titles in Foundations and Trends® in Information Retrieval

Information Retrieval: The Early Years
Donna Harman
ISBN: 978-1-68083-584-7

Bandit Algorithms in Information Retrieval
Dorota Glowacka
ISBN: 978-1-68083-574-8

Neural Approaches to Conversational AI
Jianfeng Gao, Michel Galley and Lihong Li
ISBN: 978-1-68083-552-6

An Introduction to Neural Information Retrieval
Bhaskar Mitra and Nick Craswell
ISBN: 978-1-68083-532-8

Full text available at: http://dx.doi.org/10.1561/1500000076
Deep Learning for Matching in Search and Recommendation

Jun Xu  
Gaoling School of Artificial Intelligence  
Renmin University of China  
China  
junxu@ruc.edu.cn

Xiangnan He  
School of Information Science and Technology  
University of Science and Technology of China  
China  
hexn@ustc.edu.cn

Hang Li  
Bytedance AI Lab  
China  
lihang.lh@bytedance.com

Full text available at: http://dx.doi.org/10.1561/1500000076
Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2020, Volume 14, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Full text available at: http://dx.doi.org/10.1561/1500000076
## Contents

1 Introduction 6  
1.1 Search and Recommendation 6  
1.2 Unifying Search and Recommendation from Matching Viewpoint 8  
1.3 Mismatching Challenge in Search 10  
1.4 Mismatching Challenge in Recommendation 11  
1.5 Recent Advances 12  
1.6 About This Survey 13  

2 Traditional Matching Models 16  
2.1 Learning to Match 16  
2.2 Matching Models in Search and Recommendation 22  
2.3 Latent Space Models in Search 25  
2.4 Latent Space Models in Recommendation 28  
2.5 Further Reading 31  

3 Deep Learning for Matching 33  
3.1 Overview of Deep Learning 33  
3.2 Overview of Deep Learning for Matching 49  

4 Deep Matching Models in Search 55  
4.1 Matching Based on Representation Learning 58
Deep Learning for Matching in Search and Recommendation

Jun Xu¹, Xiangnan He² and Hang Li³

¹Gaoling School of Artificial Intelligence, Renmin University of China, China; junxu@ruc.edu.cn
²School of Information Science and Technology, University of Science and Technology of China; hexn@ustc.edu.cn
³Bytedance AI Lab, China; lihang.lh@bytedance.com

ABSTRACT

Matching is a key problem in both search and recommendation, which is to measure the relevance of a document to a query or the interest of a user to an item. Machine learning has been exploited to address the problem, which learns a matching function based on input representations and from labeled data, also referred to as “learning to match”. In recent years, efforts have been made to develop deep learning techniques for matching tasks in search and recommendation. With the availability of a large amount of data, powerful computational resources, and advanced deep learning techniques, deep learning for matching now becomes the state-of-the-art technology for search and recommendation. The key to the success of the deep learning approach is its strong ability in learning of representations and generalization of matching patterns from data (e.g., queries, documents, users, items, and contexts, particularly in their raw forms).
This survey gives a systematic and comprehensive introduction to the deep matching models for search and recommendation developed recently. It first gives a unified view of matching in search and recommendation. In this way, the solutions from the two fields can be compared under one framework. Then, the survey categorizes the current deep learning solutions into two types: methods of representation learning and methods of matching function learning. The fundamental problems, as well as the state-of-the-art solutions of query-document matching in search and user-item matching in recommendation, are described. The survey aims to help researchers from both search and recommendation communities to get in-depth understanding and insight into the spaces, stimulate more ideas and discussions, and promote developments of new technologies.

