

Foundations and Trends[®] in Information Retrieval

Information Discovery in E-commerce

Suggested Citation: Zhaochun Ren, Xiangnan He, Dawei Yin and Maarten de Rijke (2024), "Information Discovery in E-commerce", Foundations and Trends[®] in Information Retrieval: Vol. 18, No. 4-5, pp 417–690. DOI: 10.1561/1500000097.

Zhaochun Ren

Leiden University
z.ren@liacs.leidenuniv.nl

Xiangnan He

University of Science and Technology of China
xiangnanhe@gmail.com

Dawei Yin

Baidu Inc.
yindawei@acm.com

Maarten de Rijke

University of Amsterdam
m.derijke@uva.nl

This article may be used only for the purpose of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval.

now

the essence of knowledge

Boston — Delft

Contents

1	Introduction	419
1.1	Motivation	419
1.2	Aims of this Survey	420
1.3	Outline	421
1.4	Our Readers	424
2	Definitions and Background	425
2.1	Background	425
2.2	User Modeling	426
2.3	Information Retrieval in E-commerce	427
2.4	Conversational AI	430
3	E-commerce Presentations and Users	433
3.1	E-commerce Presentations	434
3.2	E-commerce Users	454
3.3	Discussion	463
4	E-commerce User Modeling	465
4.1	User Behavior Modeling in E-commerce	466
4.2	User Profiling in E-commerce	480
4.3	Emerging Directions	483

5	E-commerce Search	487
5.1	Characteristics of E-commerce Search	488
5.2	Evaluation Metrics	492
5.3	Matching Strategies in E-commerce Search	496
5.4	Ranking Strategies in E-commerce Search	508
5.5	Emerging Directions	511
6	E-commerce Recommendation	518
6.1	Characteristics of E-commerce Recommendation	518
6.2	Candidate Retrieval Models	521
6.3	Candidate Ranking Models	531
6.4	Re-ranking Strategies	537
6.5	Emerging Directions	538
7	E-commerce QA and Conversations	548
7.1	Question Answering in E-commerce	548
7.2	Dialogue Systems in E-commerce	560
7.3	Emerging Directions	578
8	Conclusion and Outlook	582
8.1	Conclusion	582
8.2	Outlook	584
	Acknowledgements	590
	Appendix	592
	References	607

Information Discovery in E-commerce

Zhaochun Ren¹, Xiangnan He², Dawei Yin³ and Maarten de Rijke⁴

¹*Leiden University, The Netherlands; z.ren@liacs.leidenuniv.nl*

²*University of Science and Technology of China, China;
xiangnanhe@gmail.com*

³*Baidu Inc., China; yindawei@acm.com*

⁴*University of Amsterdam, The Netherlands; m.derijke@uva.nl*

ABSTRACT

Electronic commerce, or e-commerce, is the buying and selling of goods and services, or the transmitting of funds or data online. E-commerce platforms come in many kinds, with global players such as Amazon, Airbnb, Alibaba, Booking.com, eBay, and JD.com and platforms targeting specific geographic regions such as Bol.com and Flipkart.com. Information retrieval has a natural role to play in e-commerce, especially in connecting people to goods and services. Information discovery in e-commerce concerns different types of search (e.g., exploratory search vs. lookup tasks), recommender systems, and natural language processing in e-commerce portals. The rise in popularity of e-commerce sites has made research on information discovery in e-commerce an increasingly active research area. This is witnessed by an increase in publications and dedicated workshops in this space. Methods for information discovery in e-commerce largely focus on improving the effectiveness of e-commerce search and recommender systems, on enriching and using

Zhaochun Ren, Xiangnan He, Dawei Yin and Maarten de Rijke (2024), “Information Discovery in E-commerce”, *Foundations and Trends® in Information Retrieval*: Vol. 18, No. 4-5, pp 417–690. DOI: 10.1561/15000000097.

©2024 Z. Ren *et al.*

knowledge graphs to support e-commerce, and on developing innovative question answering and bot-based solutions that help to connect people to goods and services. In this survey, an overview is given of the fundamental infrastructure, algorithms, and technical solutions for information discovery in e-commerce. The topics covered include user behavior and profiling, search, recommendation, and language technology in e-commerce.

1

Introduction

1.1 Motivation

Over the past 20 years, we have seen an explosive growth of e-commerce portals, such as Alibaba, Amazon, eBay, and JD.com. These developments have reshaped people's shopping habits. An increasing number of customers now prefer to spend more time shopping online, generating billions of user requests per day. As part of the process of serving customer requests, large volumes of multi-modal data, including user search logs, clicks, orders, reviews, images, and chat logs, etc., are being generated. From an information retrieval point of view, discovering and employing pertinent information from the sheer volume of e-commerce data so as to enhance the performance of e-commerce services presents interesting challenges, both for academic and industrial researchers. In this survey we describe those challenges and the solutions that the community has so far proposed.

The topics of information discovery in e-commerce can be divided into several main directions:

- e-commerce presentation and users;
- user behavior and profiling;
- search in e-commerce;

- recommender systems in e-commerce; and
- question answering and dialogue systems in e-commerce.

Each of these areas comes with its own set of research challenges. For example, in e-commerce search there may be no hypertext links between products, thus excluding an important type of ranking signal that is often used in the setting of web search. But with click streams and order streams we have two parallel sources of ranking signal, a characteristic e-commerce feature that is absent from more traditional search scenarios.

E-commerce information discovery problems are wide in scope as the underlying discovery tasks concern a broad range of interaction modalities. There is a growing body of established methods in the e-commerce, aimed at developing algorithms for analyzing user behavior, for product search, for recommender systems, and for question answering and dialogue systems. These areas, and the methods developed, form the core around which most ongoing research efforts concerning information discovery for e-commerce are organized. The time is right to organize this material and to present it to a broad audience of interested information retrieval researchers, whether junior or senior, whether academic or industrial (Tsagkias *et al.*, 2020).

1.2 Aims of this Survey

A key aim of this survey is to bring together, and offer a unified perspective on, the large number of methods for e-commerce information discovery available today. To achieve this, we describe the basic architecture used for information discovery in e-commerce, algorithms for e-commerce information discovery, and evaluation principles. We supplement this with an account of available datasets and software based on these. We also introduce e-commerce applications accompanied by examples.

The survey targets practitioners and researchers from academia and industry and aims to present them with the challenges, state-of-the-art approaches, and the most urgent open questions in information discovery for e-commerce. Specifically, in terms of content, the objectives of the survey are as follows:

- To introduce tasks that constitute the information discovery problem in e-commerce, and to explain the difference between e-commerce information discovery and related work in other domains;
- To describe e-commerce information discovery algorithms in a unified way, i.e., using common notation and terminology, so that different models can easily be related to each other;
- To explain how to analyze the performance of e-commerce information discovery algorithms and why it is worth the effort;
- To present appropriate experimental and evaluation methodologies for e-commerce information discovery in both synthetic and real world settings; and
- To discuss future directions of research in e-commerce information discovery.

1.3 Outline

Information discovery aims to distill pertinent information from datasets with various modalities; it plays a role in many areas, ranging from web search to academic search and medical search. What is different about the e-commerce setting is that many traditional ranking features are either not present or present in a different form (Degenhardt *et al.*, 2017). Instead, discovery processes need to be supported based on structured information, semi-structured information, or information that might have facets such as price, ratings, title, description, seller location, etc.

1.3.1 Topics covered

We break the e-commerce information discovery problem down into five research directions: (i) e-commerce information presentation and users, (ii) user behavior and profiling in e-commerce, (iii) search in e-commerce, (iv) recommendation in e-commerce, and (v) question answering and dialogue systems in e-commerce. Below, we briefly describe each of these five directions.

The first direction concerns preliminaries about e-commerce information presentation and users. E-commerce portals provide various

modalities of information to users, e.g., rankings of products, product titles, descriptions, tips, and user reviews, etc. Multiple genres and types of text analysis can be employed to enhance e-commerce services, e.g., review filtering, review analysis, and normalization of production descriptions. User characteristics in e-commerce, e.g., browsing modules, clicks, purchases, and dwell time, generate multiple patterns for e-commerce scenarios. These two factors play fundamental roles in e-commerce information discovery. In this survey, we summarize recent work on both e-commerce information presentation and user characteristics.

The second direction concerns user behavior modeling and user profiling. Tracking and profiling users' behavior on e-commerce portals are important prerequisites for many e-commerce services, such as recommender systems, search, and online advertising. In this survey, we summarize recent work on user behavior modeling in e-commerce and introduce solutions to profiling users of e-commerce services.

The third direction of this survey concerns search in e-commerce, which examines approaches for product search scenarios on e-commerce portals. Just like, e.g., traditional web search, the target of this task is to satisfy users' needs. However, product search in e-commerce sites should be realized with different types of features than, e.g., web search, with the availability of a large number of products, query attributes, and engagement features. Moreover, calculating relevance in product search faces challenges regarding gaps between users and products. The target corpora can be structured, semi-structured, or unstructured, or a mixture of these; semantic search against such diverse sources raises interesting research challenges.

The fourth direction concerns recommendations in e-commerce. In contrast to traditional research on recommender systems that focuses on rating prediction, e-commerce recommender systems aim to tackle three challenges: the huge volume of products, sparsity, and data richness. Due to the existence of a very large number of candidate items in e-commerce portals, of which only a small fraction will attract a user's attention, e-commerce recommendation methods usually follow a two-stage recommendation framework with (i) candidate retrieval, and (ii) candidate ranking. The first phase of candidate retrieval goes through

the whole product catalog, and selects a small set of products that might match the information need. The second phase of candidate ranking ranks the candidates to present the final top- K products to the user. Given structured user behavior logs and semi-structured data about product features, e-commerce knowledge bases can be created to assist the candidate generation step. And the candidate ranking procedure ranks the retrieved candidate items for a better conversion rate or click-through rate, based on various machine learning models.

The fifth and final direction of this survey concerns question answering and dialogue systems in e-commerce. We survey recent work on e-commerce question answering and dialogue systems that have attracted increased attention. For dialogue systems, we describe both task-oriented dialogue systems, aimed at helping users complete a task in an e-commerce setting, and non-task-oriented dialogue systems aimed at generating fluent and engaging responses.

For the directions listed above, our ambition has been to cover related work up to the spring of 2023.

1.3.2 Topics not covered

E-commerce impacts large parts of our economy and society, including markets and retailers, supply chain management, and employment. With the development of data science, business intelligence studies on e-commerce marketing, e.g., sales volume forecasting and time series analysis, are receiving an increasing amount of attention. All of these areas are important, scientifically challenging, and deserving of attention from the information retrieval community. However, our focus will be limited to information discovery within the context of e-commerce. Specifically, we will not address topics such as computational advertising approaches that are irrelevant to search and recommendation, marketing strategies, forecasting, or information management in e-commerce.

1.3.3 Structure of the survey

The remainder of this survey is organized as follows. Section 2 provides key definitions and background related to e-commerce information discovery, drawing from user modeling, search, recommender systems,

question-answering, and dialogue systems. Section 3 describes preliminaries of e-commerce presentations as well as e-commerce users, including user behavior characteristics, and relevant language technologies and their use in e-commerce applications. Section 4 details user behavior modeling and user profiling approaches in e-commerce, including click behavior tracking, post-click tracking, purchase behavior modeling, and user profiling in e-commerce. Section 5 describes recent approaches proposed for e-commerce search, which we organize along two lines: research about the matching problem in e-commerce search, and about ranking strategies for e-commerce search. Section 6 presents algorithms and solutions for recommender systems in e-commerce. After introducing the two-stage recommendation framework in e-commerce portals, we organize the e-commerce recommendation studies into two groups: candidate retrieval models and candidate ranking models. We survey e-commerce question answering and dialogue systems in Section 7, where we introduce recent studies on e-commerce question answering and dialogue systems, respectively. In Section 8 we conclude this survey and identify emerging research directions and issues for future work.

1.4 Our Readers

We expect this survey to be useful to both academic and industrial researchers who either want to develop e-commerce information discovery methods, use them in their own research, or apply the methods described in the survey to improve product performance in e-commerce services. The intention is to help our audience acquire domain knowledge and to promote information discovery research activities in e-commerce.

To be able to benefit from this survey, we expect the reader to have a background in information retrieval, natural language processing, or machine learning. We recommend that readers read the material that we offer from start to finish, in the order that we offer it. However, readers who have a specific interest in search, or in recommender systems, or in conversational technology in e-commerce should read Sections 3 and 4 first before skipping ahead to Sections 5, 6, or 7, respectively.

2

Definitions and Background

The section presents definitions and background applied to e-commerce information discovery studies from the perspectives of research communities on user modeling, search, recommender systems, question-answering, and dialogue systems. We first introduce relevant concepts about user modeling, information retrieval, recommender systems, and conversational AI. We then introduce definitions and notations associated with e-commerce information discovery. Next, we explore fundamental concepts in e-commerce information discovery, including e-commerce information presentation, e-commerce search, e-commerce recommendation, and e-commerce conversational AI systems. The glossary and notations attached to these concepts are introduced in the last part of this section.

2.1 Background

E-commerce has revolutionized how consumers interact with products and services, fundamentally altering the landscape of information discovery. There are plenty of relevant research perspectives on information discovery in e-commerce. Unlike traditional retail settings, where physical exploration and interaction drive decision-making, e-commerce relies

on digital mechanisms to guide users through vast and often overwhelming amounts of information. As a result, the effectiveness of e-commerce platforms hinges on their ability to deliver personalized, relevant, and timely information to users.

Various research disciplines – such as user modeling, information retrieval, recommender systems, question-answering, and conversational AI – contribute significantly to enhancing the e-commerce experience. Each of these fields offers unique insights and methodologies for addressing key challenges in e-commerce, such as understanding user intent, predicting user preferences, addressing user concerns, satisfying user needs, and facilitating seamless product discovery. We list the fundamental concepts and research perspectives behind these areas, laying the groundwork for understanding how e-commerce platforms enable efficient and effective information discovery for users.

Information discovery in e-commerce. In the context of e-commerce, information discovery refers to the process by which users engage with relevant products or services based on their specific needs, regardless of the format or presentation of that information. This process encompasses a variety of functions, including search, recommendation, and personalized content delivery and presentation. At its core, information discovery in e-commerce involves not only retrieving relevant products but also understanding user intent and preferences to provide the most suitable results. It relies on algorithmic solutions to identify, search, recommend, and display information that aligns with user requirements. Whether through search queries, personalized recommendations, or curated content, information discovery systems enable users to efficiently navigate large product catalogs and find what they need in a seamless and engaging manner.

2.2 User Modeling

User modeling refers to the process of creating a representation of a user's characteristics, behaviors, preferences, and goals in order to personalize user-system interactions or appropriate content for that user. It is widely used in fields like information retrieval, recommender systems, and conversational AI. User modeling is a critical component

of personalizing e-commerce experiences. It involves the construction of user profiles based on behavioral data, preferences, demographics, and interactions within the system. These profiles help systems adapt content, recommendations, and interactions to suit individual needs. Meanwhile, e-commerce platforms present information in various formats, such as lists, grids, or interactive elements, which can significantly affect user engagement and conversion rates. Understanding user behavior and preferences by optimizing these presentations is important for enhancing the user experience and user satisfaction.

In e-commerce, user modeling can use implicit feedback (e.g., clicks, purchases, carting, and user engagement) and explicit feedback (e.g., ratings and reviews) to predict a user's future behavior, preferences, and profiling. Techniques like collaborative filtering, content-based filtering, and hybrid models are frequently employed in user modeling during early studies on this topic. In recent years, deep neural networks and pre-trained language models have been successfully applied to user modeling in e-commerce portals. User modeling spans across domains such as cognitive science, machine learning, and human-computer interaction, contributing to the development of systems that continuously refine the understanding of users as they interact with the platform.

2.3 Information Retrieval in E-commerce

Information retrieval (IR) has been playing a critical role in e-commerce services. Search and recommendation functionalities have been applied to e-commerce portals almost since the beginning of e-commerce. Beyond traditional search functionalities, IR techniques have evolved to support a wide range of features, including search and recommendation, making them essential for delivering a seamless user engagement. In e-commerce platforms, IR techniques are responsible for retrieving relevant items from a vast product catalog based on a user's query or search intent. This process involves not only matching query terms to product descriptions but also understanding the broader context behind the query, such as user preferences, purchase history, and real-time behaviors.

E-commerce search. E-commerce search involves techniques and algorithms used to allow users to efficiently find products or services

within an online store. This includes the use of keywords, filters, and advanced semantic search technologies. E-commerce search differs from traditional information retrieval because it often focuses on product features, pricing, availability, and user preferences. Research in this area spans areas like query understanding, ranking algorithms, and the integration of multimodal data (e.g., images and reviews).

The effectiveness of an e-commerce search engine depends heavily on how well it can interpret user queries and match them with appropriate products or services. With natural language processing methods, query understanding and expansion techniques are playing an important role to bridge the semantic gap between queries and product information during this procedure, allowing the system to understand complex, ambiguous, or conversational queries from users. For example, users might search for “affordable running shoes for winter,” which requires the platform to parse the query, infer user intent (i.e., shoes for running in cold weather), and prioritize products based on pricing and seasonal relevance. Semantic search in e-commerce techniques go beyond keyword matching by understanding the user query and correlations between key entities, enabling more context-aware results. Search engines in e-commerce also rely on machine learning models that take into account user intent, contextual data, and preferences to deliver highly relevant search results. Additionally, search engines in e-commerce must address unique challenges like scalability and diversity. With product catalogs growing rapidly, search engines must efficiently process and rank millions of items in real-time. Advanced ranking algorithms, often powered by machine learning, play a vital role in this process, optimizing for both relevance and user engagement metrics such as click-through rates or conversion rates.

E-commerce recommendations. IR in e-commerce is increasingly intertwined with recommender systems. Recommender systems are a cornerstone of e-commerce platforms, helping users discover products they might not have explicitly searched for but are likely to find appealing. These systems predict user preferences using collaborative filtering, content-based filtering, or hybrid methods that combine both approaches. While search engines retrieve items explicitly requested by

the user (i.e., through queries), recommendations anticipate user needs by suggesting products the user may not have thought to search for. These systems analyze user data, such as browsing history, purchase patterns, and interactions, to suggest relevant items. They typically use a combination of techniques, including collaborative filtering, which recommends products based on the behaviors of similar users, and content-based filtering, which suggests items with attributes similar to those the user has shown interest in. Many modern systems employ hybrid models that integrate both types of method, sometimes enhanced with techniques like deep learning, to improve accuracy and diversity. By delivering personalized recommendations, these systems not only enhance the user experience but also drive business goals by increasing engagement, conversion rates, and customer satisfaction, all while introducing users to new products that may surprise or delight them.

Recommendation systems in e-commerce analyze user data and behavioral patterns to suggest products or services that users are likely to be interested in. These systems use various algorithms, including collaborative filtering, content-based filtering, and hybrid approaches. In the context of e-commerce, recommender systems must balance relevance, diversity, novelty, and serendipity to enhance user engagement and satisfaction. By using a division into candidate retrieval and reranking stages, these systems often operate in two modes: personalized recommendations (based on individual profiles and history) and non-personalized recommendations (based on overall product popularity or trends). Research in recommender systems for e-commerce involves improving recommendation algorithms, addressing challenges like cold-start users, and optimizing recommendations for business goals such as conversion rates and customer retention.

In summary, information retrieval is foundational to the search and recommendation functions within e-commerce platforms. The convergence between search and recommendation highlights the importance of IR techniques that can balance precision (retrieving highly relevant products) with recall (offering a broader set of options that might interest the user). Hybrid models that combine collaborative filtering, content-based filtering, and neural IR approaches are commonly employed to

address this dual need. The integration of advanced IR techniques, such as search and recommendation models, allows e-commerce platforms to deliver highly personalized, efficient, and contextually relevant user experiences, ensuring that users find the products they want – and even those they did not know they wanted.

2.4 Conversational AI

Conversational artificial intelligence (AI) techniques refer to technologies that enable natural, human-like interactions between users and machines through conversational communication. These interactions can be categorized into single-turn and multi-turn scenarios, corresponding to question-answering systems and dialogue systems, respectively. During the interactions, conversational AI aims to understand user input, process context, and generate meaningful, human-like responses. Conversational AI is used in various applications, such as virtual assistants, customer support, and personal productivity tools. These systems can handle simple queries as well as complex, multi-turn conversations, adapting to user needs and improving over time through continuous learning. By mimicking human conversation patterns, conversational AI allows for more intuitive and accessible interactions, making it a valuable tool for enhancing communication between users and machines.

In e-commerce, chatbots and QA services powered by conversational AI can help users find relevant products, provide recommendations, and even complete transactions seamlessly. By offering a more engaging and interactive way for customers to interact with e-commerce platforms, conversational AI improves user satisfaction, increases engagement, and reduces the friction often associated with traditional search and navigation methods.

E-commerce question-answering. Question-answering (QA) systems are designed to deliver direct and precise responses to user questions, improving both user satisfaction and decision-making efficiency. Recently, e-commerce platforms have started to provide question-answering services. E-commerce QA systems help to enhance user experiences by enabling customers to obtain relevant, concise, and accurate answers to their product-related queries. These systems typically understand user

intent and either retrieve answers from product descriptions, reviews, and FAQs or generate responses dynamically using advanced models. E-commerce question-answering refers to the process in which users ask product-related questions on an e-commerce platform, and the system provides answers either from knowledge bases, reviews, or user-generated content. E-commerce QA systems use both retrieval-based and generative models to match or generate appropriate answers. This helps users make informed decisions based on product descriptions, user reviews, and frequently asked questions (FAQs). The goal is to reduce information overload and improve the user experience by providing relevant and concise answers.

E-commerce QA systems must handle a wide variety of queries ranging from simple fact-based questions (e.g., “What is the price of this product?”) to more complex inquiries about product specifications, reviews, or usage (e.g., “Is this laptop suitable for gaming?”). To address this diversity, QA systems often incorporate a mix of retrieval-based approaches, which search for relevant information in structured data or knowledge bases, and generative approaches, which generate answers when information is sparse or not directly available. Additionally, many e-commerce platforms enable community-based QA, where previous buyers or users of a product can contribute answers, further enriching the system’s knowledge base. The integration of QA systems into e-commerce portals helps reduce the friction often associated with product discovery and decision-making. By offering immediate answers to user queries, these systems improve the overall shopping experience, increase user engagement, and can positively impact conversion rates. E-commerce QA is a rapidly evolving field with ongoing research aimed at improving the accuracy, efficiency, and personalization of responses.

E-commerce automatic dialogue systems. Automatic dialogue systems aim to engage in natural, human-like conversations with users and are widely used in applications such as customer support, virtual assistants, and e-commerce. They provide personalized, efficient, and engaging interactions, enhancing the overall user experience. In e-commerce, automatic dialogue systems, often in the form of chatbots or voice assistants, enable natural language interactions between users and

platforms. They can support multi-turn interactions, where users ask follow-up questions, refine their preferences, or seek assistance, creating a more engaging and personalized shopping experience. These systems assist users in discovering information, finding products, completing purchases, and sharing their opinions, all through conversational interfaces. Systematically, automatic dialogue systems in e-commerce refers to the use of chatbots and virtual assistants that can simulate a human conversation to assist users in finding products, answering inquiries, and facilitating transactions. These systems use natural language processing and machine learning to provide timely and relevant assistance.

Research in conversational AI for e-commerce focuses on improving dialogue understanding, response generation, context retention across sessions, and user satisfaction. Additionally, conversational AI systems must adapt to diverse user needs and accommodate various languages and cultural contexts, making this an evolving area of study.

3

E-commerce Presentations and Users

E-commerce presentations are composed of a series of user-facing components in e-commerce portals, e.g., various pages about items and categories, titles of items, user comments on item pages, search bars, and recommendation list. Such functions provided by e-commerce portals are meant to enable interactions with users on e-commerce platforms. E-commerce users possess unique characteristics. There are multiple types of user behavior and feedback on an e-commerce platform, e.g., search, clicks, add-to-carts, purchases, returns, comments, and discussions with retailers. These unique characteristics of e-commerce users provide a rich source of information about the successes and failures of e-commerce platforms in helping users discover the items they need.

We divide this section into two parts: e-commerce presentations (Section 3.1) and e-commerce users (Section 3.2). In Section 3.1, we first introduce basic concepts of e-commerce interfaces; then we detail studies that analyze different aspects of e-commerce presentations, i.e., title analysis, item information analysis, and review analysis. In Section 3.2, we list characteristics of e-commerce users, and examine user behavior on e-commerce portals, i.e., macro behavior, micro behavior and cross-platform behavior.

3.1 E-commerce Presentations

In this section, we cover two aspects of e-commerce presentations: (i) basic concepts and types of e-commerce interface (Section 3.1.1), and (ii) e-commerce presentation analysis (Section 3.1.2).

3.1.1 Basic concepts

An *interface* refers to an interactive component of a webpage or an application (Hearst, 2009). Referring to all interactive components on e-commerce portals (e.g., search bars, navigation panels, lists of recommended items, item titles, and user reviews), e-commerce interfaces play an invaluable role for the e-commerce user experience. E-commerce interfaces dramatically impact the performance of an e-commerce platform. Depending on the nature of the stakeholders involved, most e-commerce sites can be divided into four types of business: *business-to-business*, *business-to-consumer*, *consumer-to-consumer*, and *consumer-to-business* (Nemat, 2011):

- **B2B: Business to Business** This type of e-commerce business focuses on electronic transactions of goods or services between two corporations, i.e., one company uses the e-commerce site to sell items to another company. Figure 3.1(1) shows an example of the interface used in a B2B setting.
- **B2C: Business to Consumer** B2C refers to scenarios where businesses directly sell items to consumers. Most online shopping platforms, such as Amazon, Booking.com, and JD.com belong to this type of business. Figure 3.1(2) shows an item page from Amazon as an example of the interface used by B2C businesses.
- **C2B: Consumer to Business** Instead of a business retailing items to consumers, C2B sites such as UpWork¹ cater for a scenario where consumers provide services to businesses. Figure 3.1(3) provides a screenshot from Upwork as an example of a C2B interface.
- **C2C: Consumer to Consumer** C2C refers to a type of e-commerce business where both retailers and buyers are consumers,

¹<https://www.upwork.com>

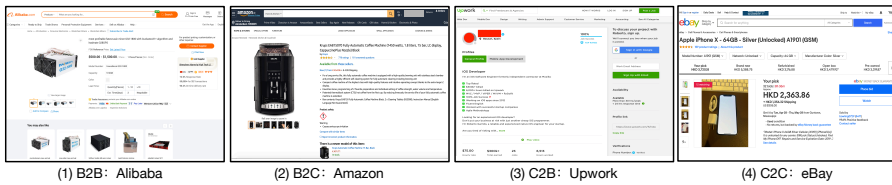


Figure 3.1: Four types of e-commerce businesses examples. Image sources: Alibaba.com, Amazon.com, UpWork.com, and EBay.com.

while the C2C site itself benefits from commission fees that are normally paid by the seller. eBay is a well-known example of C2C business. Figure 3.1(4) lists a screenshot from eBay as an example of a C2C business interface.

Like other web interfaces, e-commerce interfaces are evaluated in terms of user satisfaction (Vergo *et al.*, 2002). Thus, different communities of users are usually catered for with different interfaces that are designed to accommodate for their tastes and shopping interests. Three ingredients are shared by virtually all e-commerce interface designs:

- **Navigation options**, which refer to elements that help users reach a certain part of the e-commerce platform, e.g., search bars, paginations, and universal menus;
- **Input options**, which are elements of an e-commerce platform for which the user provides input from their end, e.g., search bars, checkboxes, dropdown lists, dropdown buttons, toggles, and other text fields; and
- **Information components**, which are composed of various types of information about the products or services listed on e-commerce platforms, such as search results, recommendation results, item titles, images, item information, question answer pairs, user reviews, and tooltips.

As we dive into the problem of information discovery in e-commerce, information components are our main focus in this section. We find that almost every e-commerce site provides six information components: *search results*, *recommendation results*, *item titles*, *item features and descriptions*, *question answer pairs*, and *user reviews of the item*. In Section 3.1.2, we describe studies on information interface analysis of these six components.

3.1.2 Analyzing information components

Many studies have been devoted to analyzing the effect of different information components. In this section, we summarize studies that focus on search results, recommendation results, titles, item descriptions, question answering, and reviews, respectively.

3.1.2.1 Search results in e-commerce

For all e-commerce information components, search and recommendation are the two main tasks in most of e-commerce platforms. E-commerce search engines are often the starting points for many online consumers (Wu *et al.*, 2018a). E-commerce sites typically feature two-stage search interfaces. As shown in Figure 3.2, in an e-commerce search session,² a consumer first searches using a query, leading to a result page, and then selects an item to click on the result page; after that, the user decides whether or not to purchase the item by examining its detailed description on the so-called item page.

E-commerce search engines provide category options with the search bar. During the early development of e-commerce search, interfaces of different types have been considered, e.g., devoted type, divided type, co-existing type, and multi-page type (Lu *et al.*, 2006). But with the development of e-commerce search, these types of interfaces have been blended by e-commerce platforms. Currently, a typical e-commerce search system includes three main components: query processing, candidate retrieval and ranking (Zhang *et al.*, 2020a). In query processing, the search engine rewrites a query from the user into a term-based representation that can be processed by downstream components. In the candidate retrieval stage, the system uses the inverted index to retrieve candidate products to match queries. Finally, the ranking component orders the retrieved candidates based on factors such as relevance, and predicted conversion ratio. We will discuss research into the principles and strategies of all three components in Section 5 in more detail.

²As defined in web-based search engines, a search session refers to all queries made by a user in a particular time period with a consistent underlying user need (Eickhoff *et al.*, 2014).

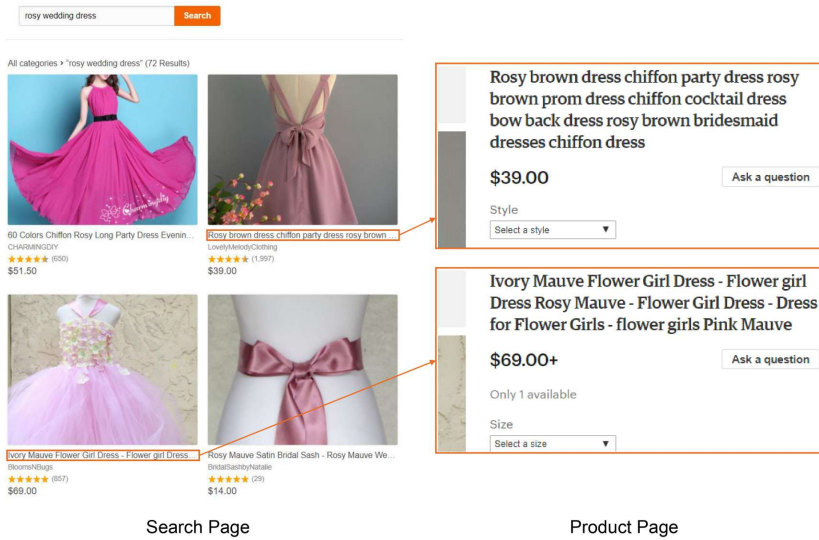


Figure 3.2: Illustration of a sample search session in an e-commerce platform. The query is “rosy wedding dress,” and the search result page is shown on the left and a portion of the item page for two items is shown on the right. This search session consists of two stages: (i) selecting an item to click from a ranked list, and (ii) deciding whether to purchase the item by reading its detailed description. Image source: Wu *et al.* (2018a).

3.1.2.2 Recommendation results in e-commerce platforms

For many e-commerce platforms, recommendations have become the most important service to help users find their needed items. E.g., recommendations have been reported to contribute to the majority of both revenue and traffic in Taobao (Wang *et al.*, 2018b), where one billion users can be connected to two billion items. To this end, the homepage on the mobile Taobao app is generated based on consumers’ past behavior via recommendation algorithms, as illustrated in Figure 3.3. Figure 3.3 shows three recommendation areas displayed on the homepage: a list of recommendation interfaces, a “popular products” list, and a promotion list, respectively. Each recommendation area is provided based on users’ past behaviors with recommendation strategies. As user behavior varies between scenarios, the recommendation strategy also needs to consider specific patterns and user preferences specific for each

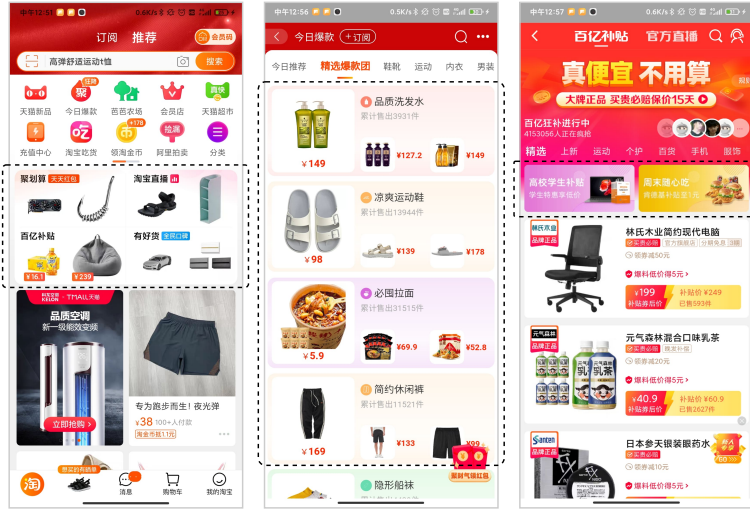
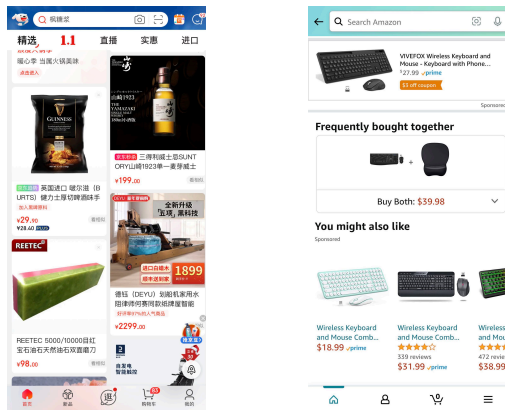


Figure 3.3: E-commerce recommendation scenarios in Taobao. The areas highlighted with dashed rectangles are personalized for users. Images and textual descriptions are also generated for better user experience. Image source: Wang *et al.* (2018b).

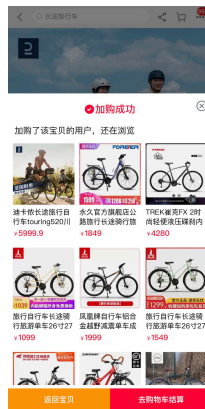
recommendation scenario. For example, on the item page, the recommendation strategy needs to provide either relevant or similar items to the item that the user is focusing on, whereas the recommendation list on the home page shows the recommendation results considering the user’s personalized preferences (Zhou *et al.*, 2018e).

Different types of recommendation results may be shown at different stages of a customer’s. Examples include “substitutes” (see Figure 3.4(a)) and “complementary items” before and after the user adds a product to their cart (see Figure 3.4(b) and 3.4(c), respectively). Once the consumer clicks a recommended product, the system will automatically jump to the product detail page, which includes product titles, product descriptions, categories, ratings, and reviews. We will discuss more details about strategies and technologies of e-commerce recommendation in Section 6.



(a) JD.com

(b) Amazon



(c) Tmall.com

Figure 3.4: Recommendation results exposed to users in three e-commerce platforms. Image sources: [JD.com](#), [Amazon.com](#), and [Tmall.com](#).

3.1.2.3 Product titles in e-commerce platforms

Product titles and their images are uploaded by suppliers to showcase their items. As most e-commerce platforms at least provide search and recommendation services based on information in the titles, retailers have applied many search engine optimization strategies to titles (Ledford, 2015). As a result, lots of item titles are lengthy, over-informative, and sometimes incorrect. Figure 3.5(b) provides an example from Tmall,



Figure 3.5: Given a query “floral-dress long sleeve women” on Tmall, the complete title cannot be displayed in the search result page unless the user proceeds to the detail page further. Image source: Wang *et al.* (2018a).

the largest B2C online shopping platform in China, where the item title is composed of more than 30 Chinese words. But when a customer browses an item on Tmall Apps, fewer than 10 Chinese words can be displayed due to screen size limitations (Figure 3.5(a)). Thus, lengthy and verbose titles are inconvenient for mobile e-commerce users to search items on e-commerce platforms. Similarly, it has been reported that item titles with less than 80 characters improve the shopping experience on Amazon, because these shorter titles make it easier for customers to find products.³ Accordingly, research on e-commerce title analysis mainly focuses on obtaining effective compression or summaries of lengthy item titles for e-commerce search.

Item title compression, also called short title extraction (Gong *et al.*, 2019), is meant to extract sufficient words from lengthy and verbose titles to produce a succinct new title to improve the user experience on

³<https://sellercentral.amazon.com/forums/message.jspa?messageID=2921001>

mobile devices (Wang *et al.*, 2018a). Inspired by neural extractive document summarization methods (Ren *et al.*, 2017), item title compression methods apply neural networks to weight the importance of each word in the item title. Gong *et al.* (2019) introduce a feature-enriched neural extractive model to extract short titles. Specifically, the authors apply a recurrent neural network as a sequential classifier with three types of features: content, attention, and semantics respectively. By using user search logs as external knowledge, Wang *et al.* (2018a) construct a multi-task learning approach for improving item title compression. The proposed method is composed of two seq2seq components which share an identical encoder. The authors combine these two components with an overall pointer neural network (Vinyals *et al.*, 2015) to automatically select the most informative words from the given item title.

Pointer neural networks easily omit key information. To tackle this problem, Sun *et al.* (2018a) introduce a multi-source pointer network model, named the multi-source pointer network (MS-Pointer), by considering two extra constraints: (i) irrelevant information reduction; and (ii) the key information retainment. Figure 3.6 provides an overview of MS-Pointer, with two encoders. In MS-Pointer, in addition to the encoder for the source title, the authors add another knowledge encoder that uses an LSTM to embed the brand name and the commodity name. As shown in Figure 3.6, MS-Pointer combines the original title “Nintendo switch console. . .” and background knowledge “brand name: Nintendo”, and then it generates the short title about the item “Nintendo switch”. More recently, Fetahu *et al.* (2023) have proposed an instruction fine-tuning strategy to summarize product titles according to various criteria such as the number of words in a summary or the inclusion of specific phrases.

The task of title generation has been proposed to extend the task of title compression into a text generation problem. Unlike title compression, which only extracts words from item titles, the task of title generation is to *generate* a short item title so as to address the problem of inaccurate item titles in e-commerce (Zhang *et al.*, 2019a). To generate a succinct and accurate short title from a long source title, Zhang *et al.* (2019a) offer a multi-modal generative adversarial network, named MM-GAN, which addresses the title generation task as a reinforcement

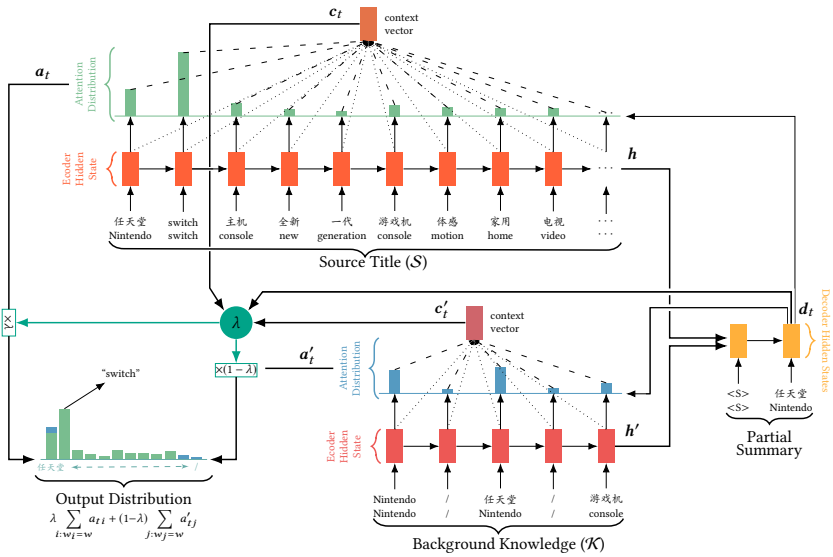


Figure 3.6: Multi-source pointer network (MS-Pointer) with two encoders for item title compression. MS-Pointer copies words from two encoders. At each decoding time step, a soft gating weight $\lambda \in [0, 1]$ is calculated to weight the probability of words from the source title, versus words from the background knowledge. The final output distribution is the weighted sum of attention distributions a_t and a'_t . Image source: Sun *et al.* (2018a).

learning problem. MM-GAN is composed of two main components, a title generator and a discriminator (Figure 3.7). Given the source title and its corresponding tags or features, the generator applies an LSTM-based network to generate a short item title. The discriminator, i.e., a binary classifier, distinguishes whether the generated short titles are human-generated or machine-generated. Thus, an adversarial learning procedure is constructed, in which the quality of the short title depends on its ability to fool the discriminator into believing it is a human-generated one, and the output of the discriminator is a reward for the generator to improve the generation performance. Recently, scene marketing has become a new marketing mode for product promotion where scene scenarios are created to demonstrate product functions (Zhao, 2020). To help the e-commerce system find scene topics, Lin *et al.* (2022) propose a topic generation method to generate scene-based titles in e-commerce.

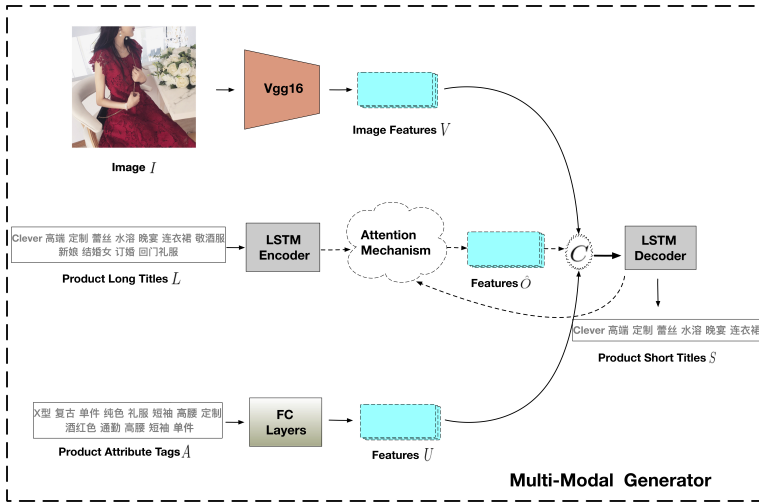


Figure 3.7: Overall framework of the MM-GAN model for short item title generation. Image source: Zhang *et al.* (2019a).

3.1.2.4 Product descriptions in e-commerce platforms

As shown in Figure 3.8, many e-commerce platforms provide a short description for each item so as to showcase the features of the item. As an important factor in content marketing, the item description is key for increasing consumer engagement. During the early years of e-commerce, item descriptions were usually written or edited by human copywriters. However, the availability of an increasing number of items in e-commerce makes this manual process too costly. Moreover, with the development of virtual assistants in e-commerce, such as Alexa and Tmall Genie, there is a growing demand for automatically generating a short description given item attributes. To address this demand, the task of item description generation has been proposed. Item description generation needs to generate an item’s description from a series of complicated attributes. Wang *et al.* (2017b) detail a statistical framework to weight the relative importance of the attributes of an item and to maintain accuracy at the same time. In Figure 3.9 we specify the framework of the proposed item description model. By combining sentence-level templates extracted from the input data with knowledge from a pre-trained dataset, the

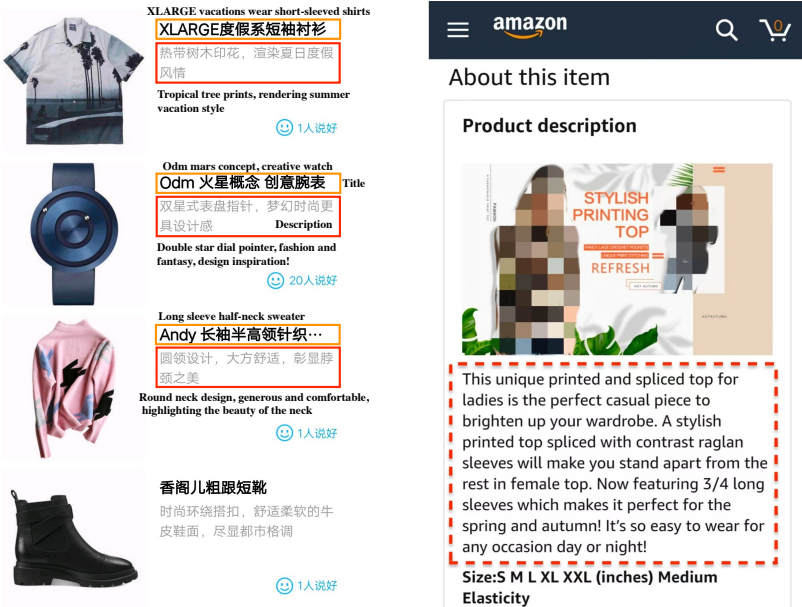


Figure 3.8: Item descriptions are widely used in e-commerce platforms, e.g., (a) Taobao and (b) Amazon. Image source: Zhang *et al.* (2019c).

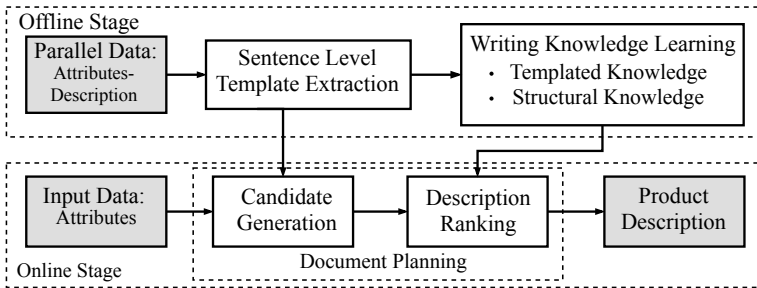


Figure 3.9: Overall framework for item description generation with pretrained writing knowledge. Image source: Wang *et al.* (2017b).

authors generate and rank candidate item descriptions through an online *document planning* stage.

Unlike early studies that focused on generating item descriptions purely from the item’s attributes, Zhang *et al.* (2019c) generate a pattern-controlled item description from multiple features, e.g., titles

and item categories. Based on the copy mechanism (Gu *et al.*, 2016), the authors propose a pattern-controlled pointer-generator network (PGPCN) to generate the description. In PGPCN, a transformer is applied in the encoder component, whereas the decoder is used to control the pattern (e.g., the category, the length, and the style of the description) of the item.

It is important that the descriptions generated for item description are grounded in facts. To generate a fact-based description, Chan *et al.* (2019) offer an encoder-decoder framework, called the fidelity-oriented product description generator (FPDG), by searching key information from keyword labels. The authors establish semantic connections between item keywords and the generated product description. As shown in Figure 3.10, FPDG has two main components: (i) a keyword encoder that stores the word and its entity label in the token memory and self-attention modules, and (ii) an entity-based generator that generates an item description based on the memory and self-attention modules.

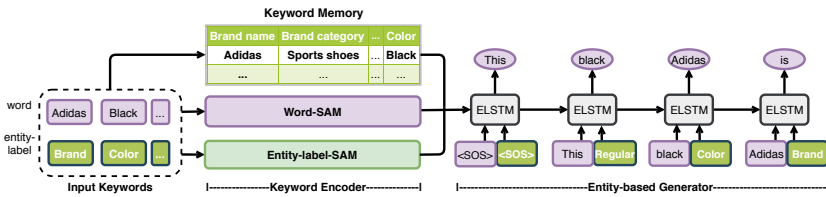


Figure 3.10: Overview of the fidelity-oriented item description generator. The whole model is divided into two components: (i) a keyword encoder, and (ii) an entity-based generator. Image source: Chan *et al.* (2019).

Personal interest is neglected by all of the above approaches that generate descriptions given attributes or keywords. To address this shortcoming, Chen *et al.* (2019d) propose a knowledge-based personalized item description generation strategy. The authors extend the encoder-decoder framework (Sutskever *et al.*, 2014) to a sequence modeling formulation using a self-attention mechanism. A large variety of item attributes, including the target user's personalized preference features, are combined in an attribute fusion component through multi-layer attention mechanisms; retrieved external knowledge is incorporated in a *knowledge incorporation* component. In Figure 3.11, we provide an example of the knowledge-based personalized item description generation.

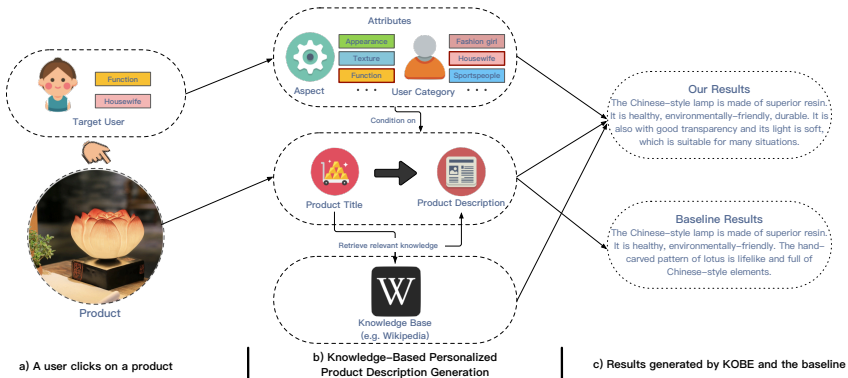


Figure 3.11: An example of knowledge-based personalized item description generation. The example is divided into three parts: (a) a user clicks an on item; (b) knowledge-based personalized item description generation; and (c) results generated by Chen *et al.* (2019d)’s proposed method and a baseline. Image source: Chen *et al.* (2019d).

3.1.2.5 Question answering in e-commerce

Question-answering (QA) systems are designed to provide direct and concise answers to user queries by understanding the intent behind a question and retrieving or generating the most relevant information. QA systems aim to deliver specific answers from structured or unstructured data sources, such as web documents or knowledge bases. QA systems are essential across a variety of domains, including web search, customer support, where users seek quick, accurate, and contextually relevant information (Radev *et al.*, 2002; Tapeh and Rahgozar, 2008; Yin *et al.*, 2016). To increase the number of sales, most e-commerce portals provide a QA service to facilitate the customers’ shopping procedure by answering their questions about products (Gao *et al.*, 2019b; Feng *et al.*, 2021). Currently, on many e-commerce sites, a user can ask a question about a product, and the QA system allows some users (e.g., customers who bought this product) to provide answers (Gao *et al.*, 2019b). In Figure 3.12, we show examples of question-answering services at Amazon and JD.com. More detailed discussions of e-commerce QA are provided in Section 7.1.



(a) Question answering service on Amazon (b) Question answering service on JD.com

Figure 3.12: Question answering services are widely applied in e-commerce platforms, (a) Amazon and (b) JD.com. Image sources: [Amazon.com](https://www.amazon.com) and [JD.com](https://www.jd.com).

3.1.2.6 User reviews in e-commerce

User reviews serve as a type of reliable information about the quality of items on e-commerce platforms. User reviews have been shown to play an essential role in determining user preference (Liang *et al.*, 2015; Huebner *et al.*, 2018; Li *et al.*, 2018b). In this section, we introduce methods for review analysis in e-commerce. Research on review analysis can be organized into four key components: (i) sentiment classification, (ii) helpfulness prediction, (iii) review summarization, and (iv) review generation.

(i) Sentiment classification in reviews. The *sentiment classification* task is to label a given text with a specific opinion label. It has received lots of attention during the past two decades (Pang *et al.*, 2002; Go *et al.*, 2009; Pan *et al.*, 2010; Kamal *et al.*, 2012; Tang *et al.*, 2014; Pontiki *et al.*, 2015; Tsytarau and Palpanas, 2016; Sun *et al.*, 2017). Work on sentiment classification in an e-commerce context has attempted to capture a user’s opinion about a specific item from reviews on an e-commerce platform (Sun *et al.*, 2017; Tang *et al.*, 2015; Xia *et al.*, 2015; Tripathy *et al.*, 2016; Li *et al.*, 2018d). Traditional approaches to sentiment classification focus on the classification problem given textual attributes of an item (Pang *et al.*, 2002), while largely ignoring the relation between users and the item. To address this problem, Tang *et al.*

(2015) introduce a neural network, the user-product neural network (UPNN), to incorporate user and item information into a document-level sentiment classification procedure. In particular, the authors jointly embed the user preference information, i.e., ratings and reviews, and item attributes. Then, a convolutional neural network is used to predict the sentiment label of the target review.

Sentiment classification using the relation between users and products or services faces two important challenges: (i) the sparseness of user-item interactions, and (ii) the information in user embedding methods. Chen *et al.* (2016a) present a fine-grained hierarchical neural network model to incorporate global user and item information into sentiment classification. Unlike many sentiment classifiers that use convolutional neural networks, the authors apply a hierarchical LSTM (Gers *et al.*, 1999) to jointly generate sentence-level representations and document-level representations. Then, user-item interaction information is applied as attention over various regions of a document to enhance the sentiment classification. Wu *et al.* (2018c) distinguish between different roles of words and sentences in user reviews: to describe the user's preferences or to describe an item's characteristics. To distinguish between these roles, the authors put forward an attention-based hierarchical neural network model to embed user and item information to generate two text representations with user attention or item attention, respectively. Fei *et al.* (2021) use fine-grained latent opinion knowledge into the sentiment classification process by using a variational reasoning method.

(ii) Helpfulness prediction. Given the fact that an item can be commented on by hundreds of thousands of consumers, the quality of reviews in e-commerce varies considerably and not all reviews are helpful. To gain insights from helpful reviews, the task of review helpfulness prediction has attracted attention from both academia and industry (McAuley and Yang, 2016). Early studies on review helpfulness prediction employ feature-aware methods, where multiple types of features, such as structural features (Kim *et al.*, 2006; Susan and David, 2010; Xiong and Litman, 2011), emotional features (Martin and Pu, 2014), semantic

features (Yang *et al.*, 2015b), argument features (Liu *et al.*, 2017), and lexical features (Xiong and Litman, 2014), are successfully applied.