Matching is not limited to search and recommendation. Similar problems can be found in paraphrasing, question answering, image annotation, and many other applications. In general, the technologies introduced in the survey can be generalized into a more general task of matching between objects from two spaces.
## List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>Partial Least Square</td>
</tr>
<tr>
<td>RMLLS</td>
<td>Regularized Matching in Latent Space</td>
</tr>
<tr>
<td>SSI</td>
<td>Supervised Semantic Indexing</td>
</tr>
<tr>
<td>BMF</td>
<td>Biased Matrix Factorization</td>
</tr>
<tr>
<td>FISM</td>
<td>Factored Item Similarity Model</td>
</tr>
<tr>
<td>FM</td>
<td>Factorization Machine</td>
</tr>
<tr>
<td>FFN</td>
<td>Feedforward Neural Network</td>
</tr>
<tr>
<td>MLP</td>
<td>Multilayer Perceptron</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Networks</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Networks</td>
</tr>
<tr>
<td>GAN</td>
<td>Generative Adversarial Network</td>
</tr>
<tr>
<td>AE</td>
<td>Autoencoders</td>
</tr>
<tr>
<td>DAE</td>
<td>Denoising Autoencoder</td>
</tr>
<tr>
<td>CBOW</td>
<td>Continuous Bag of Words</td>
</tr>
<tr>
<td>SG</td>
<td>Skip Gram</td>
</tr>
<tr>
<td>BERT</td>
<td>Bidirectional Encoder Representations from Transformers</td>
</tr>
<tr>
<td>DSSM</td>
<td>Deep Structured Semantic Models</td>
</tr>
<tr>
<td>CLSM</td>
<td>Convolutional Latent Semantic Model</td>
</tr>
<tr>
<td>CNTN</td>
<td>Convolutional Neural Tensor Network</td>
</tr>
<tr>
<td>LSTM-RNN</td>
<td>Recurrent Neural Networks with Long Short-Term Memory cells</td>
</tr>
<tr>
<td>NVSM</td>
<td>Neural Vector Space Model</td>
</tr>
</tbody>
</table>
SNRM  Standalone Neural Ranking Model
ACMR  Adversarial Cross Modal Retrieval
ARC-II Convolutional Matching Model II
DRMM  Deep Relevance Matching Model
K-NRM  Kernel Based Neural Ranking Model
DeepMF Deep Matrix Factorization
CDAE Collaborative Denoising Auto-Encoder
NAIS  Neural Attentive Item Similarity
NARM  Neural Attentive Recommendation Machine
DeepCoNN Deep Cooperative Neural Networks
NARRE Neural Attention Regression with Review-level Explanation
VBPR  Visual Bayesian Personalized Ranking
CDL  Comparative Deep Learning
ACF Attentive Collaborative Filtering
NGCF Neural Graph Collaborative Filtering
KGAT Knowledge Graph Attention Network
KPRN Knowledge Path Recurrent Network
NCF Neural Collaborative Filtering
ConvNCF Convolutional Neural Collaborative Filtering
GMF Generalized Matrix Factorization
NeuMF Neural Matrix Factorization
CML Collaborative Metric Learning

Full text available at: http://dx.doi.org/10.1561/1500000076
**TransRec**  Translation-based Recommendation

**LRML**  Latent Relational Metric Learning

**NFM**  Neural Factorization Machine

**AFM**  Attentional Factorization Machine
1

Introduction

1.1 Search and Recommendation

With the rapid growth of the internet, one of the fundamental problems in information science becomes even more critical today, that is, how to identify the information satisfying a user’s need from a usually huge pool of information. The goal is to present the user only the information that is of interest and relevance, at the right time, place, and context. Nowadays, two types of information accessing paradigms, search and recommendation, are widely used in a great variety of scenarios.

In search, documents (e.g., web documents, Twitter posts, or E-commerce products) are first pre-processed and indexed in the search engine. After that, the search engine takes a query (a number of keywords) from the user. The query describes the user’s information need. Relevant documents are retrieved from the index, matched with the query, and ranked according to their relevance to the query. For example, if a user is interested in news about quantum computing, the query “quantum computing” may be submitted to a search engine and get news articles about the topic will be returned.
1.1. Search and Recommendation

Different from search, a recommendation system typically does not take a query. Instead, it analyzes the user’s profile (e.g., demographics and contexts) and historical interactions on items, and then makes recommendation on items to the user. The user features and item features are indexed and stored in the system in advance. The items are ranked according to the likelihood that the user is interested in them. For example, on a news website, when a user browses and clicks a new article, several news articles with similar topics or news articles that other users have clicked together with the current one may be shown.