Motivated by the progress of deep neural networks, Fan *et al.* (2018) introduce a multi-task neural learning (MTNL) architecture for identifying helpful reviews. Chen *et al.* (2018a) propose a CNN-based neural network with multi-granularity (i.e., character-level, word-level, and topic-level) features for helpfulness prediction. Fan *et al.* (2019a) suggest an end-to-end deep neural architecture to capture the intrinsic relationship between the meta-data of an item and its numerous comments that could be beneficial to discover the helpful reviews. Multi-modal data has become increasingly popular in online reviews. To analyze multi-modal reviews, Liu *et al.* (2021c) introduce a multi-modal review helpfulness prediction task that is aimed at exploring multi-modal clues for review helpfulness prediction. The authors describe an item-review coherent reasoning module to capture the intra- and inter-modal coherence between the target item and the review. Han *et al.* (2022) put forward a selective attention approach, including probe mask generation and mask-based attention computation, for the multi-modal review helpfulness prediction problem. To mine the mutual information of cross-modal relations in the input, Nguyen *et al.* (2022) propose an adaptive cross-modal contrastive learning mechanism, with a multi-modal interaction module to correlate modalities' features.

(iii) Review summarization. Given a set of user reviews, the task of *review summarization* is to extract the main information from the reviews. Similar to multi-document summarization, review summarization summarizes a set of item reviews for a single item. Approaches to review summarization can be divided into two: feature-aware methods and aspect-aware methods.

Feature-aware methods are inspired by previous document summarization methods: Yang *et al.* (2010) detail a feature-based item review summarization method to satisfy the detailed information needs of customers. To jointly summarize reviews and predict ratings in a mobile environment, Liu *et al.* (2012) offer a latent semantic indexing based approach to extract features and attributes from user reviews and ratings. For new items without reviews, a probabilistic retrieval method is pro-

posed to extract relevant opinion features from other items to describe the item information (Park *et al.*, 2015). Another important part of review summarization concerns aspect extraction from user reviews (Chen *et al.*, 2014; Angelidis and Lapata, 2018; Bražinskas *et al.*, 2020), where target entities and aspects need to be extracted from opinionated text. Chen *et al.* (2014) extract prior knowledge automatically from user reviews and propose a fault-tolerant model to extract aspects guided by the knowledge. Category hierarchy information is combined with a topic model to improve the performance of aspect extraction (Yang *et al.*, 2017b). By jointly considering fine-grained aspect-topic-sentiment connections, Tan *et al.* (2017) propose a generative topic aspect sentiment model.

With the development of deep learning, item review summarization has been tackled from a range of perspectives (Ly *et al.*, 2011; Wu *et al.*, 2016a). For instance, to tackle the weakness of the “bag of words” assumption, He *et al.* (2017a) propose an unsupervised neural network model. Considering dependencies between adjacent words, the authors used an embedding method with attention mechanism to de-emphasize saliency and extract aspects. Angelidis and Lapata (2018) describe a weakly supervised neural framework for the identification and extraction of salient customer opinions that combines aspect and sentiment information. Using a small number of annotated instances with a large-scale unlabeled corpus, Bražinskas *et al.* (2020) suggest a few-shot learning framework for generating an abstractive summary. In recent years, pre-trained language models (Vaswani *et al.*, 2017; Kenton and Toutanova, 2019) have been shown to be effective in document summarization (Liu and Lapata, 2019). A domain-specific generative pre-training method, PEGASUS, has been proposed to address the e-commerce review summarization problem (Zhang *et al.*, 2021c). Inspired by Vector-Quantized Variational Auto-encoders (VQ-VAE) (Oord *et al.*, 2017), Angelidis *et al.* (2021) explain an unsupervised neural model, Quantized Transformer (QT), that uses a clustering interpretation of the quantized space to discover popular opinions among hundreds of reviews. To tackle challenges such as a lack of cross item diversity and consistency, Oved and Levy (2021) offer a method that uses strong pre-trained language models.

(iv) Review generation. The task of *review generation* has been proposed to understand how a specific user provides comments on items (Dong *et al.*, 2017). Unlike review summarization, where one extracts salient sentences or generates abstractive summaries, the task of *review generation* is to generate sentences as user reviews to represent users' intention. Sequence-to-sequence (seq2seq) neural networks have been applied to automatically generate text (Sutskever *et al.*, 2014). However, it is difficult to directly use traditional seq2seq models to generate reviews because of the following challenges: (i) the presence of unknown factors renders the generation process non-deterministic, and (ii) both implicit and explicit information need to be handled, which makes it difficult to decode reviews.

To address these problems, Dong *et al.* (2017) propose an attention-enhanced attribute-to-sequence model to generate item reviews for given attribute information. The authors introduce an attention-enhanced attribute-to-sequence model that learns to encode attributes into vectors and then uses a recurrent neural networks based on LSTM units to generate reviews by conditioning on the encoding vectors; see Figure 3.13. The model can be divided into three components: an attribute encoder, a sequence decoder, and an attention mechanism. The authors use a dataset collected from Amazon to verify the effectiveness of the proposed method, especially the attention mechanism in the review generation procedure.

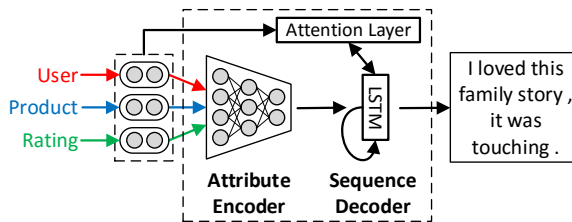


Figure 3.13: Overview of the attention-enhanced attribute-to-sequence model for review generation. Image source: Dong *et al.* (2017).

There is increasing attention for combining preference prediction with review generation. As sentiment classification plays an important role in e-commerce review analysis, Radford *et al.* (2018) describe a

representation learning strategy to detect opinions while generating reviews, where a generic sentiment tree bank was applied to represent the sentiment label in user reviews (Socher *et al.*, 2013). Many e-commerce sites provide structured information, such as aspect-sentiment scores, i.e., each review text contains sentences describing a number of aspects of the item. Focusing on generating long Chinese reviews from aspect-sentiment scores, Zang and Wan (2017) offer end-to-end sequential review generation models (SRGMs). Unlike traditional seq2seq models, SRGMs encode inputs of aspect-sentiment scores using multi-layer perceptrons. Sharma *et al.* (2018) propose an LSTM-based neural network to generate personalized reviews from multi-faceted factors, i.e., user profiles and item attributes, where an additional loss term is used to ensure consistency of the sentiment rating in the generated review.

Ni *et al.* (2017) put forward a collaborative-filtering generative concatenative network to jointly optimize item recommendation and generate personalized reviews. To generate personalized high-fidelity reviews, Ni and McAuley (2018) come up with an encoder-decoder model to use both user and item information as well as auxiliary, textual input and aspect-aware knowledge, where an attention fusion layer is introduced to control the influence of various encoders.

Some e-commerce sites have launched an interaction box called *tips* on their mobile platforms. Figure 3.14 shows examples of reviews and tips on Yelp. The left column is the review from the user “Monica H.”, and tips from several other users are shown in the right column. Tips are more concise than reviews and can reveal user experience, feelings, and suggestions with only a few words. To generate concise tips from reviews, Li *et al.* (2017b) suggest a multi-task learning framework for tip generation and rating prediction. For abstractive tip generation, gated recurrent neural networks are employed to decode user and item latent factors, whereas for rating regression, a multilayer perceptron network is employed to project user latent factors and item latent factors into ratings. A persona-aware tip generation framework has been put forward for *personalized* tip generation through adversarial variational auto-encoders (aVAE) (Li *et al.*, 2019b).

Opinion tags refer to a ranked list of tags provided by the e-commerce platform that reflect the characteristics of reviews of an item; see

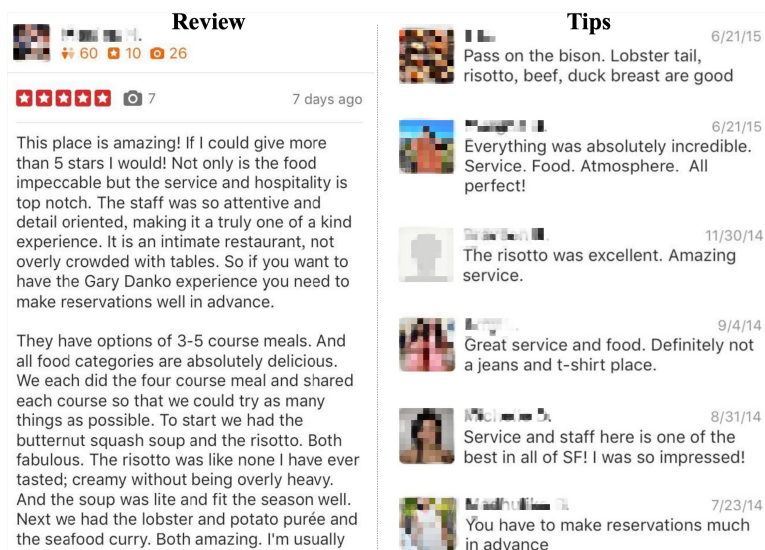


Figure 3.14: Examples of reviews and tips selected from the restaurant “Gary Danko” on Yelp. Users will get conclusions about this restaurant immediately after scanning the tips with their mobile phones. Image source: Li *et al.* (2017b).

Figure 3.15. To assist consumers to quickly grasp a large number of reviews about an item, opinion tags are increasingly being applied by e-commerce platforms. Current mechanisms for generating opinion tags rely on either manual labelling or heuristic methods, which is time-consuming and ineffective. Li *et al.* (2021c) introduce the abstractive opinion tagging task, where systems have to automatically generate a ranked list of opinion tags that are based on, but need not occur in, a given set of user-generated reviews.

The abstractive opinion tagging task comes with three main challenges: (i) the noisy nature of reviews; (ii) the formal nature of opinion tags vs. the colloquial language usage in reviews; and (iii) the need to distinguish between different items with very similar aspects. To address these challenges, Li *et al.* (2021c) come up with an abstractive opinion tagging framework, named AOT-Net, that first predicts a salience score for each review, and given the salience scores, it groups all reviews into opinion clusters and ranks opinion clusters by cluster size. With the

<p>Reviews:</p> <p>The waitress was extremely attentive and even gave us a free fried man tou dessert that came with condensed milk for dipping... I love it !! A more expensive meal but extremely satisfying and it was money wellspent. The dish and the lamb are no-limited. I love this place. Food is delicious and reasonably priced. If you're around, go here - you deserve it! All in all; was a great experience and the service is really above and beyond. The shrimp was fresh and the pork mixture was tasty. Everything was delicious! The dumplings's was thin and it was very juicy. They are consistent at each location with their great service. The environment is very tidy and clean. And the service was good though. Many lamb portions are eaten unlimitedly! Fairly quick and polite service. It worth that price!</p>
<p>Opinion tags: hospitable service, delicious food, value for money, ample food, clean environment.</p>

Figure 3.15: An example of a set of reviews and their corresponding opinion tags. Image source: Li *et al.* (2021c).

designed alignment feature and alignment loss, AOT-Net sequentially reads ranked opinion clusters and generates opinion tags with ranks. To generate opinion tags in a personalized way, Zhao *et al.* (2022b) select the information that users are interested in from reviews and then generated a ranked list of aspect and opinion tag pairs. The authors track user preferences not only using explicit feedback, i.e., reviews, but also using implicit feedback such as clicks and purchases in a heterogeneous graph neural network model.

3.2 E-commerce Users

Over 55% of online customers start to search on an e-commerce website as opposed to a generic web search engine (Zhou *et al.*, 2018e). Besides desktop clients, there are multiple e-commerce environments, e.g., mobile apps, smart watches, and interactive systems. These devices provide new means of interaction for users with e-commerce interfaces. User behavior on e-commerce platforms can be divided into two types: implicit feedback and explicit feedback (Brown *et al.*, 2003; Su *et al.*, 2018b). Implicit

feedback is captured in transaction logs and includes clicks, purchases, browses, and engagements, etc.; explicit feedback of online shopping is captured in user comments, chat logs, and questions. Following Lo *et al.* (2016), Zhou *et al.* (2018e), Gao *et al.* (2019b), and Chen *et al.* (2020c), we list eight types of user behavior information from e-commerce platforms:

- **Clicks.** As the entrance to an item page, a click on an item hints that the user is interested in the item. Click sources include the home page, shopping cart page, sale page, and the search result page, etc. Zou *et al.* (2020a) find that the more clicks, the bigger the interest from the user.
- **Purchases.** In e-commerce systems, purchases are very strong signals for recommendation. Most e-commerce platforms employ the *Gross Merchandise Volume* (GMV) as a gold standard for measuring success. GMV indicates the total amount of purchases from merchandise sales as the target of optimization of e-commerce (Anderson and Anderson, 2002; Lee *et al.*, 2001). Many recent studies use binary purchase information as the learning objective to characterize different levels of clicks (Zhou *et al.*, 2018e).
- **Browses.** On an item detail page, there are three browsable components: the main page (including basic information, title, price, pictures, etc.), the specification page (including more parameters and details), and comment page. The browsable components are helpful to understand users' interests, e.g., if a user browses the comments and specifications instead of only browsing the brief information, they have a higher probability of buying this item.
- **Add-to-carts.** Adding to cart and ordering actions offer strong signals for e-commerce search and recommendation (Su *et al.*, 2018b). Adding to cart usually reflects a strong sign of buying an item, whereas it may also reflect an interest shift phenomenon or high potential for re-purchase (Zhou *et al.*, 2018e).
- **Dwell time.** Dwell time is an effective signal to measure user engagement (Yi *et al.*, 2014). It denotes the length of time that a user spends on a web page before navigating to another page. Typically, the longer the dwell time, the more appealing the page. Dwell time is widely captured on e-commerce sites.

- **Product-aware question answering.** As explained above e-commerce platforms allow consumers to ask product-aware questions to those whom bought the same product. Correspondingly, consumers can also answer these questions asked by other users on the platform. These questions and answers provide explicit feedback and opinions of the user (Gao *et al.*, 2019b).
- **Interactions with customer services.** Provided to customers before, during, and after a purchase, customer services give direct one-on-one interactions between a consumer and the e-commerce service provider via multiple channels, e.g., dialogues, emails, and messages. Most user feedback from customer service is textual information. However, recently more and more multi-modal information, e.g., images, videos and audio messages is also included (Zhao *et al.*, 2021a).
- **Reviews and comments.** As explained above, reviews and comments are prevalent in e-commerce platforms. Reviews and comments, written by consumers, explicitly reflect their opinions about specific products and services on the e-commerce platform.

Given these types of user behavior, recent research on analyzing user behavior on e-commerce platforms focuses on answering the following questions:

- How do people make their shopping decisions? What is the process from a user's click to their purchase in e-commerce?
- What is the post-click behavior in e-commerce? What is the difference between macro-behavior and micro-behavior?

In this section, we analyze recent work on user behavior analysis in e-commerce: (i) click behavior analysis (Section 3.2.1); and (ii) user engagement and post-click behavior (Section 3.2.2).

3.2.1 From clicks to purchases

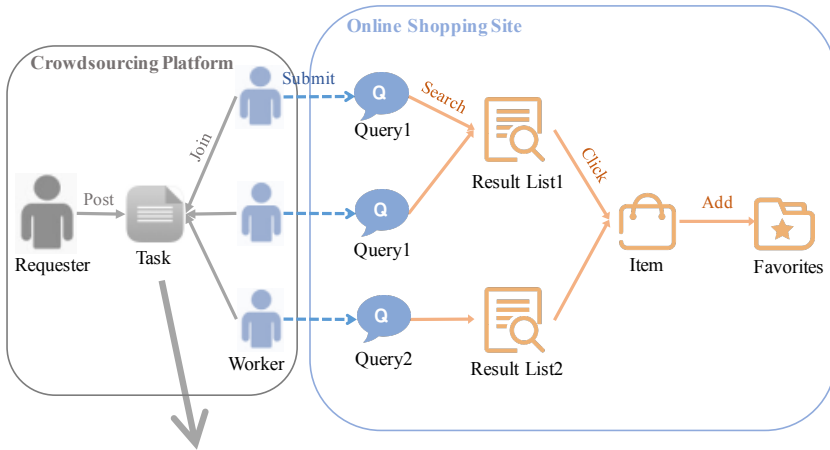
As an e-commerce user interacts with an item, they express a certain degree of interest in the item. When users browse an e-commerce platform, they may examine a specific item that is sufficiently relevant or intriguing. User clicks are an important signal for tracking a user's

interest (Chuklin *et al.*, 2015). A user's online shopping behavior can be divided into two consecutive stages: item selection/clicks, and the decision to purchase the clicked item. Users may have different intentions while shopping online, e.g., some wish to make a purchase as soon as possible while others are just looking around so as to get inspired. Therefore, Wu *et al.* (2018a) argue that clicks, as a kind of implicit feedback, should be integrated with other kinds of feedback to evaluate the "relevance" of items given a query on e-commerce portals.

During online shopping, users can add items to shopping carts and purchase them, but many platforms also facilitate additional types of activity. For instance, "adding to favorites" is a function to help users save some potentially interesting items for future purchase activities. To some extent, the degree of "adding to favorites" reflects the popularity of an item that can be exploited as a facet for item ranking for e-commerce search and recommendation (Li *et al.*, 2011).

To boost sales, some online retailers modify the ranking of their items' popularity with the usage of crowdsourcing platforms. For example, Su *et al.* (2018b) investigate and detect such kind of activities in e-commerce, e.g., crowd workers need to follow some particularly designed guidelines to disguise themselves as normal users. An example of this crowdsourcing "add to favorites" task is shown in Figure 3.16. By simultaneously manipulating a number of crowdsourcing tasks and collecting user behavior, the authors compare behavioral attributes between normal activities and spamming activities. Figure 3.17 shows these comparisons in terms of four behavioral attributes: query length, page number, browse time (time period between search and click), and dwell time (on detailed item pages).

Different recommendation scenarios on an e-commerce platforms may yield different types of user click and purchase behavior. E.g., clicks on the follow-up recommendation results after adding an item to the shopping cart, and clicks on the recommended results listed on an item's detailed page (Zhou *et al.*, 2018e). The diversity in scenarios may make it harder to interpret clicks and their relation to purchase behavior. In e-commerce search and recommendation, a purchase action is a natural ground-truth label for a click. If a user ends up purchasing an item after clicking on it, such a click indicates the user's strong interest in and



Crowdsourcing Task: Add to Favorites in Online Shopping Site


Task Information

ID: 86414 Reward: US\$ 0.72 Bidding Start: 2017-03-14 Bidding Start: 2017-03-17

Task Description

(1): Submit one of the following queries:
 a. cowhide thermal boot man
 b. cowhide corduroy boot man
 (2): Browse the results list for 3 minutes
 (3): Click on the item in the figure below
 (4): Browse the details page for 2 minutes
 (5): Click the Add to Favorites button

Task Attachments



Task Submission

(1): Screenshot of your account
 (2): Screenshots of each step

Figure 3.16: An example of crowdsourcing “add to favorites” task. Image source: Su *et al.* (2018b).

satisfaction with the item. Accordingly, the conversion rate has been proposed as an important signal (Zhou *et al.*, 2018e):

$$\text{Conversion rate} = \frac{\text{Number of clicks that ended with an order}}{\text{Number of clicks}} \quad (3.1)$$

Purchase intent represents a predictive measure of subsequent purchasing behavior. Understanding purchase intent and how it is built up over time is important for personalized and contextualized e-commerce services. In recent years, many studies have explored the conversion

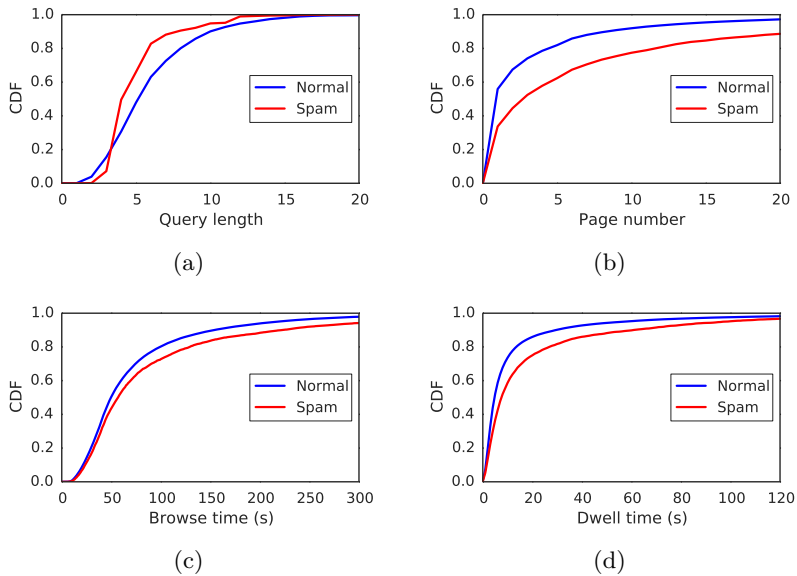


Figure 3.17: Comparisons of behavior attribute distributions between normal and spamming “add to favorites” activities. Image source: Su *et al.* (2018b).

from clicks to purchases (Wen *et al.*, 2019a; Liu *et al.*, 2020d). Moreover, to understand how user activities lead to purchase intent, both long and short-term purchase intent have been investigated (Lo *et al.*, 2016; Brown *et al.*, 2003; Kim *et al.*, 2003; Sismeiro and Bucklin, 2004; Swinyard and Smith, 2004; Suh *et al.*, 2004; Van den Poel and Buckinx, 2005; Young Kim and Kim, 2004).

Studies focusing on short-term purchase intent analysis have investigated user demographics (Young Kim and Kim, 2004), purchase patterns (Kim *et al.*, 2003), item attributes (Brown *et al.*, 2003; Van den Poel and Buckinx, 2005), and click streams (Sismeiro and Bucklin, 2004). Young Kim and Kim (2004) find that the transaction, cost, and incentive programs are important predictors for determining the short-term intention to purchase clothing, jewelry, and accessories on e-commerce portals. Furthermore, McDuff *et al.* (2015) present a large-scale analysis of the connection between facial responses and purchases.

Lo *et al.* (2016) focus on long-term purchase intent analysis. The authors perform a large-scale long-term cross-platform study of user purchase intent and how it varies over time. They focus on four kinds of signals of user actions to detect purchase intent: closing-up on a piece of content, clicking through a link to an external website, searching for content, and saving content for later retrieval. The authors find that signals for purchase intent tend to slowly build up over time, and sharply increase about three to five days before a purchase. Moreover, users with a long-term purchase intent tend to save and click-through more content; these signals may be present for weeks before a purchase is made and they are amplified in the last three days before purchase.

Social interactions can also be used to improve understanding of consumer behavior (Guo *et al.*, 2011; Gunawan and Huarng, 2015; Hajli *et al.*, 2017; Testa *et al.*, 2018). Users may consult their social network when they need to purchase something they are unfamiliar with. Thus, although social relations only provide implicit signals, they have been found to be useful to understand purchase decisions (Guo *et al.*, 2011). Bhatt *et al.* (2010) find that purchase intent from highly connected individuals is correlated with adoption by users in their social circle. However, there is little evidence of social influence by these high degree individuals. The spread of purchase intent remains mostly local to first-adopters and their immediate friends. In a 2011 study of information passing in Taobao, Guo *et al.* (2011) verify that implicit information passing is present in the network, and that communication between buyers is a fundamental driver of purchasing activity. Zhang and Pennacchiotti (2013) present a system to understand the relation between users' Facebook profiles and purchase behaviors in eBay. Extensive analyses have been done on a benchmark dataset collected from Facebook and eBay; the authors find that there are significant correlations between social network information and online purchases.

3.2.2 User engagement and post-clicks

User engagement is usually described as a combination of cognitive processes such as focused attention, affection, and interest (Mathur *et al.*, 2016). In e-commerce, there is a long line of research that analyses

user engagement (e.g., O'Brien and Toms, 2010; Vanderveld *et al.*, 2016; Wu *et al.*, 2017a; Zou *et al.*, 2020a). User engagement in e-commerce can be divided into two categories: short-term engagement and long-term engagement (Zou *et al.*, 2020a). Short-term engagement refers to the instant response (e.g., clicks and dwell time on an item page), which reflects the users' real-time preferences. However, the systems may not only want to optimize for more clicks or purchases, but also to keep users in active interaction with the system (i.e., user stickiness), which is typically measured by *delayed metrics* (Lehmann *et al.*, 2012).

Long-term user engagement is more complicated than short-term user engagement; it includes, e.g., dwell time on applications, depth of the page-viewing, and the internal time between two visits (Wu *et al.*, 2017a). Long-term user engagement reflects the user's desire to stay on the e-commerce portal longer and use the service repeatedly (Zou *et al.*, 2020a), i.e., the "stickiness."

After clicking an item via search results or recommendation results, the user enters the item page. A user's post-click refers to the user's actions within the item page after the user clicks, including inner-item clicks (i.e., clicks within the item page), purchases, service contact, and thumbnail picture views (Rosales *et al.*, 2012; Yi *et al.*, 2014; Mao *et al.*, 2014). Recent studies aim to characterize such post-click behavior on item pages as different post-click behavior has sharply different conversion rates (Zhou *et al.*, 2018e; Liu *et al.*, 2020d; Lalmas and Hong, 2018; Wu *et al.*, 2018a).

To illustrate post-click behavior on an e-commerce platform, Figure 3.18 provides an example of observed data of a user during a short period. We see that the user first enters a product page for the iPhone 7 from a search result page. After reading the detailed description and comments, this user adds the item to their shopping cart. Then, the user shifts to a page for the iPhone 6 from the search result page and reads the comments. After that, they browse a page devoted to iPhone 7 cases from the sales page and order the case. Finally, they jump to a page about the Samsung Galaxy from the home page of the e-commerce site. During this period, two kinds of post-click behavior can be found: (i) from a coarse-grained perspective, the user interacted with the iPhone 7, the iPhone 6, iPhone 7 cases, and the Samsung

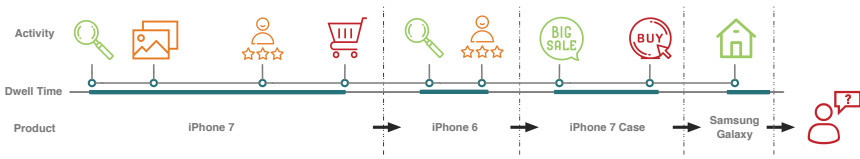


Figure 3.18: An illustrative example of post-click behavior from JD.com. Image source: Zhou *et al.* (2018e).

Galaxy; and (ii) from a fine-grained perspective, each coarse-grained interaction includes a sequence of behavior that can indicate how the user locates the item page, whether the user clicks detailed information, whether a user adds-to-cart or orders an item, and how long the user dwells on an item (Zhou *et al.*, 2018e).

As mentioned in Section 3.2, typically, there are three browsing modules on e-commerce sites: the main page (including basic information, title, price, and pictures), the specification page (including more parameters and details), and the comment page. Figure 3.19 illustrates the relations between clicks and these browsable components (Zhou *et al.*, 2018e). We see that a user is more likely to buy an item if they produce more clicks on its different browsable components, i.e., a user may gather basic information from the main item page, review feedback from the comment page, and click images to check if the item satisfies their requirements. Liu *et al.* (2020d) show how dwell time is related to clicks and browsable components; see Figure 3.20. The dwell time on an item is related to how a user locates the item. As shown in Figure 3.20(b), the longer the dwell time, the more likely a user would visit detailed components, including reading comments and specifications.

Zhou *et al.* (2018e) investigate the relation between certain types of post-click behavior and the conversion rate (CVR). They find that the post-click behavior “Cart” has the highest conversion rate, which means if a user adds an item to the cart, they are more likely to order it in the end. Similarly, if a user enters an item page from the list of items in the cart, they are also very likely to order it. When the dwell time is outside a certain range, the conversion rate drops. If the user stays much longer than they need to finish the page, they might have



Figure 3.19: The relation between clicks and browsing modules. Image source: Zhou *et al.* (2018e).

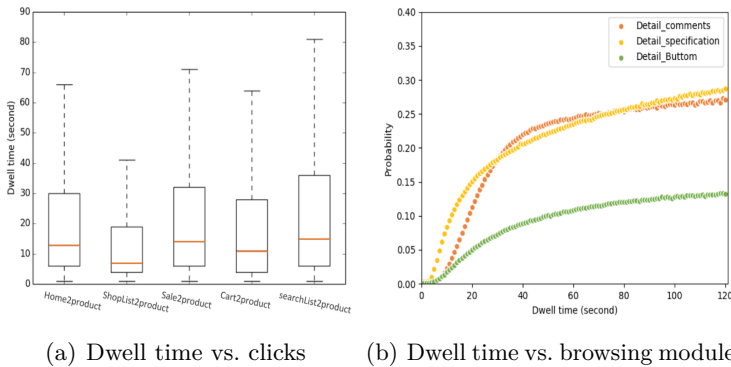


Figure 3.20: Performance of dwell time, clicks, and browsable components. Image source: Zhou *et al.* (2018e).

transferred their attention offline. It is observed that users’ interactions often exhibit a monotonic structure, i.e., the presence of a more explicit interaction (such as reviews) necessarily implies the presence of a more implicit signal (such as clicks) (Wan and McAuley, 2018).

3.3 Discussion

In this section, we have surveyed the infrastructure of e-commerce platforms, i.e., presentations and users. Specifically, we have introduced six information components that are widely applied on e-commerce platforms: search results, recommendation results, titles, product de-

scriptions, question answering, and reviews. We have highlighted studies about user behavior in e-commerce, including user clicks, purchases, engagement, and post-click behavior.

For e-commerce presentations, we have introduced basic concepts and identified key components of e-commerce interfaces. We have found that almost every e-commerce site provides six information components: search results, recommendation results, item titles, item features and descriptions, question answer pairs, and user reviews of the item. Furthermore, we have summarized recent studies that focus on analyzing the effect of these information components. Empirical studies on these information components have revealed remarkably high correlations between user behavior and information displayed in e-commerce presentations.

For e-commerce users, we have observed complex user behavior from clicks to purchases. According to empirical studies on e-commerce users, signals for purchase intent tend to slowly build up over time and sharply increase before a purchase. Studies also find that users are more likely to buy an item if they produce more clicks on its different browsable components. If a user adds an item to a cart, they are more likely to purchase it in the end. Similarly, if a user enters an item detail page from the list of items in the cart, they are also very likely to purchase it.

To gain a deeper understanding of information discovery on e-commerce platforms, we list three research questions to guide the following three sections:

- Can we model user behavior and profile users by using multiple types of user behavior, e.g., clicks, post-clicks, and purchases?
- How can we understand frameworks and components of e-commerce search through interactions between users and search engines?
- What are the principles and characteristics of e-commerce recommendations?

We will address these questions through discussions in Section 4, 5, and 6, respectively. A considerable amount of relevant work about information components will be discussed in Section 7 as they have a clear connection to question answering and dialogue generation in e-commerce.

4

E-commerce User Modeling

In Section 3, we discussed work on e-commerce information infrastructures, focusing on e-commerce presentations, and on e-commerce users. The unique characteristics of e-commerce users make modeling for e-commerce users essential when attempting to understand and support information discovery (Lo *et al.*, 2016; Huang *et al.*, 2018b). E-commerce user modeling can be separated into two types: *user behavior modeling* and *user profiling*. Given specific user behavior in various scenarios, e.g., click behavior, purchasing behavior, and post-click behavior, user behavior modeling focuses on learning a model of user behavior to predict the user's next preference. In contrast, user profiling aims to predict a user's profile (e.g., age, gender, and occupation) given the user's behavior records. In this section, we survey research on user behavior modeling and user profiling in e-commerce. First, we detail user behavior modeling approaches in Section 4.1. Next, we discuss studies on user profiling in e-commerce in Section 4.2. Lastly, Section 4.3 discusses emerging directions in e-commerce user modeling.

4.1 User Behavior Modeling in E-commerce

In this section, we describe research on e-commerce user behavior modeling, including click behavior modeling (Section 4.1.1), post-click behavior tracking (Section 4.1.2), and purchase intent modeling (Section 4.1.3).

4.1.1 Click behavior modeling

Click actions recorded in query logs have successfully been applied to extract important features in the context of ranking scenarios (Agichtein *et al.*, 2006). Regarding web search, click models have been proposed to help the search engine understand interactive user behavior (Guo *et al.*, 2009; Yilmaz *et al.*, 2010; Zhang *et al.*, 2011; Chuklin *et al.*, 2015; Borisov *et al.*, 2016). Early research on the topic aimed to track a user's behavior by using probabilistic graphical models. More recently, neural networks have been applied to improve the performance of click models by representing user behavior to capture the user's information needs (Borisov *et al.*, 2016). Focusing on improving the effectiveness by exploiting information from user-system interactions, Ferro *et al.* (2017) explore embedding dynamic interactions into learning to rank frameworks. Thereafter, curriculum learning and continuation methods have been successfully applied to exploit user interactions and facilitate rank learning (Ferro *et al.*, 2019).

Given the work mentioned above, click behavior modeling has received an increasing amount of attention in e-commerce scenarios (see, e.g., He *et al.*, 2014; Chapelle *et al.*, 2015; Chen *et al.*, 2016b; He and Chua, 2017; Li *et al.*, 2017a; Wu *et al.*, 2018a; Zhou *et al.*, 2018e; Huang *et al.*, 2019; Gong *et al.*, 2020; Bian *et al.*, 2021; Wen *et al.*, 2021). Viewing click prediction as a binary classification problem, the researchers who conducted those early studies employed logistic regression to predict whether an item will be clicked (Richardson *et al.*, 2007), where hand-crafted features are extracted from raw data to optimize a log-likelihood objective function for training. Latent factor optimization approaches, e.g., factorization machines (Rendle, 2010), have also been applied to use importance-aware and hierarchical structures purposed to manage

dynamic user behavior (Oentaryo *et al.*, 2014). In Section 6.3.2, we detail studies about factorization machines in e-commerce recommendation.

CTR prediction metric. The *click-through rate* (CTR) is a widely applied evaluation metric for click prediction that reflects the probability of a click in a trial impression. Following Regelson and Fain (2006), we established p as the probability of a click, $P = \{p_1, p_2, \dots, p_N\}$ as the set of product items, and $U = \{u_1, u_2, \dots, u_M\}$ to represent the set of users. The maximum-likelihood estimate of p refers to the number of observed successes divided by the number of trials, i.e., clicks/impressions. Given a set of search or recommendation sessions S and a query q , the probability of a product CTR($p|q$) can be formulated as follows:

$$\text{CTR}(p|q) = \sum_{s^q \in S} \frac{\Psi^{s^q}(p)}{|s^q \in S|}, \quad (4.1)$$

where s^q denotes a session with q , and Ψ^{s^q} denotes an event of a click within s^q .

From shallow to deep models. CTR has been widely applied as an evaluation metric for click modeling in e-commerce portals. Rendle and Schmidt-Thieme (2010) introduce a tensor-based method for CTR prediction; Bayesian approaches have also been used effectively for CTR prediction (Graepel *et al.*, 2010). Starting in 2015, deep learning significantly improved CTR estimation by transferring traditional architectures and developing new ones. Deep neural networks effectively capture high-order feature interactions, resulting in better CTR prediction performance. Zhang *et al.* (2016a) describe a deep neural network to learn patterns from categorical feature interactions. Similarly, Chen *et al.* (2016b) and Zhu *et al.* (2017) employ neural network models with multiple fully-connected layers to predict user clicks. Aryafar *et al.* (2017) investigate CTR prediction in promoted listings by using an ensemble learning approach to use different signals of listings. Generally, these logistic regression models can effectively achieve memorization by applying cross-product transformations over sparse features. More recent work involves representing sparse features as dense vectors, which are concatenated to form an instance vector. This vector is then passed through a multi-layer perceptron, with a sigmoid output layer, to predict

the click probability. These advancements have greatly enhanced model accuracy in CTR tasks (Zhang *et al.*, 2021b).

Wide & Deep model. Modeling the interactions between features, especially the interactions between low-order and high-order features, is essential for click prediction. The Wide & Deep model (Cheng *et al.*, 2016) considers low- and high-order feature interactions simultaneously. Wide & Deep pursues the balance between memorization and generalization. Owing to its simple structure, the “strong” features (i.e., feature combinations) of Wide & Deep allow for the assignment of larger weights during training, thus endowing the model with stronger memory. Besides the deep component based on an MLP, Wide & Deep consists of another component, the wide component. It is a generalized linear model with an input feature set that includes raw features and a feature that has been transformed by the cross-product transformation and is defined as

$$\phi_k(x) = \prod_{i=1}^d x_i^{c_{ki}}, \quad c_{ki} \in \{0, 1\}, \quad (4.2)$$

where c_{ki} is a Boolean variable that is 1 if the i -th feature is part of the k -th transformation ϕ_k , and 0 otherwise. Such a transformation allows the model to capture the interactions between the binary features, and adds nonlinearity to the wide component. The overall model architecture of Wide & Deep is shown in Figure 4.1. Wide & Deep has been shown to be effective in e-commerce recommendation scenarios; more details are provided in Section 6.3.3.

The Wide& Deep model is a representative of dual tower models for user behavior modeling. Similarly, DeepFM (Guo *et al.*, 2017), DCN (Wang *et al.*, 2017c), xDeepFM (Lian *et al.*, 2018) and Autoint (Song *et al.*, 2020a) have also been put forward for CTR prediction. The deep neural network part in these dual tower models can always be regarded as a supplementary to learn the residual signal of the feature interaction layer to approach the label, which yields stable training and the improved performance. In contrast, single-tower models like NFM (He and Chua, 2017) and the product-based neural network (Qu *et al.*, 2018) have enhanced their modeling capacity due to their more sophisticated network structures, which allow them to capture complex

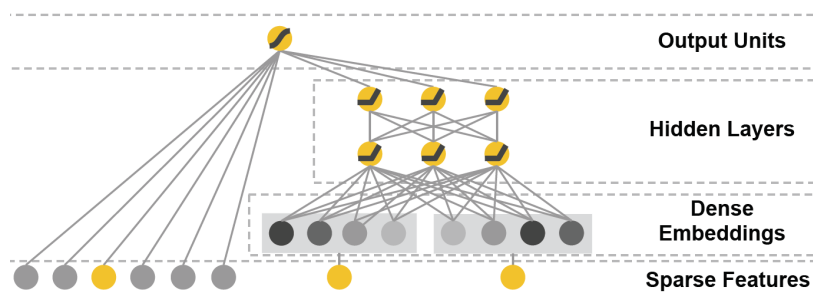


Figure 4.1: Wide & Deep model architecture. Image source: Cheng *et al.* (2016).

feature interactions. However, they often struggle with issues such as getting stuck in poor local minima and exhibit a heavy reliance on careful parameter initialization. This sensitivity to initialization can affect their training stability and convergence, making optimization more challenging than for simpler models.

Attention models for CTR. Attention neural networks have been proposed to enhance the performance of CTR prediction. The deep interest network (DIN) (Zhou *et al.*, 2018a) is the first model to introduce the attention network mechanism for user behavior modeling with CTR prediction. It assigns different weights to past behaviors based on their relevance to the target item. To capture dynamic interest evolution, the deep interest evolution network (DIEN) (Zhou *et al.*, 2019) has been proposed; it uses a two-layer GRU with an attentional update gate to model evolving user interests. Further advancements, like the behavior sequence transformer and the deep session interest network (Feng *et al.*, 2019), use self-attention to model behavior dependencies and session-based representations, showing the importance of attention mechanisms in CTR prediction (Xiao *et al.*, 2020). More recent advances in user click models with attention have focused on using deep neural networks to capture complex interactions given user profiles, item attributes, and contextual features (Hou *et al.*, 2023). These models have shown great potential in improving the accuracy and scalability.

Memory-based models. With the accumulation of large amounts of user behavior data on large e-commerce platforms, effectively han-

dling long behavior sequences is increasingly important. However, many models such as DIN (Zhou *et al.*, 2018a) struggle with the time complexity when processing such sequences. To address this, Ren *et al.* (2019a) introduce the hierarchical periodic memory network; it uses a lifelong memory mechanism with multi-layer GRUs updating at different frequencies, capturing long-term and multi-scale temporal patterns. Similarly, the user interest center and the multi-channel user interest memory network are designed to handle long-term user interest modeling, providing a more systematic, industrial-level approach (Pi *et al.*, 2019). Multi-interest networks have also been studied to improve the robustness and consistency in user click modeling (Cen *et al.*, 2020; Chang *et al.*, 2023). To mitigate noisy correlations and user intent vanishing during this procedure, attribute transition graphs and matching among various patterns need to be constructed. To this end, Liu *et al.* (2023c) characterize user intents with attribute patterns, where the frequent and compact attribute patterns serve as memory to augment session representations.

Hybrid models combining multiple factors. To address the complexity of feature interactions, various hybrid models have been proposed. For example, the gradient boosting decision tree model (GBDT; Chen and Guestrin, 2016) has been applied successfully to predict user clicks (He *et al.*, 2014). Figure 4.2 illustrates the structure of a hybrid model with GBDT and logistic regression. The model concatenates the boosted decision trees, which transforms features and the sparse logistic regression classifier. The input is a structured embedding $x = (e_{i_1}, e_{i_2}, \dots, e_{i_n})$ for each item x , where e_i refers to the i -th unit vector, and i_n is the index of the categorical features. The output of the model is a binary label $y \in \{+1, -1\}$, which indicates a click or no click. Given a labeled pair (x, y) , the authors denote the linear combination of active weights as $s(y, x, w)$, which can be calculated as follows:

$$s(y, x, w) = y \cdot w^T \cdot x = y \sum_{j=1}^n w_{j, i_j}, \quad (4.3)$$

where w is the weight vector of the click score. Using stochastic gradient descent (SGD; Saad, 1998), the authors inferred the likelihood function $p(y|x, w)$ by applying a sigmoid function over $s(y, x, w)$. Based on these

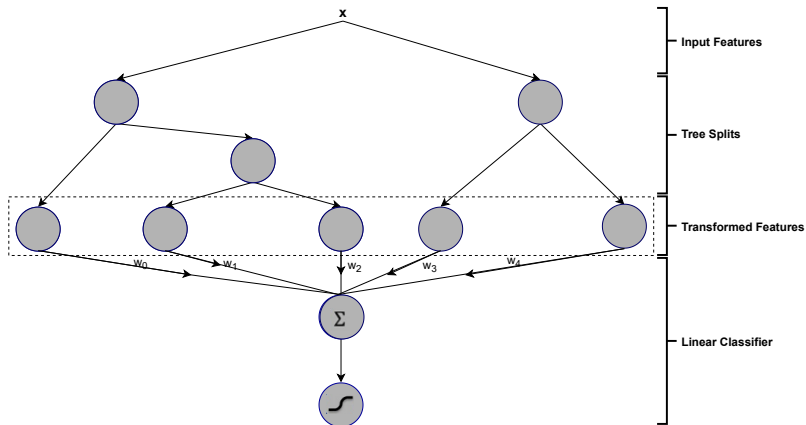


Figure 4.2: Overview of the hybrid click prediction model that uses GBDT and logistic regression. Image source: He *et al.* (2014).

transformed features, the authors applied logistic regression to predict a click or no click. Boosted decision trees are able to aggressively reduce the number of active features with only moderate prediction accuracy degradation. The hybrid click model is widely applied in e-commerce recommendations for candidate ranking (see Section 6.3.2 for more details).

To explore the feature interactions hidden in data collections, Guo *et al.* (2017) propose a neural network method, i.e., DeepFM, that combines the architectures of factorization machines and deep neural networks. As shown in Figure 4.3, DeepFM uses a wide and deep component to share the same raw input feature vector; this allows the model to learn low- and high-order feature interactions simultaneously. All parameters are jointly trained for the combined prediction model, as described by Equation 4.4:

$$\hat{y} = \text{sigmoid}(y_{FM} + y_{DNN}), \quad (4.4)$$

where $\hat{y} \in (0, 1)$ refers to the predicted CTR, y_{FM} is the output of the FM component, and y_{DNN} is the output of the deep component. The authors apply a feed-forward network in the deep component to learn higher-order feature interactions.

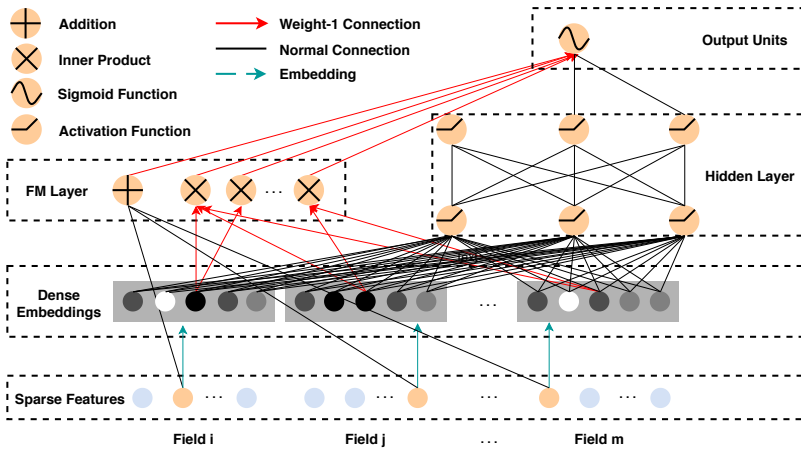


Figure 4.3: Architecture of the DeepFM model for CTR prediction. Image source: Guo *et al.* (2017).

More DeepFM model-based deep learning methods have been proposed to address the CTR prediction problem, including deep convolutional neural networks (CNNs) (Chan *et al.*, 2018) and deep interest neural networks (Zhou *et al.*, 2018a; Feng *et al.*, 2019; Li *et al.*, 2019a; Zhou *et al.*, 2019; Chen *et al.*, 2021a; Zhu *et al.*, 2021a; Guo *et al.*, 2022b; Cheng, 2022). All of the above-mentioned deep neural networks have significantly contributed to the optimization of item ranking in e-commerce search and recommendation; this will be further discussed in Sections 5.4 and 6.3.3, respectively.

To ensure consistent evaluation and comparison of CTR prediction models, benchmark frameworks such as the open benchmarking for CTR (Zhu *et al.*, 2021b) and BARS-CTR (Zhu *et al.*, 2022a) have been introduced. These frameworks provide a standardized way to evaluate model performance across different datasets, improving reproducibility and promoting further advancements in CTR prediction research.

4.1.2 Post-click behavior tracking

As we have discussed in Section 3.2.2, post-click behavior plays an important role in modeling for e-commerce users in search (Section 5) and recommendation scenarios (Section 6). Multiple studies have focused on

applying various types of interaction signals to model post-click behaviors in search and recommendation scenarios. Sculley *et al.* (2009) measure users' post-click experience by evaluating the corresponding bounce rate. The model proposed by Zhong *et al.* (2010) uses both user clicks on the search page and post-clicks beyond the search page to provide an unbiased estimation of document relevance. Lalmas *et al.* (2015) investigate how viewport time can be used to measure user attention level as an engagement metric. O'Hare *et al.* (2016) use user interactions as signals within the clicked items to enhance the search results. Wan and McAuley (2018) determine the monotonic dependency between explicit user signals and more implicit signals to improve recommender systems. Lu *et al.* (2018) propose a preference prediction model to predict user actual preferences for the clicked items by taking into account multiple post-click interactions.

Dwell time. As we have discussed in Section 3.2.2, dwell time is the most common evaluation metric for the analysis of post-click user behavior (Yin *et al.*, 2013). Accordingly, Yin *et al.* (2013) built a graphical model that focuses on using explicit user feedback and dwell time to predict user preferences in e-commerce recommendations. Yi *et al.* (2014) show that integrating dwell time into the learning objective or learning weight results in better recommendation performance than pure predictions of the CTR. Rosales *et al.* (2012) and Chapelle *et al.* (2015) use dwell time as a proxy of post-click experience in online advertising to improve the ranking performance. Bogina and Kuflik (2017) explore the value of incorporating dwell time for session-based recommendations by boosting items above the preassigned dwell time threshold. Modeling user behavior by taking into account dwell time has been shown to facilitate e-commerce recommendation performance. In Section 6.2.3, we will discuss more studies that focused on modeling sequential user dynamics by using dwell time.

User return modeling. There is a limited amount of work on modeling user returns in e-commerce, especially when user returns depend heavily on the quality of the provided service (Lo *et al.*, 2016; Zhou *et al.*, 2018e). The model developed by Zhou *et al.* (2018e) provides rich user interfaces after a user clicks an item. For instance, it encourages

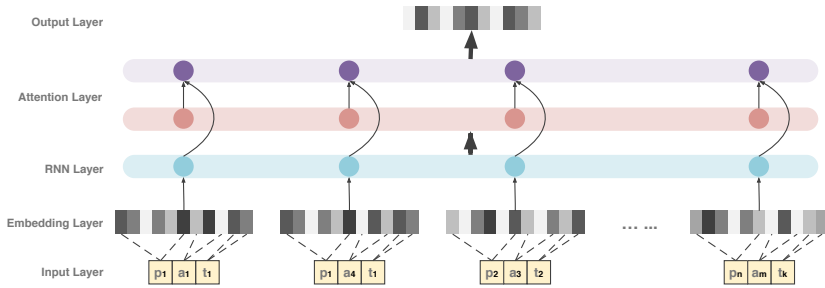


Figure 4.4: Micro-behavior modeling framework. Image source: Zhou *et al.* (2018e).

users to visit different item modules or sub-pages, i.e., to read comments or click on pictures embedded within the item page; this generates a large amount of heterogeneous post-click behavior. Three problems pose a challenge for attempts to model micro-behavior on e-commerce platforms: (i) Sparseness and high dimensionality of the user representation; (ii) Sequential information of micro-behavior; and (iii) Diverse effects of micro-behavior. To address these three challenges, Zhou *et al.* (2018e) propose the framework shown in Figure 4.4, which consists of five layers: an input layer, an embedding layer to solve the problems of sparseness and high dimensionality, an RNN layer to model sequential information, an attention layer to capture the diverse effects of micro-behavior, and an output layer. The input of the model comprises the data of a user, u , with a sequence of micro-behavior. Formally, the authors define it as the sequence $S_u = \{x_1, x_2, \dots, x_n\}$, where each x_i is a tuple, i.e.,

$$x_t = (p_v, a_m, d_k), \quad (4.5)$$

where $p_v \in \mathbb{R}^V$ is a one-hot indicator vector where $p_v(i) = 1$ if x_i is about the i -th product and other entities are zero. Similarly, $a_m \in \mathbb{R}^M$ and $d_k \in \mathbb{R}^K$ are indicator vectors for activities and dwell time, respectively. Each indicates a unique element in the product set P , activity set A , and dwell time set D , respectively. Here, the vocabulary sizes of P , A , and D are V , M , and K , respectively, and there are $V \times M \times K$ tuples in total. To address sparseness and high-dimensionality problems, the authors design an embedding layer to transform the input x_t into a low-dimensional dense vector e_t :

$$e_t = \text{concatenate}(W_P p_v, W_A a_m, W_D d_k), \quad (4.6)$$

where $W_P \in R^{d_P \times V}$, $W_A \in R^{d_A \times M}$, and $W_D \in R^{d_D \times K}$, where $d_P \ll N$, $d_A \ll M$, and $d_D \ll K$ are the number of latent dimensions for products, activities, and dwell time, respectively. The initial weights of W_P , W_A , and W_D are trained by applying word2vec (Mikolov *et al.*, 2013). Additionally, the final embedding of x_t is the concatenation of three embeddings. To capture the sequential information of micro-behavior, the authors build an RNN layer. The output of the embedding layer e_t is the input of the RNN layer. The t -th hidden state unit output is calculated as

$$h_t = \sigma(W_{eh}e_t + W_{hh}h_{t-1} + b_t), \quad (4.7)$$

where $\sigma(\cdot)$ is a non-linear activation function, e.g., ReLU, sigmoid, or tanh; $W_{eh} \in R^{d_h \times d_e}$, $W_{hh} \in R^{d_h \times d_h}$, and $b_i \in R^{d_h}$. To capture the effects of micro-behavior, the authors introduce an attention layer (Mnih *et al.*, 2014) that assigns proper weights to each hidden unit; this helps to obtain a more balanced output. The attention weight is mapped from the hidden layer vector to a real valued score by the function $\sigma(\cdot)$. To achieve sufficient expressive ability, the function $\sigma(\cdot)$ is typically implemented by a neural network layer. Then, the final output is an attention weighted pooling of the RNN layer. To exploit the different transition patterns between items and operations in micro-behavior modeling, Meng *et al.* (2020c) incorporate item knowledge into a joint user modeling framework including a recurrent neural network and a graph neural network. To incorporate the micro-behavior information in the iterative process of user behavior modeling, Yuan *et al.* (2022a) model a user session as a fine-grained sequence of micro-behaviors and proposed a self-attention mechanism to encode the dyadic relations of micro-behaviors.

Experimental results have confirmed that post-click user modeling can provide deeper insights into user behavior, which is used to advance e-commerce search and recommender systems by successfully modeling the sequential dynamics in the candidate retrieval stage (Zhou *et al.*, 2018e). However, post-click behaviors are often sparse in real-world scenarios, making it challenging to supplement large-scale implicit feedback. To address this, recent studies have integrated post-clicks with other user behaviors. Wen *et al.* (2019b) describe a generic probabilistic framework

to fuse click and post-click feedback in recommender systems. Wang *et al.* (2021) reveal the importance of mitigating the clickbait issue from click behaviors, and apply causal inference to establish a causal graph to reformulate the process. In Section 6.2.3, we discuss further studies that have focused on integrating post-click tracking into e-commerce recommendation.

4.1.3 Purchase-intent modeling

Purchase-intent prediction is another important task in e-commerce modeling (Qiu *et al.*, 2015; Kooti *et al.*, 2016; Lo *et al.*, 2016; Wan *et al.*, 2017). According to Bellman *et al.* (1999), the volume of online activities of a customer proves useful when predicting the occurrence of a future purchase. Statistical models of customer purchase behavior have been studied for decades. Early research on purchase behavior modeling was based on statistical approaches, e.g., negative binomial distribution models (Bearden and Netemeyer, 1999; Kooti *et al.*, 2016). Features related to information gathering and the purchase potential (e.g., monetary resources and product values) also help to predict the purchase intention (Bearden and Netemeyer, 1999; Hansen *et al.*, 2004; Pavlou and Fyngenson, 2006).