Table 1.1 summarizes the differences between search and recommendation. The fundamental mechanism of search is “pull”, because users first make specific requests (i.e., submit queries) and then receive information. The fundamental mechanisms of recommendation is “push”, because users are provided information which they do not specifically request (e.g., submit queries). Here “beneficiary” means the people whose interests are to be met in the task. In a search engine, the results are typically created solely based on the user’s needs, and thus the beneficiary is the users. In a recommendation engine, the results usually need to satisfy both the users and providers, and thus the beneficiary is all of them. However, the distinction is becoming blurred recently. For example, some search engines mix search results with paid advertisements, which benefits both the users and the providers. As for “serendipity”, it means that conventional search focuses more on information that is clearly relevant. Conventional recommendation, on the other hand, is allowed to offer unexpected but useful information.

Table 1.1: Information-providing mechanisms of search and recommendation

<table>
<thead>
<tr>
<th></th>
<th>Search</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query available</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Delivery model</td>
<td>Pull</td>
<td>Push</td>
</tr>
<tr>
<td>Beneficiary</td>
<td>User</td>
<td>User and provider</td>
</tr>
<tr>
<td>Serendipity</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
1.2 Unifying Search and Recommendation from Matching Viewpoint

Garcia-Molina et al. (2011) pointed out that the fundamental problems in search and recommendation are to identify information objects satisfying users’ information needs. It is also indicated that search (information retrieval) and recommendation (information filtering) are the two sides of the same coin, having strong connections and similarities (Belkin and Croft, 1992). Figure 1.1 illustrates the unified matching view of search and recommendation. The goal in common is to present to the users the information they need.

Search is a retrieval task, which aims to retrieve the documents that are relevant to the query. In contrast, recommendation is a filtering task, which aims to filter out the items that are of interest to the user (Adomavicius and Tuzhilin, 2005). As such, search can be considered as conducting matching between queries and documents, and recommendation can be considered as conducting matching between users and items. More formally, both the matching in search and recommendation can be considered as constructing a matching model \( f: \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{R} \) which calculates the matching degree between two input objects \( x \) and \( y \), where \( \mathcal{X} \) and \( \mathcal{Y} \) denote two object spaces. \( \mathcal{X} \) and \( \mathcal{Y} \) are the spaces of queries and documents in search, or the spaces of users and items in recommendation.

Under the unified matching view in Figure 1.1, we use the term information objects to denote the documents/items to retrieve/recommend, and use information needs to denote the queries/users in the respective task. By unifying the two tasks under the same view of matching and
comparably reviewing existing techniques, we can provide deeper insights and more powerful solutions to the problems. Moreover, unifying the two tasks also has practical and theoretical implications.

Search and recommendation have already been combined in some practical applications. For example, at some E-commerce sites, when the user submits a query, a ranking list of products are presented based on not only relevance (query-product matching) but also user interest (user-product matching). In some lifestyle apps, when the user searches for restaurants, the results are returned based on both relevance (query-restaurant matching) and user interest (user-restaurant matching). There is a clear trend that search and recommendation will be integrated into a single system at certain scenarios to meet users’ needs better, where matching plays an essential role.

Search and recommendation already have many shared technologies because of their similarities in matching. Some search problems can be solved by using recommendation techniques (Zamani et al., 2016), and vice versa (Costa and Roda, 2011), on the basis of matching. With the use of deep learning technologies, the matching models for search and recommendation bear even more resemblance in architecture and methodology, as reflected in the techniques: embedding the inputs (queries, users, documents, and items) as distributed representations, combining neural network components to represent the matching function, and training the model parameters in an end-to-end manner. Moreover, search and recommendation can be jointly modeled and optimized if they share the same set of information objects (as in the above examples of E-commerce sites and lifestyle apps) (Schedl et al., 2018; Zamani and Croft, 2018a, 2020). Therefore, in order to develop more advanced ones, it is necessary and advantageous to take a unified matching view to analyze and compare existing search and recommendation technologies.

The matching tasks in search and in recommendation face different challenges in practice. The underlying problem is essentially the same, however, that is, the mismatch challenge. Next, we introduce the key challenges of the two tasks, respectively.
1.3 Mismatching Challenge in Search

In search, queries and documents (usually their titles) are taken as texts. The relevance of a document to a query is mainly represented by the matching degree between the two. The document is considered relevant to the query if the matching degree is high. Natural language understanding by computer is still challenging, and thus the calculation of matching degree is still limited to the text level but not at the semantic level. A high match degree at the text level does not necessarily mean high relevance at the semantic level, and vice versa. Moreover, queries are issued by users, while documents are compiled by editors. Due to the ambiguity of natural language, users and editors are likely to use different language styles and expressions for presenting the same concepts or topics. As a result, the search system may suffer from the so-called query-document mismatch problem. Specifically, when the users of a search engine and the editors of the documents use different texts to describe the same concept (e.g., “ny times” vs. “new york times”), query-document mismatch may occur. This is still one of the main challenges for search. Moving to the cross-modal IR (e.g., using text queries to retrieve image documents), the query-document mismatch problem becomes even more severe, because different modalities have different types of representations. In cross-modal retrieval, one major challenge is how to construct a matching function that can bridge the “heterogeneity gap” amongst the modalities.

To address the query-document mismatch challenge, methods have been proposed to perform matching at the semantic level, referred to as semantic matching. The key idea in the solutions is either to perform more query and document understanding to better represent the meanings of the query and document, or to construct more powerful matching functions that can bridge the semantic gap between the query and document. Both traditional machine learning approaches (Li and Xu, 2014) and deep learning approaches (Guo et al., 2019b; Mitra and Craswell, 2018; Onal et al., 2018) have been developed for semantic matching.
1.4 Mismatching Challenge in Recommendation

The mismatching problem is even more severe in recommendation. In search, queries and documents consist of terms in the same language, making it at least meaningful to conduct direct matching on their terms. In recommendation, however, users and items are usually represented by different types of features, for example, the features of users can be the user ID, age, income level, and recent behaviors, while the features for items can be the item ID, category, price, and brand name. Since the features of users and items are from the spaces of different semantics, the naive approaches based on the matching of superficial features do not work for recommendation. More challengingly, the items can be described by multi-modal features, e.g., images of clothing products and cover images of movies, which could play a pivotal role in affecting the decision-making of users. In such visually-aware scenarios, we need to consider the cross-modal matching between users and multi-modal content.

To address the mismatching challenge in recommendation, the collaborative filtering principle has been proposed (Shi et al., 2014). Collaborative Filtering (CF), which works as the fundamental basis of almost all personalized recommender systems, assumes that a user may like (consume) the items that are liked (consumed) by the similar users, for which the similarity is judged from the historical interactions (Sarwar et al., 2001). However, directly evaluating the similarity between users (items) suffers from the sparsity issue, since a user only consumed a few items in the whole item space. A typical assumption to address the sparsity issue is that the user-item interaction matrix is low-rank, which thus can be estimated from low-dimensional user (and item) latent feature matrix. Then the user (item) similarity can be more reliably reflected in the latent feature matrix. This leads to the effectiveness of matrix factorization for collaborative filtering (Koren et al., 2009; Rendle et al., 2009), which becomes a strong CF method and an essential design for many recommender models. Besides matrix factorization, many other types of CF methods have been developed.

\(^1\)Here we do not consider cross-language information retrieval.
like neural network-based methods (He et al., 2017c; Liang et al., 2018) and graph-based methods (Wang et al., 2019b; Ying et al., 2018).

To leverage the various side information beyond the interaction matrix, such as user profiles, item attributes, and the current contexts, many generic recommender models that follow the standard supervised learning paradigm have been proposed. These models can be used in the (re-)ranking stage of a recommendation engine, e.g., by predicting the click-through rate (CTR) of an item. A representative model is factorization machine (FM) (Rendle, 2010), which extends the low-rank assumption of matrix factorization to model feature interactions. Since the expressiveness of FM is limited by its linearity and second-order interaction modeling, many later efforts complement it with neural networks for nonlinear and higher-order interaction modeling (He and Chua, 2017; Lian et al., 2018; Zhou et al., 2018). These neural network models have now been intensively used in industrial applications. Batmaz et al. (2019) and Zhang et al. (2019) reviewed deep learning methods for recommendation systems.

Please note that though query-document matching and user-item matching are critical for search engines and recommendation systems, these systems also include other important components. Besides matching, web search engines also include crawling, indexing, document understanding, query understanding, and ranking, etc. Recommendation systems also include components such as user modeling (profiling), indexing, caching, diversity controlling, and online exploration, etc.