Feature-based methods. User-aware features have been successfully applied to predict purchase intent in e-commerce scenarios. Qiu *et al.* (2015) put forward a pipeline-based purchase-prediction approach that includes three main components. First, the authors use associations between products to predict the needs of customers; then, they combine collaborative filtering and a hierarchical Bayesian discrete choice model enable customer preference learning; lastly, they construct a support vector regression-based model to calculate the popularity of products. After analyzing user behavior on Pinterest, Lo *et al.* (2016) propose a predictor to detect a user's purchase intent. The authors apply five kinds of feature: demographics, activity, action-type, content, and temporal features. Wan *et al.* (2017) also introduce a three-stage model to predict the purchase behavior on a real-world e-commerce portal. To identify who can be converted to regular loyal buyers and then targeted to reduce promotion cost, Liu *et al.* (2016a) describe a solution for repeat

buyer prediction; they collect a large number of features to capture the preferences and behavior of users, characteristics of merchants, brands, categories, and items, and the interactions among them. Hendriksen *et al.* (2020) analyze the potential of long-term historical records (from logged-in users) to more accurately and reliably predict purchase intent. Additionally, Ariannezhad *et al.* (2021) show that data that provides information on customer behavior in one channel (e.g., online) can help to predict purchase intent in other channels (e.g., offline). Social media has become another important source of information to help explore consumer purchase intentions (Mishne and de Rijke, 2006a; Zhang and Pennacchiotti, 2013; Ding *et al.*, 2015). Zhang and Pennacchiotti (2013) explore whether users' social media information is correlated with their e-commerce profiling categories. Accordingly, the authors use correlations to build machine learning algorithms to predict user purchase behavior.

CVR prediction methods. *Conversion rate* (CVR) prediction is another way to predict the user purchase intention on e-commerce platforms. CVR calculates the proportion of users who will eventually convert after clicking (Lee *et al.*, 2012; Yang *et al.*, 2016; Lu *et al.*, 2017; Wen *et al.*, 2019a; Su *et al.*, 2020b; Wen *et al.*, 2020; Yasui *et al.*, 2020; Yang *et al.*, 2021a; Hou *et al.*, 2021; Li *et al.*, 2021a). Because conversions are extremely rare, CVR modeling is very challenging. CVR can be split into the following two categories: post-view conversion and post-click conversion, i.e., conversion after viewing an item without having clicked it, and conversion after having clicked the item, respectively. Most approaches focus on the task of post-click conversion. Wen *et al.* (2019a) propose a decision tree ensemble model, i.e., *ldcTree*, that exploits deep cascade structures and applies cross-entropy based feature representations. Nonetheless, there are still three challenges in CVR estimation: *data sparsity*, *sample selection bias*, and *delayed feedback*. The data sparsity problem reflects the insufficiency of click samples in training data. Sample selection bias refers to the systematic difference in the data distribution between the training space and inference space (Wen *et al.*, 2020). Figure 4.5 illustrates the sample selection bias problem related to the development of an efficient industrial-level

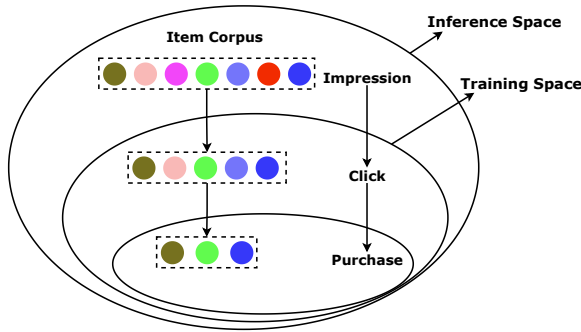


Figure 4.5: Illustration of the sample selection bias problem in conventional CVR prediction, where the training space only consists of clicked samples, whereas the inference space is the entire space of all items. Image source: Wen *et al.* (2020).

recommender system. The delayed feedback problem indicates that the e-commerce platform can receive feedback with a delay after an item is impressed or clicked by a user (Chapelle, 2014; Su *et al.*, 2020b).

To address the data sparsity problem, Ma *et al.* (2018) consider the entire space multi-task model, which aims to apply multi-task learning to accomplish two subtasks of predicting the post-view click-through rate and post-view conversion rate. Wen *et al.* (2020) observe that users always engaged in abundant purchase-related actions after clicking. Thus, they propose a deep neural network method within a multi-task learning framework to decompose post-click behavior to predict CVR. The authors distinguish between purchase-related actions and other actions, which, taken together, can be used to form a probabilistic sequential user behavior graph.

The above multi-task strategies are also helpful in alleviating the selection bias problem. Furthermore, Zhang *et al.* (2020b) detail a doubly robust estimation method to debias CVR prediction. Yasui *et al.* (2020) provide a dual learning method to simultaneously address the delayed feedback problem and the selection bias problem. To reduce the variance of doubly robust loss to enhance model robustness, Guo *et al.* (2021) enhance a more robust doubly robust approach for debiasing post-click conversion rate estimation. But the authors do not directly control the bias and the variance in an effective way. To address this problem, Dai *et al.* (2022) propose a generalized framework of doubly

robust learning, which unifies the existing doubly robust methods. Based on this framework, two new doubly robust methods were proposed to control the bias and mean squared error.

Chapelle (2014) describe a delayed feedback model to optimize CVR as a joint probability over the predicted CVR and the delayed time distribution. Yoshikawa and Imai (2018) extend this delayed feedback model to a non-parametric model. Su *et al.* (2020b) focus on post-click calibration in CVR modeling. The authors extract pre-trained embeddings from impressions/clicks to enhance the conversion models; they propose an inner/self-attention mechanism to capture the fine-grained personalized product purchase interests. To estimate unbiased CVR in the online settings, Yang *et al.* (2021a) introduce a elapsed-time sampling delayed feedback model to track relations between the observed conversion distribution and the true conversion distribution. Li *et al.* (2021a) use an idealized dataset for training a prophet model that can use the data properly, and then learn the actual model by imitating the prophet. Chen *et al.* (2022) confirm the importance of dividing observed samples in a more granular manner, and hence propose an unbiased importance sampling method with two-step optimization to address the delayed feedback issue.

These purchase-intent modeling strategies have been applied in e-commerce searches and recommendations; we will discuss these strategies in more detail in Sections 5.4 and 6.2.

CLTV prediction methods. Lastly, *customer lifetime value* (CLTV) prediction has also received attention in recent years. CLTV is an important task in e-commerce search and recommendation models. CLTV is defined as the sales, net of returns, of a customer over a 1-year period. The objective of CLTV prediction is to improve three key business metrics: (i) the average customer shopping frequency, (ii) the average order size, and (iii) the customer churn rate. With CLTV prediction, e-commerce retailers can rapidly identify and nurture high-value customers (Vanderveld *et al.*, 2016). Classic work on CLTV prediction applies handcrafted features and ensemble classifiers (e.g., GBDT, Chen and Guestrin, 2016).

4.2 User Profiling in E-commerce

A *user profile* refers to personal information about a specific user. Personalization plays an important role in web search and recommender systems. In e-commerce portals, user profiling is a critical module for e-commerce information discovery tasks, as it provides personalized content in search or recommendation results. User profiling can be defined as the process of exploring information about a user's interest domain (Dong *et al.*, 2014; Kanoje *et al.*, 2015). Information about a user can be used by e-commerce search and recommender systems to enhance the system effectiveness because it enables better user understanding (see Section 5.3.5 and 6.4). Given that it originated from work on the prediction of user purchase intention, research into user profiling in e-commerce has continuously garnered attention over the years (see, e.g., Solomon and Behavior, 1994; Braynov, 2003; Hollerit *et al.*, 2013; Zhang and Pennacchiotti, 2013; Gupta *et al.*, 2014; Rahdari *et al.*, 2017; Huang *et al.*, 2018b). Early work on user profiling for e-commerce mainly focuses on information filtering, social media analysis, web searches, and fraud detection (Solomon and Behavior, 1994; Fawcett and Provost, 1996; Adomavicius and Tuzhilin, 1999; Kuflik and Shoval, 2000; Braynov, 2003). Most of these are rule-based strategies. For example, Fawcett and Provost (1996) employ a rule-based user-profiling method to uncover indicators of fraudulent behavior. These indicators were used to create user profiles that were then applied as features of their proposed system, which combines evidence from multiple profilers to generate high-confidence alarms.

4.2.1 Types of user profiling

User profiling can be classified as either *profile extraction* and *profile learning* (Tang *et al.*, 2010). Profile extraction focuses on extracting information about a user, such as demographic data (e.g., age, gender, location) or basic behavior patterns. However, in e-commerce, profile extraction is often less critical because it provides only a fixed snapshot of the user, lacking the adaptability needed to capture evolving preferences and real-time behavioral changes. In contrast, profile learning is more

significant as an e-commerce modeling tool (Kufflik and Shoal, 2000; Cufoglu, 2014). Profile learning methods can be grouped into three categories: (i) content-based methods, (ii) collaborative methods, and (iii) hybrid methods (Cufoglu, 2014). Content-based methods infer user profiles based solely on the users' own previous behavior (Kufflik and Shoal, 2000). Collaborative methods in user profiling focus on applying collaborative filtering approaches to infer profile information based on the behavior of users in a suitably defined neighborhood of similar users. Collaborative filtering methods can be classified as either *memory-based methods* or *model-based methods* (Godoy and Amandi, 2005; Cufoglu, 2014). Memory-based solutions estimate ratings for a user based on the entire collection of previous ratings of similar users (Adomavicius and Tuzhilin, 2005; Su and Khoshgoftaar, 2009).

In contrast to memory-based collaborative filtering methods for user profiling, model-based methods use the collection of ratings to learn a model to estimate user profiles (Su and Khoshgoftaar, 2009). Hybrid methods have garnered attention because they combine content-based methods and collaborative methods (Godoy and Amandi, 2005). Regarding research on personalized recommendations, to capture users' information and interest, Liu *et al.* (2010) introduce a dynamic collaborative filtering method for news recommendation where the recommender constructs user profiles based on their past click behavior. The authors conduct a log analysis of the changes in user interest in news topics over time. By classifying users' news interests as either *genuine interests* or the *influence of local news trends*, the authors are able to (i) construct a Bayesian framework to model a user's genuine interests based on their past click behavior, and (ii) predict current interests by jointly analyzing the genuine interest and the local news trends.

4.2.2 User profiling with social media

Social media is playing an important role in e-commerce user profiling. On the one hand, social media provides a source for generating user profiles, especially when addressing the *cold start* problem. Mishne and Rijke (2006b) provided an early example of this idea, using a combination of text analysis and external knowledge sources to estimate

the commercial tastes of bloggers from their posts. On the other hand, profiling information learned from social media data can be applied to explore the user's purchase intentions on e-commerce platforms. Targeting users with no history on an e-commerce site, Zhang and Pennacchiotti (2013) focus on predicting the purchase behavior by proposing a feature-selection method to predict the product categories from which a user will buy. Representing user purchase intent as textual information that indicates a desire to purchase a product or service in the future, Gupta *et al.* (2014) propose a binary classification approach to identify the user purchase intention based on their social media posts. Ding *et al.* (2015) explore relationships between a user's consumption habits and their social media data. They propose a consumption intention mining model (CIMM) based on CNNs. Lo *et al.* (2016) analyze user activities in social media to build a time-varying model to predict user purchase intent. The authors analyze Pinterest¹ data to understand how the usage of an e-commerce platform relates to future user shopping behavior. They find that indicators of purchase intent tended to gradually build up over time and sharply increase 3 to 5 days before purchase. Multi-modal information has also been applied to infer user profiles. Gelli *et al.* (2017) focus on automatically discovering actionable images for users according to their personality. By applying their model to a large-scale dataset, the authors find a significant correlation between personality traits and affective visual concepts in the image content.

4.2.3 Graph-based user profiling

Most approaches to user profiling only use a single type of information. In e-commerce modeling, heterogeneous graphs are also being used to work with user profiles. Figure 4.6 shows an example of a graph with heterogeneous information; particularly, three kinds of nodes were applied to represent three types of data, i.e., users, items, and attributes, respectively. Chen *et al.* (2019f) focus on applying rich interactions among data instances, i.e., co-click and co-purchase behavior on e-commerce platforms, to enhance user-profiling performance in e-commerce models. Neighborhood features provide useful information that helps to

¹<https://www.pinterest.com>

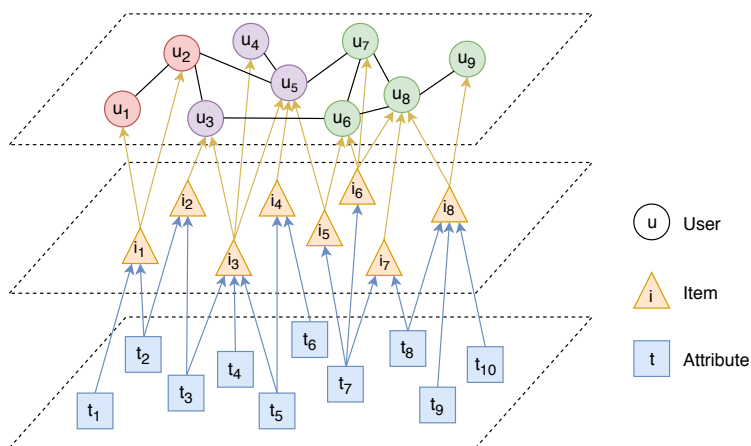


Figure 4.6: User profiling results in the form of a heterogeneous graph. Image source: Chen *et al.* (2019f).

infer user profiles, e.g., users that have similar co-purchase behavior on e-commerce platforms are likely to be of a similar age. The authors propose a heterogeneous graph attention network to infer user profiles within a multi-type data environment; the network is able to model unsupervised information in a heterogeneous graph by encoding the graph structure and node features. Gu *et al.* (2020b) construct a hierarchical profiling framework to model users' real-time interests at different granularities. A pyramid recurrent neural network model for hierarchical user profiling is constructed based on users' micro-behavior; it is subsequently applied to model the types and dwell times of behavior to enable an effective formulation of users' real-time interests. These graph-based user-profiling methods have been applied to enhance the re-ranking results in e-commerce recommendation systems. More details are discussed in Section 6.4.

4.3 Emerging Directions

4.3.1 Graph learning for user behavior modeling

Graph neural networks use neural networks to represent graph information (Scarselli *et al.*, 2009; Bruna *et al.*, 2014; Battaglia *et al.*, 2016).

Graph convolutional networks extend CNNs to graph structured data; they have been shown to be effective on a range of graph classification tasks (Bruna *et al.*, 2014; Defferrard *et al.*, 2016; Zhao *et al.*, 2021c; Liu *et al.*, 2022c), and when applied for semi-supervised classification (Hamilton *et al.*, 2017; Kipf and Welling, 2017) and link prediction (Ma *et al.*, 2020c). Most of these models have been designed for static graphs. However, Ma *et al.* (2020c) propose a dynamic graph neural network model that can model dynamic information. Xu *et al.* (2021a) enhance the capacity for learning complicated temporal dependencies in a graph, by proposing a transformer-style relational reasoning network with a dynamic memory updating mechanism. Graph neural networks that applied to recommender systems learn item embeddings within a large-scale item relationship graph (Ying *et al.*, 2018) that describes users, items, and pairwise relations in e-commerce scenarios (Ma *et al.*, 2020c; Gao *et al.*, 2022). Graph learning can also be applied for post-click modeling and user purchase-intent modeling. To predict fine-grained post-click CVR, Bao *et al.* (2020) design a model to represent user micro-behavior as a purchase-related micro-behavior graph. The authors apply a multi-task learning framework to construct a graph-based micro-behavior conversion model that can capture the correlation between different types of micro-behavior. The proposed multi-task learning and inverse propensity weighting modules mitigate the data sparsity- and sample selection bias-related problems. To prediction efficient and accurate CVR, Wen *et al.* (2021) propose a graph neural network to hierarchically model both micro- and macro-behaviors in a unified framework. It seems likely that graph neural networks will continue to facilitate new ways of modeling, gaining insights into, and predicting, e-commerce user behavior in search (Section 5.3.3) and recommendation scenarios (Section 6.2.2).

4.3.2 Dynamic user behavior modeling and profiling

Most studies on e-commerce user behavior modeling and user profiling are conducted under the assumption that a snapshot of user behavior is recorded on e-commerce portals. However, a user's personal interests and behavior may change over time (Koren, 2009; Gao *et al.*, 2013;

Yin *et al.*, 2014; Huang *et al.*, 2015; Yin *et al.*, 2015; Jagerman *et al.*, 2019; Wang *et al.*, 2021; Huang *et al.*, 2022). Thus, modeling e-commerce users' temporal behavior is important for e-commerce search and recommendation system development. Social networks and social media provide a rich source of information for temporal models of user behavior (Mislove *et al.*, 2010; Gao *et al.*, 2013; Yin *et al.*, 2014; Yin *et al.*, 2015). Yin *et al.* (2014) design a latent mixture model, which they named a temporal context-aware mixture model, to account for the intentions and preferences that drive user behavior. It models the topics related to users' intrinsic interests, and the topics related to temporal context' it then jointly analyzes the influences of the two factors to model user behavior in a unified way. To enable the dynamic learning of user profiles, Cao *et al.* (2017a) developed a model that considers multiple information sources and their relations. Similarly, Liang *et al.* (2018b) proposes a streaming profiling algorithm that initially applies a user-expertise tracking model to track the changes in the dynamic expertise of users; it then uses a keyword diversification algorithm to produce top- k diversified keywords that allow the users' dynamic expertise to be profiled at a specific timestamp.

Time and temporal phenomena are valuable sources of information that can facilitate the understanding and prediction of user behavior; this area of research is likely to continue to garner much attention in the near future.

4.3.3 User modeling with insufficient data

As we have discussed in this section, a wide range of click models and ranking methods can be applied to model user click behavior (Chuklin *et al.*, 2015; Borisov *et al.*, 2016; Ferro *et al.*, 2017; He and Chua, 2017; Wu *et al.*, 2018a; Ferro *et al.*, 2019; Liu *et al.*, 2022c; Vardasbi *et al.*, 2022). However, the models developed to date tend to require fully labeled data to train the ranking models, although, in realistic e-commerce scenarios, not all of a user's behavior can be recorded. Similarly, regarding post-click modeling, purchase-intent prediction, and user-profiling tasks, the reality that there is a limited amount of behavioral data makes it difficult to work with existing click modeling

and ranking solutions. For example, in the case of CVR tasks, the data sparsity problem arises when the number of training samples for the sequential behavior of the form “click \rightarrow purchase” is insufficient to fit the large parameter space of the CVR task (Wen *et al.*, 2020; Guo *et al.*, 2021). How to enhance user modeling performance under the conditions of a limited amount of imperfectly labeled data remains an important open problem in e-commerce.

5

E-commerce Search

E-commerce search, or simply “product search,” represents a special retrieval scenario where users submit queries to retrieve products using a search engine (Ai *et al.*, 2017). E-commerce search portals are gaining in popularity as many consumers choose e-commerce search on an e-commerce platform rather than generic web search (Li *et al.*, 2011). Unlike in web search, in e-commerce search there can be millions of results to surface for a given search query (Wu *et al.*, 2018a). We have discussed user behavior modeling and user profiling in Section 4 to understand how to explore and exploit information from user behavior. In this section, we focus on the other side of the coin, on search technologies that are based on users’ interactive behavior. To learn about e-commerce search solutions, we discuss research on query understanding and ranking technologies for e-commerce search. In Section 5.1 we summarize characteristics of e-commerce search. In Section 5.2 we recall key metrics used for evaluating e-commerce search. In Section 5.3 we present studies on representing e-commerce search queries. Then, we detail e-commerce ranking approaches in Section 5.4. Finally, we discuss emerging research directions in e-commerce search in Section 5.5.

5.1 Characteristics of E-commerce Search

Before detailing related work, we highlight characteristics of e-commerce search. We divide this section into two parts: an overview of e-commerce search, and challenges in e-commerce search.

5.1.1 Overview of e-commerce search

Early e-commerce search approaches are based on traditional information retrieval theory and faceted search models (Yee *et al.*, 2003; Jansen and Molina, 2006). Jansen and Molina (2006) explore the difference between ad-hoc search and e-commerce search. A grocery retrieval system has been developed by considering a discrepancy between consumers' shopping lists and retailers' stock information (Nurmi *et al.*, 2008). Early e-commerce search systems rely on information that retailers make available: either semantic markup on unstructured HTML documents or a data feed provided in some predefined structured format. Product resolution (Balog, 2011) focuses on recognizing webpages that represent the same product. Based on the task of product resolution, Duan *et al.* (2013a) propose a probabilistic mixture model for mining and analyzing product search logs. Similar setups can be also found in a product-aware keyword search system (Duan *et al.*, 2013b).

Unlike traditional ad-hoc retrieval, e-commerce search relies on a decision mechanism about consumers' purchase behavior in e-commerce portals (Li *et al.*, 2011). There are two main stakeholders in e-commerce search, consumers and business owners, whose interests align but also conflict to a certain extent (Tsagkias *et al.*, 2020). In e-commerce search, customers do not just browse relevant items, but also try to locate an item that satisfies their specific purchase intent (Li *et al.*, 2011). While consumers aim to find the best quality at the lowest price, businesses want to maximize profit, which translates into higher prices for customers or lower costs for businesses. E-commerce search typically requires more structured information (e.g., brands, categories, shops, etc.) than web search and more diversified personal definitions of "relevance" during search sessions. On the one hand, as we have discussed in Section 3.2, users come to e-commerce websites with a

wide spectrum of intents. Hence, multiple user behavior discussed in Section 3 and 4, e.g., clicks, post-clicks, and purchases, etc., should be integrated to model the “relevance” in e-commerce search. On the other hand, only a few products are actually purchased by the consumers and different individuals have different opinions even about the same product (Ai *et al.*, 2017). Thus, e-commerce search should consider users’ differences to satisfy the needs of all consumers. In general, there are four unique characteristics in e-commerce search:

- **Consumer query intent.** Similar to web search, queries in e-commerce search can be divided into three classes: navigational, informational and transactional (Li *et al.*, 2011). However, e-commerce search queries take a different form. Specifically, navigational queries are product serial numbers and inquiries for customer support; informational queries include leaves in the product taxonomy and product attributes; and transactional queries are a mix of navigational and informational queries. Unlike traditional web search, there are three query intents for e-commerce search: *target finding*, *decision making*, and *exploration* (Su *et al.*, 2018a). Following the well-known web-search taxonomy due to Broder (2002), Su *et al.* (2018a) describe a hierarchical e-commerce search taxonomy to explore consumers’ shopping intents, with *shallow exploration*, *targeted purchase*, *major-item shopping*, *minor-item shopping*, and *hard-choice shopping*. The authors find that consumers tend to conduct more focused searches in target finding sessions compared to those in the decision making and exploration sessions. In target finding sessions, consumers tend to issue a few specific queries and browse only top ranked results; in decision making sessions, consumers tend to issue short queries, browse deep, and click more results; and in exploration sessions consumers issue many diverse queries but do not click often. Given these search intents, customized search approaches for each type of search queries can be developed to improve the utility of e-commerce search.
- **Heterogeneous consumer behavior.** As we have described in Section 3, multiple types of user behavior can be observed in e-commerce platforms. During an online shopping journey, a

consumer may have multiple targets at different stages. Blake *et al.* (2016) observe that, during a journey, e-commerce search proceeds as a kind of “funnel” where, initially, search is along broad categories, and then it becomes more refined to obtain an item at the lowest cost given a consumer’s cost of search. Hence, e-commerce search approaches should be aware of the stages in each consumer’s journey. Meanwhile, the overall impact of heterogeneous consumer behavior also makes e-commerce search different from traditional web search. Users’ micro behavior, post-click behavior, and engagement make the search intent dynamic and complicated during search sessions (Zhou *et al.*, 2018e; Wu *et al.*, 2018a).

- **Online and offline ranking.** Traditional learning to rank methods sort documents according to their relevance to the query. E-commerce search has an intrinsic difference in the relevance in rankings: the notion of “relevance” is blurred (Wu *et al.*, 2018a). Users come to e-commerce platforms with a wide spectrum of intents. Some users wish to make a purchase as soon as possible while others are just wandering around the platform to get inspired. Hence, various kinds of signals, including clicks, favorites, adding carts, purchases, etc., should be integrated to model relevance in e-commerce search. E-commerce businesses having both online and physical presence bring a unique blend of infrastructure challenges (Ariannezhad *et al.*, 2021). Thus, users’ shopping experiences lie in smooth transitions from offline to online and vice versa (Tsagkias *et al.*, 2020).
- **Business criteria and metrics.** Most e-commerce platforms apply *Gross Merchandise Volume* (GMV) as the gold standard for measuring success, which indicates the total amount of sales during e-commerce activities. Thus, one of the main targets of an e-commerce search algorithm should be to maximize the value of purchases per search session (Wu *et al.*, 2018a). Many e-commerce search engines apply a two-stage framework to resolve the whole process into two successive subtasks: a ranking problem and a classification problem (Wu *et al.*, 2018a). This two-stage search process on e-commerce platforms makes the optimization more

complicated in contrast to web search. In e-commerce, regulatory and business constraints decide which products can be shown to which consumers, whereas competing brands can have agreements with an online retailer to restrict showing their products with those of their competitors. Therefore, it is important to understand consumers' inventory gaps and provide alternatives in e-commerce search (Tsagkias *et al.*, 2020).

5.1.2 Challenges in e-commerce search

Based on the above criteria, we see two main challenges in e-commerce search (Rowley, 2000; Jansen and Molina, 2006; Li *et al.*, 2011; Duan *et al.*, 2013a; Ai *et al.*, 2017; Trotman *et al.*, 2017; Wu *et al.*, 2018a). First, there exists a mismatch between users' queries and product representations where both use different terms to describe the same concepts (Li *et al.*, 2011). This mismatch problem is even more severe in personalized search when more personalized information needs to be considered during retrieval. Second, the ranking problem in e-commerce search is challenging: multiple types of information sources make ranking products in e-commerce search more complicated than in web search. Diverse relevance factors make it difficult to use traditional static-ranking evaluation metrics, e.g., NDCG and MAP, to measure the quality of rankings in e-commerce search.

Recent studies on e-commerce search that aim to tackle the above challenges, focus on one of two aspects (Trotman *et al.*, 2017; Wu *et al.*, 2018a): (i) matching optimization in e-commerce search, i.e., the vocabulary gap problem, representation-based matching, interaction-based matching, and matching in personalized search; and (ii) ranking optimization in e-commerce search, i.e., learning to rank methods and evaluation metrics. For real-world e-commerce search, a joint online and offline search framework with both semantic matching and ranking optimization modules is able to outperform traditional search systems at both semantic retrieval and personalized ranking scenarios (Li *et al.*, 2019c). In Section 5.3 and 5.4, we summarize recent work on e-commerce search that is aimed at tackling the two research challenge

listed above. Prior to that, we introduce the evaluation metrics used to assess e-commerce search in Section 5.2.

5.2 Evaluation Metrics

Evaluation in web search focuses on the relevance to a given query of documents. E-commerce search provides multiple signals to judge the saliency of items. Besides for relevance of an item to a given query, revenue-aware features are also considered in e-commerce search evaluation (Wu *et al.*, 2018a).

5.2.1 Relevance-based metrics

Relevance-aware evaluation metrics are in various information retrieval domains (Manning *et al.*, 2008), including in e-commerce search. It is common to see studies compute evaluation metrics based on the top 100 items retrieved by each e-commerce search model (Ai *et al.*, 2017; Van Gysel *et al.*, 2016a; Van Gysel *et al.*, 2018). Mean average precision (MAP), hit ratio (HR), mean reciprocal rank (MRR), and normalized discounted cumulative gain (NDCG) are four widely-used relevance-aware metrics (Van Gysel *et al.*, 2018; Wu *et al.*, 2018a). These metrics are also widely applied in various information retrieval domains (Manning *et al.*, 2008). Average precision (AP) computes the average value of Precision over the interval from 0 to 1. Given k candidate items, $AP@k = \sum_{k=1}^n P@k \cdot rel(k)$, where $P@k$ refers to Precision@ k ; $rel(k)$ indicates 1 if the k -th item is relevant and 0 otherwise $rel(k) = 0$. Based on that, MAP calculates the mean of AP for all queries:

$$MAP = \frac{1}{Q} \sum_{q=1}^Q AP_q, \quad (5.1)$$

where Q denotes the number of queries. $HR@k$ refers to the fraction of queries for which the relevant item is included in the top- k results, so we have:

$$HR@k = \frac{|Q_{rel}^k|}{Q}, \quad (5.2)$$

where $|Q_{rel}^k|$ denotes the number of queries for which the relevant item is included in the top- k results. MRR, also known as average reciprocal

hit ratio (Radev *et al.*, 2002), evaluates processes where a list of possible responses to a sample of queries ordered by probability of correctness:

$$MRR = \frac{1}{Q} \sum_{i=1}^Q 1/rank_i, \quad (5.3)$$

where $rank_i$ refers to the rank position of the first relevant document for the i -th query. DCG@ k evaluates the relevance of a document based on its position in the top- k results:

$$DCG_k = \sum_{i=1}^k \frac{rel_i}{\log(i+1)}, \quad (5.4)$$

where rel_i indicates the relevance of the document at position i . Ideal DCG (IDCG) is the DCG score for the ideal ranking, which is ranking the items top down according their relevance up to position k . NDCG allows one to compare the performance across different queries, using normalization of DCG by IDCG:

$$NDCG@k = \frac{DCG_k}{IDCG_k}, \quad (5.5)$$

5.2.2 Revenue-aware metrics

GMV indicates the total income amount transacted from merchandise sales, whereas the overall revenue generated for the e-commerce site is proportional to the GMV (Wu *et al.*, 2018a). Revenue-aware metrics are applied to e-commerce search evaluation to evaluate how the methods can improve the actual revenue of search sessions. To calculate the average revenue for every impression in each e-commerce search session, Wu *et al.* (2018a) introduce the average revenue metric, $Avg.Rev(i, q)$, for a query-item pair (i, q) as follows:

$$Avg.Rev(i, q) = \frac{price(i) \times purchase(i, q)}{|S_q|_{i \in S_q}}, \quad (5.6)$$

where $purchase(i, q)$ denotes the number of times that the item i has been purchased in a search session for a query q ; and $|S_q|$ is the set of search sessions for query q where the item i is impressed.

Wu *et al.* (2018a) specify a metric named $Rev@k$ to evaluate the revenue in e-commerce rankers. $Rev@k$ calculates the average revenues that a prediction algorithm would generate for each session. Specifically, $Rev@k$ is calculated as follows:

$$Rev@k(\rho) = \sum_{s \in S} \sum_{r_s \leq k} price(r_s^{-1}) \Phi(r_s^{-1}), \quad (5.7)$$

where ρ is the ranking order and $r_s \leq k$ denotes the top- k ranked positions in the session s . r_s^{-1} denotes the corresponding item at the position r_s , $price(i)$ indicates the price of item i , while Φ denotes a purchase event. Based on $Rev@k$, it is able to evaluate the revenue influence of the candidate rankers.

Empirical studies have been performed in benchmark e-commerce search datasets to find differences between relevance-based metrics and revenue-aware metrics (Wu *et al.*, 2018a). Relevance-based metrics primarily measure the success of retrieving relevant data from user logs, while revenue-aware metrics provide a clearer understanding of how ranking methods influence actual revenue in e-commerce scenarios. This difference shows that, although relevance is important, incorporating revenue-aware metrics offers more practical insights into the financial impact of search ranking methods within e-commerce settings.

The lack of annotated real-world benchmarks poses a challenge for conducting large-scale empirical studies in e-commerce scenarios. To address this issue, recent studies have started using both synthetic and semi-synthetic datasets in their experiments, allowing for more controlled analysis while approximating real-world conditions (Xu *et al.*, 2022a).

5.2.3 User engagement metrics

As we discussed in Section 3.2.2, user engagement is defined as the quality of user experience in interaction with a system, characterized by various attributes, e.g., positive affect, aesthetic and sensory appeal, attention, novelty, and perceived user control (Mathur *et al.*, 2016). In recent years, engagement metrics have been applied to evaluate the quality of interaction between the user and e-commerce search engines (Vanderveld *et al.*, 2016). User engagement metrics in e-commerce

can be divided into two categories: short-term engagement metrics and long-term engagement metrics (Zou *et al.*, 2020a).

To evaluate the quality of short-term user-system interactions, short-term engagement metrics about instant clicks, purchases, and dwell time are used in e-commerce search evaluation. As we discussed in Section 4.1 and 4.1.3, the *click-through rate* (CTR) and *conversion rate* (CVR) are the two most widely applied metrics to evaluate instant click and purchase prediction. Meanwhile, dwell time and bounce rate are the two main metrics used in short-term post-click evaluations (Lalmas *et al.*, 2015).

As we discussed in Section 3.2.2, long-term user engagement reflects the user's desire to stay on the e-commerce portal longer and use the service repeatedly (Zou *et al.*, 2020a), i.e., the "stickiness." Long-term user engagements measure versatile user behaviors based on a very large number of environmental interactions. Multiple long-term engagement metrics have been applied in previous studies. Wu *et al.* (2017a) employ cumulative clicks over time to estimate the long-term interactions between the user and the system. Zou *et al.* (2020a) apply three evaluation metrics to evaluate the long-term user interaction behaviors: (i) average clicks per session: the average cumulative number of clicks over a user visit; (ii) average depth per session: the average browsing depth that the users interact with the recommender agent; and (iii) average return time: the average revisiting days between. In multilingual scenarios, a transformed query converts a secondary language query into a semantically equivalent query in the primary language, allowing it to fully use the search engines' abilities. To evaluate transformed queries in multilingual e-commerce search, a set of behavior metrics based on user engagement specific to the existing search system was developed. Using these metrics, a query transformation system has been built and tested both offline and through online A/B tests on the Amazon platform, showing improvements in the multilingual search experience for customers (Hu *et al.*, 2020b).

5.3 Matching Strategies in E-commerce Search

In this section, we showcase recent research on matching in e-commerce search, especially concerning e-commerce query understanding and processing. We divide this section into five parts: (i) we recall the vocabulary gap problem in Section 5.3.1; (ii) we detail representation-based matching approaches in Section 5.3.2; (iii) we detail interaction-based matching approaches in Section 5.3.3; (iv) we detail hybrid matching approaches in Section 5.3.4; and (v) Section 5.3.5 describes studies on query processing in personalized e-commerce search.

5.3.1 Vocabulary gap

Users' shopping lists often differ from the product information maintained by retailers (Nurmi *et al.*, 2008). Duan *et al.* (2013a) find that while query languages such as SQL can be successfully applied to search in these product databases, their usage is difficult for non-experienced end users. In direct search scenarios in e-commerce platforms, most consumers formulate queries using characteristics of the product they are interested in (e.g., terms that describe the product's categories, brands, and shops, etc.). Hence, it is common to see a mismatch in e-commerce search between queries and product representations, where different tokens are used to describe the same product, i.e., the *vocabulary gap problem* (Van Gysel *et al.*, 2016a). To address this problem, early studies focus on rewriting verbose queries in e-commerce search (Bendersky and Croft, 2009; Xue *et al.*, 2010; Singh *et al.*, 2012). Selecting a subset of the original query (i.e., "sub-query") has been shown to be effective for improving these queries. Xue *et al.* (2010) formally model the distribution of sub-queries, where the sub-query selection procedure is modeled as a sequential labeling problem. A conditional random field model was applied to track the local and global dependencies between query words. Singh *et al.* (2012) present techniques to reduce long queries to effective shorter ones that lack superfluous terms. The authors describe a system that provides high quality product recommendations for null queries, where time-based relevance feedback is used to improve the fidelity of rewrites. However, these approaches focus only on the query space and

overlook critical information from the product space and the connection between the two spaces (Van Gysel *et al.*, 2016a).

To address the vocabulary gap problem, a series of studies have been proposed to match queries and product information in e-commerce search. Following Sarvi *et al.* (2020), matching solutions can be divided into *representation-based*, *interaction-based*, and *hybrid* approaches. All of them have been shown to be effective for matching consumers' queries and product information (Li *et al.*, 2019c; Yao *et al.*, 2022).

5.3.2 Representation-based matching

In early work, approaches to the task of *entity finding* have been applied to address the vocabulary gap problem between queries and products, where products are viewed as retrievable entities (Balog *et al.*, 2010; Gäde *et al.*, 2016; Van Gysel *et al.*, 2016a). There are two problems in entity finding for e-commerce search (de Vries *et al.*, 2007). First, entity finding retrieves entities of a particular type from multi-domain knowledge bases, whereas e-commerce search systems operate within a single but dynamic domain. Second, queries in e-commerce search contain a lot of free-form text (Rowley, 2000), whereas in entity finding most queries are semi-structured with relational constraints (Balog *et al.*, 2010).

To address the above two problems of matching optimization, representation learning methods have been applied to obtain better representations for each text associated with products by encoding both the query and the product title into single embedding vectors. Representation learning helps by flexibly adapting to the dynamic nature of e-commerce domains. It also effectively encodes the free-form text of queries and product descriptions into a common semantic space. Numerous neural text matching methods have been developed (Onal *et al.*, 2018; Mitra and Craswell, 2017; Lin *et al.*, 2021a). DSSM is one of the earliest deep learning-based models in text matching, in which each text is vectorized separately by a five-layer network (Huang *et al.*, 2013); CDSSM replaces the full connection layer with a convolution layer and a pooling layer to generate text vectors (Shen *et al.*, 2014). Hu *et al.* (2014) proposed ARC-I, where convolution operations represent two

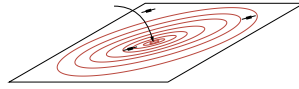


Figure 5.1: Illustrative example of how entities are ranked in vector space models w.r.t. a projected query. Image source: Van Gysel *et al.* (2016a).

concatenated texts for matching using a linear transformation. CNTN also adopts convolution neural network to represent two texts, and it proposes the neural tensor network to model the similarities between two texts (Qiu and Huang, 2015). MVLSTM obtains representations for each text and adopts an interactive method to measure similarities between two texts using 3 similarity operations, i.e., cosine, bilinear, and tensor layer (Wan *et al.*, 2016). As there are important differences between web search and e-commerce search, learning query and product representations is not a solved problem. Hence, to discriminate products based on textual descriptions, the importance of learning semantic representations of products was soon realized (Demartini *et al.*, 2009; Van Gysel *et al.*, 2016b).

Van Gysel *et al.* (2016a) propose an unsupervised distributed representation learning approach, namely latent semantic entities (LSE), to learn a unidirectional mapping between words and entities, as well as distributed representations of both words and entities. Given a set of entities, the authors assume that each entity has a set of associated documents. LSE then learns a function that maps a sequence of words in the query from the vocabulary to an entity vector space. Thereafter, cosine similarity is applied to calculate a relevance score between candidate entities and the query. In Figure 5.1, we see how entities are then ranked according to the projected query. Specifically, the authors take the representation of a string of words to be the average of the representations of the words it contains. Then, a projection matrix is employed to map the average one-hot representations to a distributed representation. Similarly, distributed representations of entities are mapped to the same space. Figure 5.2 provides a schematic overview of the LSE model. All the parameters are learned by using gradient descent. Based on the LSE model, an increasing number of representation learning

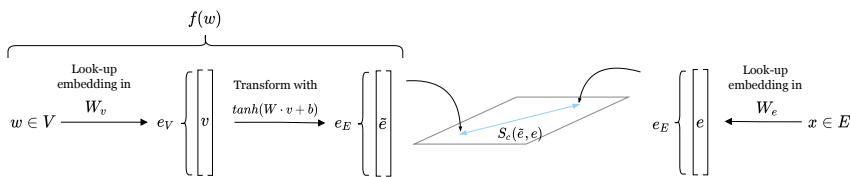


Figure 5.2: Schematic representation of the Latent Semantic Entities model for a single word w . Image source: Van Gysel *et al.* (2016a).

approaches have been proposed to address the vocabulary gap problem in e-commerce search (Xu *et al.*, 2018; Van Gysel *et al.*, 2018; Zhang *et al.*, 2019d).

In addition to learning semantic representations of products, e-commerce search, due to its specific task nature, also has multiple methods for optimizing matching results. *Substitutable* products are those that are interchangeable in e-commerce platforms (Wang *et al.*, 2018e). The substitutability relation among products can be determined in multiple ways. Van Gysel *et al.* (2018) find that product substitutability relations can facilitate the retrieval of relevant products impacted by the vocabulary gap problem, i.e., product substitutability can be integrated into product search either extrinsically or intrinsically. The authors propose a two-stage framework to combine a textual matching method (the LSE model) with substitutability to infer representations of queries and products. Unlike previous work on entity representation learning, the authors integrate relations among entities within the latent semantic space inference.

Users often browse multiple search results pages and make comparisons before purchase. Relevance feedback (RF) approaches have been proposed to extract the relevant topic (Jin *et al.*, 2013). However, the mismatch problem between queries and products still exists in the task of multi-page e-commerce search. Bi *et al.* (2019b) analyze different context dependency assumptions in multiple search result pages, and propose a context-aware embedding model to capture different types of dependency. The authors introduced three types of context dependency: (i) long-term context dependency, (ii) short-term context dependency, and (iii) long-short-term context dependency. Given these three types of

context dependency, a context-aware embedding model is proposed. By assuming that the users' preferences are associated with their implicit feedback, the embedding model, namely CEM, captures user preferences from their clicked items, which are implicit positive signals.

Another important aspect of e-commerce search concerns query reformulations by users. Based on query logs of eBay's search engine, Hirsch *et al.* (2020) offer a large-scale and in-depth study of users' query reformulations in e-commerce search. The authors analyze many aspects of search sessions composed of query reformulations, e.g., the number of reformulations and the distribution of their types, changes of search results pages as a result of the reformulations, clicks and purchases. An approach is proposed to predict if a query will be reformulated in an e-commerce search session. The authors find that post-retrieval features and query performance predictors contribute the most to the prediction of reformulation. By incorporating these features, the accuracy in predicting whether users will reformulate their queries can be significantly enhanced. Based on an attention mechanism, the MMAN model is proposed to enhance query representations by extending category information; it includes three main modules: self-matching, char-level matching, and semantic-level matching (Yuan *et al.*, 2023). Experiments show that these modules improve query representation, effectively handle long-tail queries, and achieve better semantic disambiguation.

Graph neural networks (GNNs) (Scarselli *et al.*, 2009; Kipf and Welling, 2017) have been applied to e-commerce search to help infer query and product embeddings. GNNs derive node representations by aggregating features appearing in the neighborhood. Niu *et al.* (2020) put forward a GNN-based method to learn representations of users and shops in e-commerce search. The authors propose a dual hierarchical graph attention network for e-commerce search. A heterogeneous graph is constructed to perform graph-based representation learning for both shops and queries, which includes both first-order and second-order proximities from various user interactions in e-commerce. The proposed method can help to relieve the semantic gap between user queries and shop names by borrowing item neighbor title text. The proposed neighbor proximity loss provides strong additional guidance for learning graph topological structure. A large-scale offline evaluation and online

A/B tests demonstrate the significant superiority of this approach. Chang *et al.* (2021) transfer the matching problem into an extreme multi-label classification problem (Yu *et al.*, 2022b), aiming to tag input instances (i.e., queries) with the most relevant output labels of products. The authors suggest a tree-based sparse linear model with n-gram TF-IDF features to augment the diversity of the matching results. For multi-lingual search scenarios in e-commerce, Lu *et al.* (2021) detail a graph-based model with a graph convolution layer to fill the vocabulary gap.

Transformer-based pre-trained models like BERT (Kenton and Toutanova, 2019) use stacked encoder layers that rely on a self-attention mechanism. BERT and BERT-like pre-trained models have been applied in product search (Peeters *et al.*, 2020; Liu *et al.*, 2022d; Qiu *et al.*, 2022; Wang *et al.*, 2023a). Qiu *et al.* (2022) apply dual-tower pre-training strategies to optimize both user intent detection and embedding retrieval in e-commerce search. Liu *et al.* (2022d) examine the performance of multiple pre-training embedding methods and observed that query representation learning remains a bottleneck compared to product representation learning when using these semantic search training objectives. Large-scale pre-trained language models, such as GPT-3 (Brown, 2020), also demonstrate promising performance across several benchmarks (Kim *et al.*, 2022a).

5.3.3 Interaction-based matching

Encoding queries and products in representation-based matching methods are independent of each other. Mapping individual queries and products into fixed-dimensional vectors may lose fine-grained matching information. To tackle this challenge, interaction-based matching has been proposed. This kind of method first matches different parts of the query with different parts of the document and then aggregates the partial evidence of relevance. In contrast to representation-based matching methods, interaction-based approaches usually build an interaction matrix between two documents and optimize it for matching (Hu *et al.*, 2014; Pang *et al.*, 2016). Interaction-based matching has first been applied in web-based retrieval. Hu *et al.* (2014) build an interac-

tion matrix to conduct several convolution and pooling operations to extract matching features. Guo *et al.* (2016) mention three factors in relevance matching – exact matching signals, query term importance, and diverse matching requirements – and design the architecture of their deep matching model. Similarly, MatchPyramid constructs an interaction matrix to capture matching patterns (Pang *et al.*, 2016). Mitra *et al.* (2017) employ both distributed representations and local representations to obtain the final matching score. Follow-up research has proposed a series of matching approaches specifically for e-commerce. Guo *et al.* (2019a) introduce a model based on MatchZoo. It is meant for short-text matching and replaces the matching histogram with a top- k max pooling layer. Li *et al.* (2020a) describe a product matching model, PMM, to make use of the information contained in titles and attributes of products. PMM consists of a product title matching module, and a product attributes matching module. Bi *et al.* (2020b) detail an end-to-end context-aware embedding model that can incorporate both long-term and short-term contexts to predict purchased items, unlike most approaches that focus on relevance feedback.

5.3.4 Hybrid matching

Several hybrid matching models have been proposed to combine the strengths of representation- and interaction-based models. Mitra *et al.* (2017) propose a matching model, DUET, that integrates both local (interaction-based) and distributed (representation-based) features to calculate query-document relevance. Pre-trained language models, such as BERT (Kenton and Toutanova, 2019), have further advanced hybrid approaches by capturing both contextualized token-level interactions and global semantic representations, achieving promising performance in tasks like search and recommendation (Sun *et al.*, 2019a; Nogueira *et al.*, 2019; Lin *et al.*, 2021a). Tracz *et al.* (2020) introduce a BERT-based model to use both types of matching in a similarity learning framework for product matching in e-commerce. Yao *et al.* (2022) explore the deployment of BERT in online retrieval systems by distilling it into a representation-based architecture, while still maintaining the advantages of interaction-based processing for more precise matching.

5.3.5 Matching in personalized search

One of the primary characteristics of e-commerce search is its highly personal variance in queries. First, multiple items could be topic-related with a consumer's query, but only a few are actually purchased, i.e., different individuals have different opinions even on the same product. Hence e-commerce search without personalization is unlikely to satisfy consumers. Second, personalization has explicit benefits for e-commerce platforms by exhibiting the products that consumers would like to purchase. As we have seen in Section 5.1, the definition of "relevance" for e-commerce search is not the same as for web-based search as most e-commerce platforms apply gross merchandise volume (GMV) as the gold standard for measuring success. To the best of our knowledge, Jan-nach and Ludewig (2017a) have been the first to attempt to personalize product search by using personalized recommendation approaches. However, the matching problem in personalized e-commerce search is more challenging as most platforms have approaches to matching products to queries that are far from perfect. To address this problem, an increasing number of matching studies have been proposed. Matching approaches in personalized product search can be classified into query-independent and query-dependent ones (Liu *et al.*, 2022a).

Query-independent matching. Query-independent matching methods embed users into a general profiling vector in the offline training stage (Ai *et al.*, 2017; Ai *et al.*, 2019b; Liu *et al.*, 2020c). Ai *et al.* (2017) design a deep neural network and jointly learn latent representations for queries, products, and users. A hierarchical embedding model is proposed for personalized e-commerce search. As illustrated in Figure 5.3, the authors project both queries and consumers into a single latent space and explicitly control their weights in a personalized product search model. Following Van Gysel *et al.* (2016a), the authors design latent representations of queries and users to have good compositionality so that the personalized search model can be directly computed as a linear combination of query models and user models. Both queries and users are projected into a single latent space. Given a query q , the corresponding query intent is represented in R^α ; similarly, the user preference is represented in R^α given a user u . As shown in Figure 5.3, the

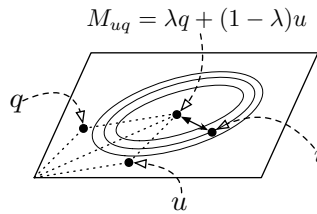


Figure 5.3: Personalized product search in a latent space with query q , user u , personalized search model M_{uq} and item i . Image source: Ai *et al.* (2017).

personalized product retrieval model is defined as $M_{uq} = \lambda q + (1 - \lambda)u$, where λ is a hyper-parameter that controls the weight of the query model q and the user model u . To alleviate the mismatch problem during personalized product search, the authors put forward a hierarchical embedding approach to reflect the distributed representations of users, items, and queries. Their experiments are conducted with synthetic queries generated from product category information.

External structured knowledge graphs have been applied to enhance the personalized matching procedure. Structured relationships among users, products, and queries have been jointly considered in graph neural network approaches. Ai *et al.* (2019b) introduce a unified knowledge graph on multiple types of product data and conduct retrieval with it. A dynamic relation embedding module constructs a session-based knowledge graph and a soft matching algorithm extracts explainable paths with knowledge embeddings. Liu *et al.* (2020c) exploit the structured representation learning scheme from user-query-product interactions with conjunctive graph patterns. Geometric operations, such as projections and intersections, are applied in the proposed graph neural networks. Derived from knowledge graph embeddings, Liu *et al.* (2022a) use multiple vectors to encode the diverse preferences of users. The authors used the category information to aggregate the multiple interests of users with category indications as references. To exploit collaborative signals among products, users, and queries, Cheng *et al.* (2022) offer a hypergraph-based method from the ternary user-product-query interactions, by considering high-order features of neighbors. Query-independent matching models can calculate user embeddings and store

them in advance, which makes it convenient and efficient to apply in real-world search engines. To tackle inconsistent user behavior in multi-stage e-commerce search systems, Wang *et al.* (2023a) employ external information to refine query-item matching. By mining various user interactions (ordered, clicked, unclicked items) within a post-fusion strategy, they generate more accurate semantic representations. This approach not only enhances retrieval efficiency, but also improves both offline recall and online conversion rates.

Query-dependent matching. To capture users' dynamic interests given a query, query-dependent matching approaches have been proposed. To address the problem of when and how to conduct search personalization in product retrieval, Ai *et al.* (2019a) conduct an empirical analysis of the potential of personalization in product search with large-scale search logs sampled from a real-world e-commerce search engine. To analyze query specificity, the authors compute the purchase entropy of each query in the sampled e-commerce search logs as $Entropy(q) = -\sum_{i \in I_q} P(i|q) \log P(i|q)$, where I_q refers to the candidate item set for a query q . In Figure 5.4(a), the authors provide the purchase entropy of queries on *Beauty* products (e.g., facial cleanser) in the sampled search logs. They rank queries according to their frequencies on a logarithmic scale and split them into three groups: the queries with low frequency (LowFreq), with medium frequency (MedFreq), and with high frequency (HighFreq). Queries with high frequencies have more potential for personalization, as more purchases on different items can be observed when the number of sessions increases. The authors evaluate the differences between $P(i|q, u)$ and $P(i|q)$ by using the familiar *MRR* metric, i.e., $MRR(q) = \sum_{u \in u} RR(P(i|q), P(i|q, u)) \cdot P(u)$, where $RR(P(i|q), P(i|q, u))$ reflects the the reciprocal rank of a ranked list produced by ranking with $P(i|q)$, using $P(i|q, u)$ as the ground truth. In Figure 5.4(b) the authors show the $MRR(q)$ of queries on *Beauty* products; the similarity between $P(i|q, u)$ and $P(i|q)$ is not monotonically correlated with query frequency. Accordingly, the potential of personalization varies significantly in different queries, which requires sophisticated models for personalization in product search.

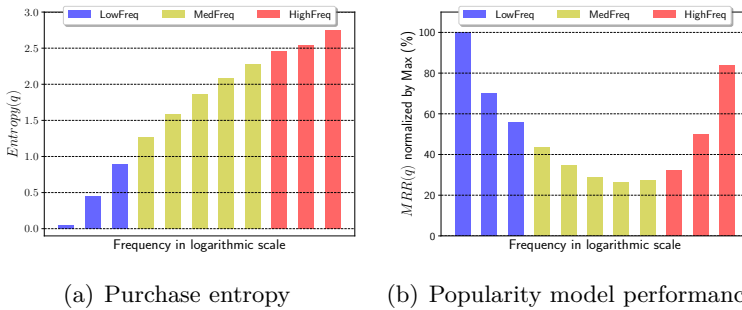


Figure 5.4: The purchase entropy $Entropy(q)$ and popularity model performance in $MRR(q)$ for queries with different frequencies. Image source: Ai *et al.* (2019a).

To tackle this challenge, Ai *et al.* (2019a) introduce a zero-attention neural network model (ZAM) for personalized product search that conducts differentiated personalization for different query-user pairs. This network builds on an embedding-based generative framework. Query-dependent personalization is implemented by constructing user profiles with a zero attention strategy that enables it to automatically decide when and how to attend in different search scenarios. Likewise, Guo *et al.* (2019b) design a dual attention-based network to capture users’ current search intentions and their long-term preferences. Bi *et al.* (2019b) study relevance feedback based on both long-term and short-term context dependencies in multi-page product search with an end-to-end personalized product search model. Xiao *et al.* (2019) propose a streaming Bayesian method to explicitly and collaboratively learn representations of different categories of entities in a joint metric space over time.

Sparseness is another critical challenge in tracking users’ sequential behaviors in product search. Graph-based methods have been proposed to tackle this challenge. By using short-term user behavior, Fan *et al.* (2022) extend a structural relationship representation learning scheme (Liu *et al.*, 2020c) to explore both local and global user behavior patterns. Zhu *et al.* (2022b) integrate cross-domain transfer learning with a knowledge graph to establish the underlying interest correlation. Their proposed method performs interest alignment across domains by

explicitly modeling the long-term and short-term interactions between users and items, which capture the dynamics of product properties and user interests.

Pre-trained language models have also been applied to query-dependent matching in personalized product search. To conduct differential personalization in different contexts, Bi *et al.* (2020a) propose a transformer-based embedding model (TEM); in TEM, personalization can vary from no to full effect. In contrast with ZAM, TEM takes into consideration the interactions between purchased items so that it learns better dynamic representations of queries and items, which leads to better attention weights in personalized product search. Based on TEM, a review-based transformer model, abbreviated as RTM, is designed to match user intents and items at the level of finer-grained information (e.g., their associated reviews) (Bi *et al.*, 2021). The authors conduct user-item matching at the review level so that the reason an item is ranked at the top can be explained; also, the importance of each user and item review during matching is dynamically adapted. RTM represents users and items based only on their reviews without the need for their identifiers, which improves the generalization ability during product search. Dai *et al.* (2023) describe a contrastive learning framework, CoPPS, to enhance user representations for personalized product search. By pre-training a sequence encoder with contrastive sampling and fine-tuning, CoPPS employs multiple data augmentation strategies to improve user modeling.