1.5 Recent Advances

Though traditional machine learning was successful for matching in search and recommendation, recent advances in deep learning have brought even more significant progress to the area with a large number of deep matching models proposed. The power of deep learning models lies in the ability to learn distributed representations from the raw data (e.g., text) for the matching problem, to avoid many limitations of hand-crafted features, and to learn the representations and matching networks in an end-to-end fashion. Moreover, deep neural networks have sufficient capacity to model complicated matching tasks. They have
the flexibility of extending to cross-modal matching naturally, where the common semantic space is learned to represent data of different modalities universally. All these characteristics are helpful in handling the complexity of search and recommendation.

In search, the mismatch between query and document is more effectively addressed by deep neural networks, including the feed-forward neural networks (FFNs), convolutional neural networks (CNNs), and Recurrent neural networks (RNNs), because they have stronger capabilities in representation learning and matching function learning. Most notably, Bidirectional Encoder Representations from Transformers (BERT) has significantly enhanced the accuracy of matching in search and stands out as the state-of-the-art technique now.

In recommendation, recent focus has shifted from behavior-centric collaborative filtering to information-rich user-item matching as in sequential, context-aware, and knowledge graph enhanced recommendations, which are all practical scenario-driven. In terms of techniques, graph neural networks (GNNs) become an emerging tool for representation learning (Wang et al., 2019a,b), because recommendation data can be naturally organized in a heterogeneous graph and GNNs have the capability to exploit such data. To handle user behavior sequence data, self-attention and BERT are also adopted, which demonstrates promising results in sequential recommendation (Sun et al., 2019; Yuan et al., 2020).

1.6 About This Survey

This survey focuses on the fundamental problems of matching in search and recommendation. State-of-the-art matching solutions using deep learning are described. A unified view of search and recommendation from matching is provided. The ideas and solutions explained may motivate industrial practitioners to turn the research results into products. The methods and the discussions may help academic researchers to develop new approaches. The unified view may bring researchers in the search and the recommendation communities together and inspire them to explore new directions.
The survey is organized as follows: Section 2 describes the traditional machine learning approaches to matching for search and recommendation; Section 3 gives a general formulation of deep matching methods; Section 4 and Section 5 describe the details of the deep learning approaches to search and recommendation respectively. Each section includes the representation learning-based approaches and matching function learning-based approaches; Section 6 summarizes the survey and discusses open problems. Sections 2, 3, 4, and 5 are self-contained, and the readers can choose to read on the basis of their interest and need.

Note that deep learning for search and recommendation is a very hot topic of research. As such, this survey does not try to cover all related works in the fields of information retrieval and recommender systems. Instead, we discuss the most representative approaches of the two fields from the viewpoint of matching, aiming to summarize their key ideas which are general and essential. In particular, this survey covers the representative work before 2019.

Several previous FnTIR issues have given detailed introductions to related topics. One issue (Li and Xu, 2014) introduces the traditional machine learning approaches to the semantic matching problem, particularly in web search. Our survey in this issue is very different from it in the sense that (1) it focuses on the newly developed deep learning methods, and (2) it considers both search and recommendation. Mitra and Craswell (2018) conducted a comprehensive survey on deep neural networks for information retrieval, referred to as Neural IR. Bast et al. (2016) carries out a survey on the techniques and systems of semantic search, which means search with keyword queries, structured queries, and natural language queries, to documents, knowledge bases, and their combinations.

Several surveys and tutorials have been made on deep learning for information retrieval and recommendation. For example, Onal et al. (2018) have explained neural models for ad-hoc retrieval, query understanding, question answering, sponsored search, and similar item retrieval. Zhang et al. (2019) reviews deep learning-based recommendation methods according to the taxonomy of deep learning techniques, e.g., MLP, CNN, RNN, autoencoder-based, and so on. Other related
surveys and tutorials include Kenter et al. (2017), Li and Lu (2016), Guo et al. (2019b), Batmaz et al. (2019), and Zhang et al. (2019). They all quite differ from this survey, which summarizes existing work from the perspective of matching (e.g., input representations and the way for matching).

This survey focuses on state-of-the-art matching techniques using deep learning. We expect that the readers have a certain knowledge of search and recommendation. Those who are not familiar with the areas may consult existing materials (e.g., Adomavicius and Tuzhilin, 2005; Croft et al., 2009; Li and Xu, 2014; Liu, 2009; Ricci et al., 2015). We also assume that the readers have sufficient knowledge of machine learning, particularly deep learning.


References


References


References


References


References


References


References