For real-world applications, Zhang *et al.* (2020a) jointly consider semantic matching and personalized matching at JD.com. The authors introduce a deep personalized and semantic retrieval model (DPSR) with a two-tower architecture: a multi-head design of a query tower and an attention-based loss function. Similarly, Magnani *et al.* (2022) describe a semantic retrieval model with a two-tower architecture in e-commerce search production at Walmart.com. The authors select negative examples for training a large semantic retrieval model and use an approximate metric to evaluate the performance.

5.4 Ranking Strategies in E-commerce Search

Ranking optimization is another core task in e-commerce search. In this section, we introduce ranking strategies in e-commerce search.

As we have discussed in Section 5.1.1, learning to rank approaches have been applied to e-commerce search. Unlike web search scenarios based on relevance judgments, e-commerce search has multiple implicit and explicit signals that need to be integrated during the ranking process. In e-commerce search, learning to rank methods need to tackle a series of challenges (Karmaker Santu *et al.*, 2017). Learning to rank methods assume that the scoring of items to be ranked is a parameterized function of multiple features computed based on the given query and the items, whereas parameters are used to control the weights of features. A large number of product and query features, such as brands, rating, categories, etc., are important to obtain useful representations in e-commerce learning to rank models. Also, product popularity related features have been shown to be effective in optimizing ranking results of e-commerce search. Karmaker Santu *et al.* (2017) study the performance of multiple learning to rank strategies in e-commerce search, and find that LambdaMART (Burgess, 2010) is able to learn a balance between these two kinds of features. Detailed comparisons are listed in Table 5.1. In addition, user engagement behavior, such as clicks and orders, also plays an important role in optimizing ranking results (Karmaker Santu *et al.*, 2017; Wu *et al.*, 2018a). In contrast to web search which only has clicks to judge the relevance, e-commerce search contains four prominent relevance feedback behavior: clicks, cart-adds, orders, and revenue. Thus, various training objectives can be considered in optimizing e-commerce ranking results.

Another main challenge for e-commerce learning to rank methods is the lack of labeled information. In web search, high-quality labeled information obtained by eliciting relevance ratings from human experts or crowdsourcing, makes learning to rank methods effective. However, in the context of e-commerce search it is infeasible to determine a standard method to obtain ground-truth information. Intuitively, crowd sourcing annotations seem to provide labels for e-commerce learning to rank methods. Karmaker Santu *et al.* (2017) and Alonso and Mizzaro (2009)

Table 5.1: Comparison of ranking algorithms in terms of NDCG@10 for target variable “Click Rate”, “Cart Add Rate”, “Order Rate”, and “Revenue.” Regu. is an abbreviation for Regulation, whereas lo. is an abbreviation for loss. Table source: Karmaker Santu *et al.* (2017).

Algorithm	Click Rate		Cart Add Rate		Order Rate		Revenue	
	Train	Test	Train	Test	Train	Test	Train	Test
RankNet (Burgess <i>et al.</i> , 2005)	0.6857	0.6855	0.4399	0.4402	0.7158	0.7142	0.7577	0.7578
RankBoost (Freund <i>et al.</i> , 2003)	0.5899	0.5904	0.4073	0.4043	0.5007	0.4994	0.5663	0.5639
AdaRank (Xu and Li, 2007)	0.6877	0.6857	0.4464	0.4401	0.7334	0.7349	0.757	0.7566
Random Forest (Breiman, 2001)	0.6378	0.6125	0.4588	0.4296	0.5707	0.5288	0.6463	0.5959
LambdaMART (Burges, 2010)	0.8426	0.8291	0.7664	0.7324	0.7728	0.7687	0.8183	0.7998
Logistic Regression (L1 regu.) (Fan <i>et al.</i> , 2008)	0.6284	0.6272	0.4274	0.4252	0.6677	0.6632	0.6873	0.6822
Logistic Regression (L2 regu.) (Fan <i>et al.</i> , 2008)	0.5889	0.5866	0.4066	0.4025	0.5045	0.4983	0.5751	0.5675
SVM Classifier (L1 regu.+L2 lo.) (Fan <i>et al.</i> , 2008)	0.6366	0.6317	0.4348	0.4331	0.6870	0.6794	0.7105	0.7059
SVM Classifier (L2 regu.+L1 lo.) (Fan <i>et al.</i> , 2008)	0.4596	0.4594	0.3274	0.3219	0.4281	0.4289	0.4503	0.4462
SVM Regressor (L2 regu.+L2 lo.) (Smola and Vapnik, 1997)	0.2358	0.2341	0.1909	0.1914	0.2100	0.2087	0.2030	0.2027
SVM Regressor (L2 regu.+L1 lo.) (Smola and Vapnik, 1997)	0.2876	0.2865	0.2110	0.2096	0.2078	0.2038	0.2093	0.2121

study the reliability of the relevance judgements provided by crowd workers for e-commerce queries. The authors find that crowdsourcing fails to provide reliable relevance judgements for e-commerce queries. Thus implicit feedback from e-commerce users is an important source of information that helps reveal the saliency of products for a given query.

Wu *et al.* (2018a) divide the ranking optimization problem into two successive stages: click optimization and purchase optimization. We have discussed previous studies on click modeling and purchase modeling in Section 4.1.1 and 4.1.3, respectively. Inspired by these approaches, Wu *et al.* (2018a) apply a list-wise learning to rank model (Cao *et al.*, 2007) to optimize clicks by jointly considering item positions and query-level structures; a binary classification approach is applied to predict the purchase behavior.

Online learning to rank approaches have been applied to optimize ranking results in e-commerce search (Hu *et al.*, 2018b). Unlike traditional static learning to rank models, online learning to rank methods optimize the production ranker interactively by exploiting users' implicit feedback (Zoghi *et al.*, 2015; Oosterhuis and de Rijke, 2017; Schuth *et al.*, 2016). Online learning to rank approaches can be divided into two groups: the first is to learn the best ranking function from a function space (Hofmann *et al.*, 2016; Hofmann *et al.*, 2013; Yue and Joachims, 2009); the second group directly learns the best list under some model of user interactions (Radlinski *et al.*, 2008; Schuth *et al.*, 2016; Oosterhuis and de Rijke, 2017). As part of the first group of online learning to rank methods, Hu *et al.* (2018b) propose a reinforcement learning method for ranking optimization in e-commerce searching scenarios. The authors formulate the multi-step ranking procedure in e-commerce search as a search session Markov decision process (SSMDP). An algorithm, named deterministic policy gradient with full backup estimation (DPG-FBE), is then proposed for the problem of high reward variance and unbalanced reward distribution of SSMDP. To integrate search results from heterogeneous sources, Takanobu *et al.* (2019) introduce a search result aggregation method that formulates a semi-Markov decision process, where a low-level policy is applied to represent items and a high-level policy is used to select rankers. Based on a large-scale real-world dataset,

Anwaar *et al.* (2020) employ a counterfactual risk minimization (CRM) approach to directly optimize the ranking list from the log data.

In recent years, deep neural-based retrieval models (Mitra and Craswell, 2017; Kenter *et al.*, 2017; Onal *et al.*, 2018) have been used in the ranking step of e-commerce search. Magnani *et al.* (2019) enhance the deep neural network model using different types of text representation and loss function at Walmart.com. Zhang *et al.* (2019d) detail an e-commerce ranking model with interaction features between the query and a graph of products maintaining product interactions. Pre-trained foundation models have been shown to be effective in real-world web search scenarios (Zou *et al.*, 2021; Lin *et al.*, 2021a; Chu *et al.*, 2022). In e-commerce search, Wu *et al.* (2022a) apply BERT for product ranking within a multi-task learning framework. The authors use the probability transfer method in the framework to model multiple sequential engagement behaviors. By integrating semantic matching features output by the domain-specific BERT, the authors confirm the effectiveness of their proposed approach on real-world e-commerce search data.

5.5 Emerging Directions

5.5.1 Multi-modal e-commerce search

Along with the rise of deep neural networks, multi-modal search has increasingly received attention (Zhang *et al.*, 2013; Mao *et al.*, 2015; Yang *et al.*, 2017a; Liu *et al.*, 2018a; Balaneshin-kordan and Kotov, 2018; Guo *et al.*, 2018; Zhang *et al.*, 2018a; Qu *et al.*, 2021; Wei *et al.*, 2022; Tan *et al.*, 2022). The key of multi-modal search is to find an effective mapping mechanism to project data from different modalities into a common latent space. Multi-modal search approaches can be classified into hashing-aware models and semantic-aware models. The former type of methods map various modalities in the original space to a Hamming space using hash functions (Cao *et al.*, 2017b; Luo *et al.*, 2018; Zhu *et al.*, 2020a; Tan *et al.*, 2022), whereas the latter ones project the multi-modal data into a low-dimensional space by learning a mapping function (Wang *et al.*, 2016a; Laenen *et al.*, 2018; Qu *et al.*, 2021; Wang *et al.*, 2022c).

As we discussed in previous sections, most approaches to e-commerce search focus on the textual matching problem. With the rise in online photos and openly available image datasets this is changing. Yang *et al.* (2017a) propose a novel end-to-end approach for scalable visual search infrastructure at Ebay. Similar platforms can be found at Alibaba (Zhang *et al.*, 2018a) and JD.com (Li *et al.*, 2018a; Wang *et al.*, 2020b). Early studies into multi-modal e-commerce search have mostly focused on the fashion category (Yang *et al.*, 2017a; Zhang *et al.*, 2018a; Li *et al.*, 2018a). However, Wang *et al.* (2020b), Dagan *et al.* (2021), and Liu *et al.* (2022b) confirmed that multi-modal search is widespread across many e-commerce categories, especially categories that involve aspects that are harder to express verbally, but can naturally be captured visually, such as style, type, and pattern.

In multi-modal search, multiple modalities, such as text and images, can be found in both queries and products. Most multi-modal search solutions optimize a function to project multi-modal data into a low-dimensional space after unified representation learning. Laenen *et al.* (2018) present a multi-modal search paradigm for e-commerce search that results in an improved shopping experience. The authors reason with both images and languages through a common embedding space. Guo *et al.* (2018) formulate the e-commerce personalized search problem based on the relevance between images and text with respect to the query and the user preferences from both textual and visual modalities. A transition-based product search method has been proposed, where the multi-modal feature space is initialized based on the textual and visual features of products. An interpretable multi-modal e-commerce retrieval framework has been proposed for fashion products (Liao *et al.*, 2018); the authors bridge the gap between deep features and meaningful fashion concepts. They propose a hierarchical similarity function to accurately characterize the semantic affinities among fashion items.

Dagan *et al.* (2021) highlight various differences between visual and textual search, which can be summarized in the following challenges in multi-modal e-commerce search:

- **Heterogeneous resources.** User queries submitted on e-commerce platforms can be real-world images shot while the user is engaging with the platform.

- **Large-scale sparse data.** Large multimedia corpora make scalability and efficiency key requirements for e-commerce search. Visual queries are more specific than text queries, which results in a smaller number of retrieved results and sparse coverage of categories.
- **Limited user engagement.** In multi-modal e-commerce search, user engagement behavior, e.g., clicks, dwell time, purchases, etc., is substantially sparser than in textual search.
- **Ambiguous user intents.** Two different main use cases exist in multi-modal e-commerce search: target finding and decision making (Dagan *et al.*, 2021). The two use cases reflect the navigational intent and informational intent, respectively.

Additionally, annotating new product images and retraining a new feature representation model on large-scale data is expensive. To address this problem, few-shot multi-modal product search has been proposed to update the feature representation and index model with few-shot data and employ a fast learning strategy for new categories. Wang *et al.* (2020b) describe a framework for few-shot incremental search via meta-learning, with a multi-pooling feature extractor to extract discriminative multi-modal features. Based on pre-trained language models, Liu *et al.* (2022b) detail an effective contrastive learning framework to learn representations of multi-modal search sessions based on multi-view heterogeneous graph networks. Liu *et al.* (2023b) employ BERT with a self-distillation framework for product understanding by integrating visual and textual information.

Cross-modal search is receiving more and more attention (Qu *et al.*, 2021; Wang *et al.*, 2022c; Tan *et al.*, 2022). Unlike multi-modal search, cross-modal search aims to solve the discrepancy problem between different modalities in search sessions (Wang *et al.*, 2022c). Although much progress has been made in bridging multiple modalities, it still remains challenging because of the difficult intra-modal reasoning and cross-modal alignment (Qu *et al.*, 2021). As complicated relations exist among products in e-commerce, it is more difficult to explore these multi-hop interactions in cross-modal product search scenarios. Moreover, the extremely high computational costs brought by multi-modal input

limits its usefulness in large-scale cross-modal retrieval in e-commerce applications.

5.5.2 Conversational e-commerce search

Originating from early studies on interactive information retrieval (Croft and Thompson, 1987; Belkin *et al.*, 1995), conversational search refers to the process of interacting with a dialogue system to search for information. Conversational search allows users to express their information needs by directly conducting conversations with search engines. Unlike traditional query-based search engines, conversational search systems capture users' intent by taking advantage of the flexibility of mixed-initiative interactions and by providing useful information more directly using human-like responses (Radlinski and Craswell, 2017; Vtyurina *et al.*, 2017; Ren *et al.*, 2021; Vakulenko *et al.*, 2021). User studies have been conducted to study whether conversational search is needed and what it should look like. Trippas *et al.* (2018) conduct a laboratory-based observational study, where pairs of people perform search tasks communicating verbally. The authors find that conversation search paradigms are more complex and interactive than traditional search scenarios. Moreover, it is difficult to simulate human-human interactions in a conversational search session. To address this problem, Ren *et al.* (2021) collect a dataset of human-human dialogues about conversational search in a wizard-of-oz fashion, namely wizard of search engine (WISE), where two workers play the role of seeker and intermediary, respectively.

In a conversational search session, the user first initializes the conversation with a request, then the conversational agent iteratively asks the user about their preferences and estimates user interest based on their feedback. Finally, the agent retrieves the information and generates the response. Many studies focus on the second stage, including generating clarifying queries (Zamani *et al.*, 2020; Liu *et al.*, 2021e; Ghanem *et al.*, 2022) and understanding users' intent (Wu and Yan, 2019; Gao *et al.*, 2021a; Trippas *et al.*, 2020). Radlinski and Craswell (2017) consider the question of what properties would be desirable for a system so that the system enables users to answer a variety of information needs in a natural and efficient manner. Azzopardi *et al.* (2018) outline the

actions and intents of users and systems and explain how these actions enable users to explore the search space and resolve their information needs. Ghanem *et al.* (2022) propose a method to generate user queries for story-based reading comprehension skills. The response generation problem in conversational search is also being addressed. Ren *et al.* (2021) describe a modular end-to-end neural architecture to transfer the output from intent understanding to improve response generation. Ye *et al.* (2022) design an interconnected network to co-generate structured and natural responses that allow for bidirectional semantic associations to generate responses.

In e-commerce, conversational search systems can help consumers access products through instant conversational interactions on mobile phones or other smart devices. Also, conversational e-commerce search alleviates the burden of reformulating queries and browsing through dozens of products in e-commerce search (Zou *et al.*, 2022a). Moreover, conversational e-commerce search provides a natural way to collect explicit feedback from users to understand their preferences (Zamani *et al.*, 2022). As an early attempt, Zhang *et al.* (2018c) detail a paradigm for conversational product search to ask users about their preferred values of an aspect and adopt a memory network to retrieve search results. Based on that, Bi *et al.* (2019a) focus on product-seeking conversations and proposed an aspect-value likelihood model for negative feedback, with a multivariate Bernoulli distribution to generate explainable e-commerce aspects. To improve the quality of representation in conversational product search, Zou *et al.* (2022a) integrate the representation learning of user, query, item, and conversation into a unified generative framework.

There are several open questions and research directions for future work in conversational e-commerce search. First, obtaining data in real-world scenarios is a challenging problem. Several studies, including, e.g., ConvPS (Zou *et al.*, 2022a), have applied simulated data in their experiments. It would be useful to extend observational experiments to a wizard-of-oz setting by establishing a real-world dataset in the future. We also observe that the ongoing trend in simulating users in task-oriented dialogues could provide more useful insights (Sun *et al.*, 2021b). Second, modeling user-system interactions is a crucial aspect of conversational e-commerce search. Vakulenko *et al.* (2021) reveal the complicated

situation for large-scale dialogue analysis specifically focusing on the patterns of mixed-initiative. In contrast with traditional web search scenarios, e-commerce search needs to analyze more complicated user engagement behavior (see Section 5.1.2 and 3.2). Hence, in future work, both short-term and long-term user engagement interactions may need specialized attention in conversational e-commerce search. Third, how to evaluate the performance of conversational e-commerce search is still an open question. We have demonstrated a wide range of evaluation metrics applied in e-commerce search, e.g., engagement-based metrics and revenue-based metrics (see Section 5.2). Therefore, it is even more challenging to measure the success of conversational e-commerce search in terms of all these various kinds of metrics. Measuring interactivity is critical in conversational information seeking (Zamani *et al.*, 2022). Keeping track of how well a user understands the system and vice versa can be another important research direction. Fourth, personalization conversational e-commerce search needs more attention in future work.

5.5.3 Generative retrieval in e-commerce

Generative retrieval has emerged as a novel retrieval paradigm in recent years. During the training phase, documents and their corresponding document identifiers (docids) are encoded into the model's parameters. Given a query, the model can directly generate the relevant DocIDs (Tay *et al.*, 2022). This approach uses the model's ability to memorize and use document representations, enhancing retrieval efficiency and accuracy by bypassing traditional retrieval pipelines.

For indexing strategies, the approach typically involves using a sequence-to-sequence (seq2seq) model to learn the mapping from queries to docids. In addition, a range of tasks have been introduced to enhance the indexing performance, including learning a mapping from documents to docids (Tay *et al.*, 2022), constructing pseudo-queries for documents and mapping them to docids (Wang *et al.*, 2022b), as well as training the model to rank different documents corresponding to the same query (Li *et al.*, 2024).

Document representations are another key component of the generative retrieval model. Since docids are the direct output of model,

the quality of these representations directly determines the prediction accuracy of generative models. Existing docids can generally be categorized into two types: numeric-based and text-based. Tay *et al.* (2022) introduce three types of numeric docids: (i) unstructured atomic docids, where each document is assigned a random and unique integer identifier, without any structural or semantic information; (ii) naively structured string docids, where each document is assigned a random and unique numeric string; and (iii) semantically structured docids, which use a hierarchical k -means method, allowing relevant documents to share the same prefix, thereby introducing semantic structure. Wang *et al.* (2022b) propose that the meaning of a docid depends on both the position and the prefix context. Therefore, they propose a prefix-aware weight-adaptive decoder to adapt to different docids. Sun *et al.* (2024) detail learnable document representations by using a discrete autoencoder to encode documents into short, discrete docids. It optimizes these docids by converting documents into docids through an encoder, then training the model to minimize the reconstruction loss of converting docids back to the original documents. Text-based docids are another popular approach, as they can inherit the language model’s text generation ability and do not require learning a new docid vocabulary. The title can be regarded as an intuitive abstract text that represents the content of a document. De Cao *et al.* (2020) describe how to use the document title as a docid and achieve good retrieval performance. However, an arbitrary document may not have an informative and structured title like those in Wikipedia; Zhou *et al.* (2022b) combine keywords from both the URL and the title to form a docid. Additionally, some work has attempted to incorporate more content into text-based docids, including titles, URLs, document substrings, pseudo-queries, and more (Li *et al.*, 2023c). The substrings in the document can also represent the information stored in the document. Bevilacqua *et al.* (2022) use arbitrary n -grams in documents as docids, and retrieve documents using a pre-built FM-indexer.

6

E-commerce Recommendation

Recommendation methods refer to information filtering techniques, which can be traced back to the 1980s (Malone *et al.*, 1987), i.e., even before the web had been developed. Significant research efforts have been devoted to recommendation in various domains, such as email, movies, books, and music. E-commerce has its unique properties; recommendation methods developed for other domains may not be suitable for e-commerce scenarios.

In this section, we focus on recommendation techniques for e-commerce. First, we summarize the key characteristics of e-commerce recommendation, for which a two-stage framework is developed. This forms the mainstream solution for e-commerce recommender systems. We then review models developed for the two stages. We discuss evaluation methodologies for e-commerce recommendation and conclude with a description of future research directions to help build better e-commerce recommendation solutions.

6.1 Characteristics of E-commerce Recommendation

Recommendation techniques have been applied extensively in e-commerce, in many different scenarios. They play a crucial role in e-

commerce by facilitating users to find desired products and to help boost revenue. E-commerce has three key characteristics that affect recommendation techniques:

- **Large volume of products.** The first key characteristic of e-commerce is the large volume of products, which brings challenges to algorithm scalability, as recommendation algorithms need to quickly scan products and select the ones that are of interest to a user. It is common that an e-commerce platform contains millions of products (Kersbergen and Schelter, 2021). This sheer number of items inevitably poses a computational challenge to recommender systems. The training of models needs to be efficient to be able to quickly (e.g., hourly or at least daily) refresh the model given new user behavior data and latency needs to be sufficiently low in order to be able to expose the large catalog to a large number of users.
- **Sparsity.** The second characteristic is the sparsity of behavior, since a user can only consume (e.g., purchase or click) a few products. The main aim of e-commerce recommendation is to satisfy the information needs of users in viewing or purchasing products and services. Behavioral data (e.g., browsing, purchases, and clicks) is an important and useful data source for learning the personalized tastes of users. Given the huge volume of products, it is impossible for a user to interact with most items. The products with which a user actually interacts are typically just a small fraction of the entire catalog (Li *et al.*, 2023b). This results in a significant sparsity issue that forces recommendation methods to learn from user behavior on a limited set of products and to generalize their predictions to all other products.
- **Data richness.** The third characteristic is the richness of product and user data (see Section 3). In addition to user behavior data that directly reflects user preferences, rich information about products (e.g., product name, description, categories, images, etc.) and users (e.g., age, gender, occupation, income level, etc.) is available in e-commerce scenarios. Moreover, there is much contextual information associated with user behavior, such as time, location, last purchase, and submitted query in the session

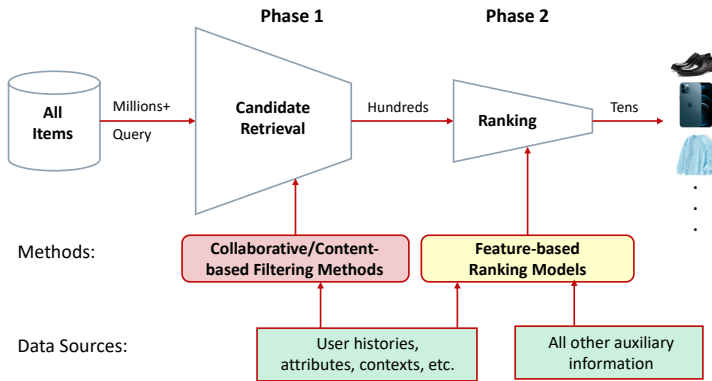


Figure 6.1: Illustration of a two-stage recommendation solution with candidate retrieval and candidate ranking. Image source: Ren *et al.* (2018).

etc. (Chen *et al.*, 2017d). This auxiliary information is useful for inferring why a user chooses an item, and is particularly beneficial for cold-start scenarios, where a user (or an item) has very few interactions.

As we will see below, many developments in e-commerce recommendation have addressed the three characteristics listed above, resulting in a large number of technical achievements that can be directly put to practical use.

We present a two-stage solution that has been commonly used in industrial recommender systems (Cheng *et al.*, 2016; Wang *et al.*, 2018e; Covington *et al.*, 2016). Figure 6.1 illustrates a two-stage recommendation framework with two main phases: *candidate retrieval* and *candidate ranking*. Given an information need (either expressed explicitly, e.g., by a user expressing interest in a product category, or implicitly, e.g., through the event of a user visiting a recommender system), the first phase of candidate retrieval goes through the entire product catalog, which may contain millions of products, and selects a small set of products that best match the information need. Then the selected candidates (usually in the order of hundreds) are passed to the second phase of candidate ranking, which ranks the candidates to produce the final top- k products to the user. Such a two-stage architecture can strike a balance between efficiency and effectiveness – it not only supports fast retrieval from

a large-scale catalog by using an efficient and light-weight candidate retrieval model, but also maintains good recommendation performance as the final ranking is determined using a potentially fine-grained ranking model. We survey studies into each stage in Section 6.2 and 6.3, respectively.

It is worth mentioning that the pipeline sketched in Figure 6.1 may involve a third stage, re-ranking, to refine the list produced by the ranking model to meet additional criteria or constraints, such as diversity, novelty, or fairness (Abdollahpouri, 2019; Wilhelm *et al.*, 2018; Gogna and Majumdar, 2017; Pei *et al.*, 2019; Ai *et al.*, 2018). We briefly introduce recent progress in re-ranking models in Section 6.4.

6.2 Candidate Retrieval Models

In this section, we survey candidate retrieval models ranging from traditional heuristic methods (Section 6.2.1) to recent advanced embedding-based methods (Section 6.2.2).

6.2.1 Heuristic methods

Heuristic-based methods are commonly used for candidate retrieval because they are simple and easy to implement. These methods are based on a manually heuristic rather than on optimization with an objective function. Early studies have presented various retrieval strategies to search candidate products. For example, selecting high sale or promotion items has been widely adopted in practical recommender systems, whereas several researchers use principles from economics to perform item selection for candidate retrieval (Zhao *et al.*, 2017; Zhang *et al.*, 2016b).

There are many heuristic methods that aim at mining item or user relations for candidate retrieval. We detail approaches to these methods from three perspectives: (i) neighborhood-based methods, (ii) graph-based method and (iii) methods based on complementary and substitutable items.

Neighborhood-based methods. Neighborhood-based methods first compute similarity between items or users, and then make recom-

mendations by aggregating information from similar users or items. Neighborhood-based methods can be classified into *user-oriented* and *item-oriented*. The paradigm of user-oriented methods (Resnick *et al.*, 1994; Konstan *et al.*, 1997) can be summarized as follows (Adomavicius and Tuzhilin, 2005):

$$\hat{r}_{u,i} = \text{Agg}(\{r_{u',i}\}_{u' \in S_u}). \quad (6.1)$$

where $\hat{r}_{u,i}$ is the rating of user u for the target unrated item i that we seek to estimate, S_u is the set of similar users of u , and $r_{u',i}$ is the rating of similar user u' on the target item i . $\text{Agg}(\cdot)$ is a function that aggregates information from similar users. The process consists of two steps: (i) finding similar users, and (ii) aggregating the information of similar users. Recent work has explored various approaches for similarity computation, including but not limited to Pearson correlation coefficients, information entropy, and mean squared difference (Shardanand and Maes, 1995) between the ratings given by two users (Resnick *et al.*, 1994; Shardanand and Maes, 1995). Spearman rank correlation has been used to measure item rank similarity rather than value similarity (Herlocker *et al.*, 1999). For the aggregation function $\text{Agg}(\cdot)$, a linear weighted combination is a commonly used strategy (Resnick *et al.*, 1994; Shardanand and Maes, 1995; Herlocker *et al.*, 1999). The similarity score is usually used for setting combination weights, as similar users naturally deserve a larger contribution to a prediction.

Inspired by the success of user-oriented methods, item-oriented methods have also been explored (Sarwar *et al.*, 2001; Karypis, 2001). Item-oriented methods have a quite similar process as user-oriented methods except that they aim at mining item similarity rather than user similarity. Recent studies empirically show that item-oriented methods achieve better performance than user-oriented methods (McLaughlin and Herlocker, 2004; Sarwar *et al.*, 2001; Karypis, 2001).

In summary, neighborhood-based methods have shown several advantages: they are simple, efficient, highly explainable, and easily deployed. However, their disadvantages are also obvious: they have a heavy reliance on human expertise, lack of flexibility, and they suffer from data sparsity. Although they may not perform as well as recent, more ad-

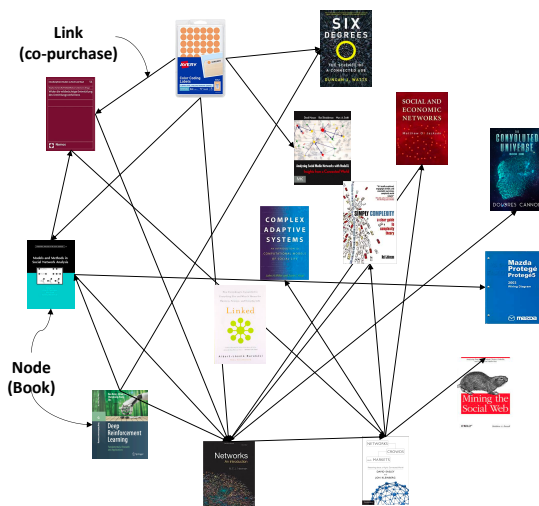


Figure 6.2: An item graph. Each node represents an item and each edge indicates the existence of a co-occurrence relation between the two items. Image source: Leem and Chun (2014).

vanced methods, they usually serve as a benchmark for candidates generation in real-world recommender systems.

Graph-based methods. Another type of heuristic strategy is to explore similar items through an item graph. As shown in Figure 6.2, the graph can be constructed from item-item co-occurrence or user-item interaction information, where each node represents a user or an item while each edge indicates a certain relation between these objects. Based on the graph, it is easier and more effective to evaluate the similarity between items via the closeness of two items in the graph. This has inspired a number of recent publications on graph-based retrieval. For example, Leem and Chun (2014) conduct information propagation on the graph to calculate item similarity, whereas Eksombatchai *et al.* (2018) directly perform random walks from seed nodes to select relevant items.

Complementary and substitutable items. Beyond item similarity, more complicated item relations concerning complementarity of items and substitutability of one item for another have also been considered (Wang *et al.*, 2018e; Zhang *et al.*, 2018b; Chen *et al.*, 2020d; Liu

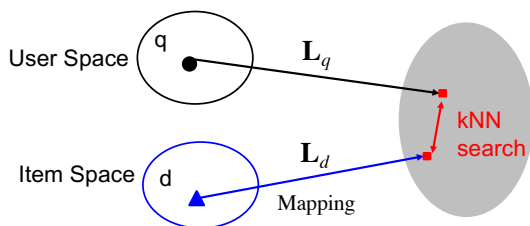


Figure 6.3: Illustration of the workflow of embedding-based methods: They first map users and items into a common embedding space; then they use the KNN algorithm to search the candidate items having the smallest embedding distance with the target user. Image source: Andoni and Indyk (2006).

et al., 2020d). As explained in Section 3.1.2.2, complements indicate items that might be purchased together, while substitutes indicate items that are interchangeable. Mining complements and substitutes is beneficial to satisfy the true need of users, and further increases the click-through rate and user stickiness (Wang *et al.*, 2018e; Liu *et al.*, 2020d). However, it is challenging as we often lack ground-truth labels of such relations. To deal with this problem, Wang *et al.* (2018e) use co-view and co-purchase statistics, as weak relation signals to supervise the learning of the complements and substitutes. They further integrate the learned relations into a vanilla recommendation model and observe improvements in recommendation performance. Liu *et al.* (2020d) introduce a graph convolutional neural network that decouples item semantics for inferring complementary and substitutable items. The decoupled graph neural network contains a two-step knowledge integration scheme. Chen *et al.* (2020d) design attribute-aware collaborative filtering to perform substitute recommendation by addressing issues from both personalization and interpretability perspectives.

6.2.2 Embedding-based methods

Embedding-based methods are another type of candidate retrieval methods. As shown in Figure 6.3, this kind of method first maps users and items into a common embedding space, and then uses approximate the K-Nearest Neighbor (KNN) algorithm (Andoni and Indyk, 2006) to search for candidate items with the smallest embedding distance to

the target user. The key of these methods lies in learning high-quality embeddings for each user and item. In what follows, we introduce techniques for learning embeddings.

Matrix factorization. Matrix factorization (MF) is a classic embedding-based method. The basic assumption behind MF is that the user-item interaction matrix has a low-rank structure. MF delineates each user and item as an embedding vector and then predicts the preference between each user-item pair as the inner product of their embedding vectors. Let $p_u \in \mathcal{R}^k$, $q_i \in \mathcal{R}^k$ denote the embedding vector of user u and item i , respectively. MF makes a prediction for the user-item pair (u, i) as follows:

$$\hat{r}_{ui} = p_u^\top q_i. \quad (6.2)$$

MF can be optimized by minimizing the deviation between the predictions and the user-item interactions. Formally, we have the following objective function:

$$\min_{P, Q} \sum_{(u, i)} l(r_{ui}, \hat{r}_{ui}) + \lambda L_{reg}(P, Q), \quad (6.3)$$

where $r_{u,i}$ denotes a user-item interaction. This can be explicit feedback (e.g., user ratings), which directly reflects the user preference, or implicit feedback (e.g., purchases and clicks), which indicates whether the user interacts with the item. $l(\cdot, \cdot)$ denotes the selected error function between the prediction and the ground truth label. It can be selected from mean squared error (MSE) loss (Koren *et al.*, 2009), binary cross-entropy loss (He *et al.*, 2017b), hinge loss (Wu *et al.*, 2016b), and Poisson likelihood (Gopalan *et al.*, 2015). $L_{reg}(\cdot, \cdot)$ denotes a regularizer for the embeddings to avoid over-fitting. Here we collect p_u (and q_i) for each user (and item) as a matrix P (and Q).

Compared with heuristic-based methods, MF is a more generic method that can adaptively learn user preferences from their history behaviors and require no manually crafted heuristic design. The use of MF has brought a revolution, pushing research attention from previous heuristic-based methods towards embedding-based methods.

Information-enhanced embedding models. Despite the prevalence of MF, it is still insufficient to yield accurate embeddings. The reason

is that MF directly projects user/item IDs to an embedding space, making MF reliant on the behavioral signal from the objective function. Hence, MF models do not perform well for inactive users or items with very few interactions. To deal with this problem, several methods have been proposed to enrich the representation with supplementary information. We can divide them into three groups: (i) neighborhood-enriched embedding methods (ii) feature-enriched embedding methods and (iii) graph-enriched embedding methods. We discuss each of these groups in detail as follows:

(i) Neighborhood-enriched embedding. Beyond the user ID, Bell and Koren (2007), Koren (2008), and Kabbur *et al.* (2013) enrich a user’s representation with their rating history. This kind of method generates user-item preference scores as follows:

$$\hat{r}_{ui} = q_i^\top \left(p_u + |\mathcal{N}(u)|^{-\alpha} \sum_{j \in \mathcal{N}(u)} y_j \right), \quad (6.4)$$

where $\mathcal{N}(u)$ denotes a set of items with which the user u has interacted. Each item $j \in \mathcal{N}(u)$ is mapped into a common embedding space to get a vector representation q_j . Neighborhood-enriched embedding methods can be considered as a combination of MF-based methods and neighborhood embedding methods. MF-based and neighborhood methods make predictions from different perspectives, which results in different strengths and weaknesses. Neighborhood methods struggle to detect all associations captured by interactions, whereas MF-based methods are poor at detecting associations among sparse neighborhoods. Hence, endowing MF with neighborhood methods fosters its merits and circumvents its weaknesses.

(ii) Feature-enriched embedding. Another way to improve MF is to use rich user and item features, e.g., a user’s age, gender, education, revenue, product tags, category, and price. These features are valuable in enriching the representations of users and items, which further boosts the recommendation performance. Figure 6.4 summarizes the architecture of feature-enriched candidate retrieval methods as a two-tower structure (Yi *et al.*, 2019; Fan *et al.*, 2019c; Xu *et al.*, 2016). The left tower is the user tower that translates a user’s features into the user’s

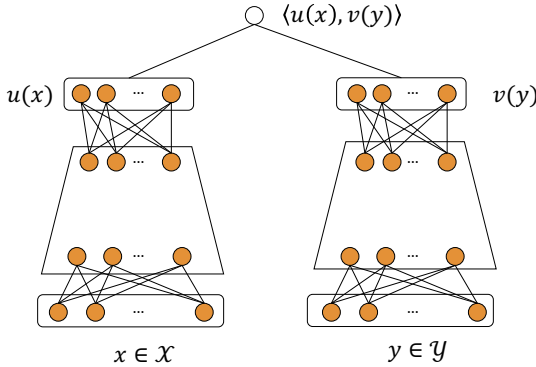


Figure 6.4: Illustration of the two-tower architecture for learning users' and items' representations. Image source: Xu *et al.* (2016).

embedding representation; the right tower is the item tower generating the item representation. Let x_u denote the features of user u , while x_i denotes the features of item i . The function of the two towers can be depicted as follows:

$$p_u = f_U(x_u), q_i = f_I(x_i), \quad (6.5)$$

where $f_U(\cdot)$ and $f_I(\cdot)$ denote the translation function with regard to the implementation of the two towers. Linear models and neural networks (e.g., MLP (Huang *et al.*, 2013), LSTM (Song *et al.*, 2016), CNN (Shen *et al.*, 2014) and auto-encoders (Wu *et al.*, 2016b; Liang *et al.*, 2018a)) can be used as the translation function. Given user and item embeddings, the two-tower model makes a prediction for each user-item pair. Cosine similarity and inner product are usually adopted to measure the embedding distance, so we have:

$$\begin{aligned} \text{Cosine: } \hat{r}_{ui} &= \frac{q_i^\top p_u}{\|p_u\| \cdot \|q_i\|}, \\ \text{Inner product: } \hat{r}_{ui} &= q_i^\top p_u. \end{aligned} \quad (6.6)$$

(iii) Graph-enriched embedding. A drawback of the aforementioned methods is that they regard each user or item as an “island” and fail to explicitly encode their relations into representations. These relations, e.g., item-item occurrences or user-item interactions, are important in

revealing correlations between users and items, which provide valuable signals for recommendation. To address this problem, several studies have established specific graphs between users and items, and then conduct graph representation learning to generate user and item embeddings. One representative method is EGES (Wang *et al.*, 2018b). EGES constructs an item-item graph based on user behavior sequences, where two items are connected if they consecutively occur in one sequence. EGES then applies DeepWalk (Perozzi *et al.*, 2014) on the item graph to generate item representations. Similarly, other related methods perform graph convolutional networks to enrich the representation learning of both users and items (Wang *et al.*, 2019c; Ying *et al.*, 2018; He *et al.*, 2020). A knowledge graph with informative relations between items has been exploited to learn better embeddings (Wang *et al.*, 2019b; Wang *et al.*, 2020d).

LightGCN (He *et al.*, 2020) is a light graph neural network (GNN) model for recommendation, in which only the item and user embedding need to be learned, whereas non-linear operations are not considered. The basic concept of LightGCN is to learn the user or item representation by aggregating the information from multi-order neighbors in the user-item interaction graph. Assuming that the user and item embedding are \mathbf{e}_u^0 and \mathbf{e}_i^0 for user u and item i respectively, LightGCN takes the following aggregation to get the representations of different layers:

$$\mathbf{e}_u^{k+1} = \sum_{i \in \mathcal{N}_u} \mathbf{e}_i^k, \mathbf{e}_i^{k+1} = \sum_{u \in \mathcal{N}_i} \mathbf{e}_u^k, \quad (6.7)$$

where k represents the k -th layer, \mathcal{N}_u represents the neighbors of user u , and \mathcal{N}_i represents the neighbors of item i . Then, the final representation of users and items are computed as follows:

$$\mathbf{e}_u = \sum_{k=0}^K \alpha_k \mathbf{e}_u^k, \mathbf{e}_i = \sum_{k=0}^K \alpha_k \mathbf{e}_i^k, \quad (6.8)$$

where α_k is a hyper-parameter to control the contributions of different layers. Eventually, it takes the dot product of the two representations as the prediction score.

The quality of data affects the upper bound of the performance of the learning-based models. However, user behavior data in e-commerce

recommendation is usually very sparse. To address this problem, data augmentation is a common strategy to increase the diversity of data. SGL (Wu *et al.*, 2021b) designs three types of methods to augment the training data for graph-based methods. The augmented data is used for an additional unsupervised task which maximizes the agreement of positive pairs.

6.2.3 Session-based recommendation

Modeling sequential dynamics is important for candidate retrieval in e-commerce recommendation. Sequential (or session-based) recommendation takes behavior sequences as the input, and then predicts the user's next click or purchase behavior. Sequential methods can be grouped into two classes: Markov chain-based and neural network-based sequential recommendation methods.

Early studies on sequential recommendation methods often use Markov chains by assuming that the user's next action only depends on the previous one. As a representative method, the factorizing personalized Markov chains model (FPMC) extends MF by modeling the effects of sequential-consecutive actions (Rendle *et al.*, 2010). Since the previous action is a critical factor affecting the user's next decision, FPMC achieves a significant gain over MF-based models. Following FPMC, He and McAuley (2016a) use a higher-order Markov chain to capture the sequential dependence among non-consecutive behavior. It is hard to use Markov chain-based methods for capturing complicated and long-term dependencies in sequential data, and this limits their performance in e-commerce recommendation.

More recently, deep neural networks have been used in sequential recommendation due to their powerful expressive ability on capturing behavior dependences (discussed in Section 4). Generally speaking, neural networks for sequential recommendation can be divided into three types: (i) RNN-based methods that model sequential dependence with RNN (or improved versions such as LSTM and GRU) to capture both long-term and short-term dependencies (Quadrana *et al.*, 2017; Jannach and Ludewig, 2017b; Hidasi *et al.*, 2016); (ii) CNN-based methods that concatenate the embedding of the previous item in a sequence as to

a matrix (Tang and Wang, 2018); and (iii) attention-based methods that introduce attention mechanisms into sequential recommendation by considering various types of user behavior with different influences (Kang and McAuley, 2018; Sun *et al.*, 2019a).

6.2.4 Next-basket recommendation

The next basket recommendation (NBR) task uses information from previous sessions. It is defined as recommending a group of items to a user based on their shopping history, where the history is a time-ordered sequence of baskets that they have purchased in the past (Li *et al.*, 2023b). Each basket is a set of items with no particular order. This formulation fits the grocery shopping setting well, where a user's purchase history occurs naturally in the form of such baskets. Variations of the NBR task where the order of items in a basket may be relevant, can, e.g., be found in music (playlist recommendation), in travel (recommending holiday packages), and in research and education (recommending reading lists). Two main characteristics of the grocery shopping scenario make the NBR task in this domain distinct from other retail domains: (i) users shop for grocery items repeatedly and on a regular basis, and (ii) grocery items have a short life time and are repurchased frequently by the same user (Liu *et al.*, 2019a).

In the grocery shopping domain, it has been found to be useful to distinguish between *repeat items*, i.e., items that a user has consumed before, and *explore items*, i.e., items that a user has not consumed before (Ariannezhad *et al.*, 2022; Li *et al.*, 2023b). In particular, for repeat items, the set of candidate items that needs to be considered for an individual user is usually in the low hundreds (Ariannezhad *et al.*, 2021) as opposed to the full item catalog that needs to be considered for explore items, whose size may exceed 50,000 items in grocery shopping (Ariannezhad *et al.*, 2020; Sprangers *et al.*, 2023). This fact makes the task of retrieving repeat items to be included in the next basket to be recommended to a user considerably easier than the task of retrieving explore items. Frequency-based, nearest neighbor-based, and deep learning-based methods have all been used for the NBR task, and for the candidate retrieval phase in particular (Li

et al., 2023b). As a rule of thumb, being biased towards the easier repetition task is an important strategy that helps to boost the overall NBR performance. Deep learning-based methods do not effectively exploit the repetition behavior. Indeed, they achieve a relatively good exploration performance, but they are not able to outperform simple frequency-based baselines in several cases. Some recent state-of-the-art NBR methods are skewed towards the repetition task and outperform frequency-based baselines. However, the improvements they achieve are limited, especially considering the complexity and computational costs, e.g., for the training process (Yu *et al.*, 2020) and for hyper-parameter search (Faggioli *et al.*, 2020; Hu *et al.*, 2020a).

A number of variations of the NBR task have recently been considered, each with their own challenges for retrieving candidate items. The *within-basket recommendation* task uses information from previous sessions as well as information from an incomplete basket to which additional items could potentially be added (Ariannezhad *et al.*, 2023). Another variation concerns an *item-centered* scenario (as opposed to the familiar user-centered scenario), where the input is an item and the task is to identify users who might be interested in consuming the item (Li *et al.*, 2023a).

6.3 Candidate Ranking Models

Given generic feature vectors as input, work on candidate ranking strategies has mainly focused on modeling interactions between features. According to the types of interaction function they adopt, existing methods can be divided into linear, polynomial, and neural network models.

6.3.1 Linear models

Early studies on candidate ranking usually apply linear models, such as *logistic regression* (LR) (Kleinbaum *et al.*, 2002; Hosmer Jr *et al.*, 2013) and naive Bayesian methods (Hastie *et al.*, 2009). In contrast to other complicated models, linear models are straightforward, efficient, and explainable. Although linear models may not perform as well as deep

neural networks, they indeed lay the foundation for recent advances in e-commerce recommendation (Peng *et al.*, 2002; Kiseleva *et al.*, 2016; Bernardi *et al.*, 2015). In e-commerce recommendation, logistic regression (LR) is one of the most popular methods that formulate the task as a classification task to rank through predicting the probability of an item to be interacted. LR first collects features of users (e.g., age and gender) and items (e.g., price and categories) into a number of feature vectors, and then applies a linear combination function to map the feature vector into the final predicted score. Similarly, Kiseleva *et al.* (2016) employ a naive Bayesian ranking strategy in e-commerce recommendation by considering contextual user profiling.

6.3.2 Polynomial models

The performance of linear models is limited because of high space complexity and the inability of high-level feature modeling. To address these two problems, factorization machines (FM) have been proposed (Rendle, 2010). Factorization machines factorize parameters $w_{i,j}$ into an inner product of two latent vectors, i.e., $w_{i,j} \equiv \langle v_i, v_j \rangle$, where v_i denotes a latent vector of the i -th feature. With different types of knowledge, the feature interactions across multiple fields should have different weights in recommendation. To tackle this challenge, field-aware factorization machines (FFM) (Juan *et al.*, 2016) have been proposed to capture field-aware weights and distribute a single latent vector to multiple fields.

A drawback of FM and FFM is that they only capture second-order feature interactions but neglect higher order interactions, which are widely observed in e-commerce scenarios. As described in Section 4.1.1, He *et al.* (2014) proposed a hybrid model by combining GBDT and logistic regression for click-behavior modeling. The hybrid model is able to use boosted decision trees (i.e., GBDT) to conduct feature interactions into logistic regression for e-commerce recommendation. In this hybrid model, GBDT adaptively conducts feature selection and higher-order feature interactions.

6.3.3 Neural network models

Deep neural networks have been used in e-commerce recommendation because of their powerful expression ability of capturing complicated feature interactions. Up to now, research on neural network models can be categorized into the following three research directions: (i) The first direction aims at developing feature interaction modules based on neural networks, e.g., adding more neural network layers or combining the superiority of different neural networks. (ii) The second direction aims at enhancing the expression by using deep neural networks using FMs. (iii) The third direction aims at using the attention mechanism in capturing diverge and dynamic contributions of feature interactions. Below, we will detail the recent advances along these research directions.

Neural feature interactions. Shan *et al.* (2016) propose the *deep crossing* model, which can be considered as the first end-to-end deep learning framework for recommender systems. Deep crossing enjoys the merits of deep learning in coping with various features and capturing complex feature interactions. It consists of the four components: an embedding layer, a stacking layer, multiple residual units layer, and a scoring layer. The goal of the embedding layer is to transform per individual sparse features into dense vectors in latent space via neural networks. The stacking layer concatenates different embedding features from the embedding layer and generates a new vector for all features. The scoring layer servers as an output layer with logistic regression to generate the final predicted score.

Collaborative filtering can be reconsidered from the perspective of deep learning. Traditional collaborative filtering methods employ the inner product of a user's latent vector and an item's latent vector for rating prediction. Neural collaborative filtering (NCF) (He *et al.*, 2017b) has been proposed to replace the inner product operation with a neural network.

By learning frequent co-occurrences of features, deep neural networks may have poor memorization, i.e., these models easily over-generalize and recommend less relevant items when user-item interactions are sparse. To address this problem, Cheng *et al.* (2016) introduce the Wide&Deep model for recommendation. The detailed model architecture

of Wide&Deep has already been presented in Section 4.1.1. Wide&Deep maintains a balance between memorization and generalization: its wide component can effectively memorize sparse feature interactions, while the deep neural networks can generalize the previously unseen feature interactions through low-dimensional embeddings.

Follow-up studies improve either the wide component or the deep component in the Wide&Deep model. The Deep&Cross Network (DCN) (Wang *et al.*, 2017c) replaces the wide component with a well-designed cross network. DCN applies an explicit feature crossing mechanism with multiple cross layers. Later studies seek to apply automating machine learning (AutoML) to model the selection of feature interactions (Su *et al.*, 2021; Meng *et al.*, 2021b; Zhao *et al.*, 2021b). SIF (Yao *et al.*, 2020) uses one-shot architecture search methods to search proper interaction functions (e.g., inner product, plus/minor, max/min pooling, outer product, and concatenation) for collaborative filtering models. AutoFIS (Liu *et al.*, 2020b) continuously searches effective feature interactions by incorporating architecture parameters to identify important feature interactions. AutoGroup (Liu *et al.*, 2020a) groups useful features into sets using AutoML and then generates interactions from each set.

Endowing factorization machines with neural networks. Many models have been proposed to integrate factorization machines (FMs) with deep neural networks to make full use of their advantages in feature combination. Neural factorization machines (NFMs; He and Chua, 2017) enhance FMs by modeling higher-order and non-linear feature interactions. NFM introduces a bi-interaction pooling operation in neural network modeling, and presents a new neural network perspective for FMs. Given the embedding set of all features \mathcal{V}_x , the bi-interaction layer in NFM converts \mathcal{V}_x to one vector:

$$f_{BI}(\mathcal{V}_x) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n (x_i v_i) \odot (x_j v_j), \quad (6.9)$$

where \odot denotes an element-wise product of two vectors. The output of bi-interaction pooling is a d -dimensional vector that encodes the second-order interactions between features in the embedding space. By stacking non-linear layers above the bi-interaction layer, NFM can effectively

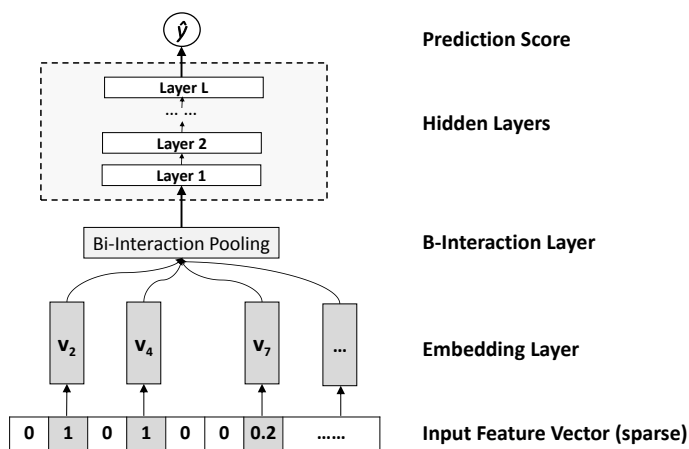


Figure 6.5: Overview of neural factorization machines. Image source: He and Chua (2017).

model higher-order and non-linear feature interactions. In contrast to traditional deep learning methods that simply concatenate or average embedding vectors in the low level, the use of bi-interaction pooling encodes more informative feature interactions. Figure 6.5 illustrates the architecture of NFM.

As demonstrated in Section 4.1, DeepFM (Guo *et al.*, 2017) aims to learn both low and high-order feature interactions, and consists of two components: the FM component and the deep component. Compared with Wide&Deep, DeepFM replaces its wide component with FM to remedy its shortcoming in automatic feature combination. Another difference is that DeepFM shares the feature embedding between the FM and deep component. Besides NFM and DeepFM, many other neural networks have been proposed based on FMs: FNN (Zhang *et al.*, 2016a) directly stacks FMs with neural networks; PNN (Qu *et al.*, 2016) models both bit-wise interactions and vector-wise feature interactions; and xDeepFM (Lian *et al.*, 2018) extends deepFM with explicit high-order feature interactions.

Attention mechanisms. Attention mechanisms have been applied in recommender systems and achieved great success (Xiao *et al.*, 2017; Li *et al.*, 2017a; Zhang *et al.*, 2019b). AFM (Xiao *et al.*, 2017) is an early

attempt to introduce an attention mechanism to recommendation. It can be regarded as an extension of NFM. The sum pooling operation in NFM treats all pairwise feature interactions equally, which may produce suboptimal results. To address this problem, AFM uses the attention mechanism on feature interactions by performing a weighted sum on the interacted vectors. The output of the attention-based pooling layer is projected into the prediction score. In session-based or sequential recommendation scenarios, Li *et al.* (2017a) propose an encoder-decoder model, neural attentive recommendation machine (NARM), to emphasize a user's main purpose in the current session. The authors adopt a hybrid encoder structure with a global component for modeling long-term purposes and a local component for modeling short-term purposes. Based on the combined session representation, a bi-linear matching scheme is then applied to compute the recommendation scores for each candidate item.

Following NARM, Ren *et al.* (2019c) put forward the RepeatNet model to deal with the phenomenon of repeat consumption behavior. The authors incorporate a repeat-explore mechanism into neural networks, which can select items from a user's history and suggests them at the appropriate moment. In standard embedding paradigms, user features are compressed into fixed-length representation vectors. However, fixed-length vectors limit capturing the diverse interests of a user from historical behavior. (Zhou *et al.*, 2018a) introduce the deep interest network (DIN) to tackle this challenge by designing a local attention unit. The local attention unit in DIN adaptively learns the representation of user interests from historical behavior by taking account of the relevance of historical behavior. Cheng and Xue (2021) unify CTR prediction models using a discrete choice model based on the self-attention mechanism. The authors regard feature interaction as the individual's comprehensive measurement of the influence of different factors on the decision-making process.

6.3.4 Retraining strategies

E-commerce recommender systems rely on knowledge gleaned from historical interactions. As the number of collected interactions grows,

recommendation models must be regularly retrained to reflect users' dynamic preferences. There are two intuitive heuristic retraining methods:

- **Full retraining** simply merges the old data and new data to perform a full model training. The method is designed to capture both short-term and long-term features of the recommender system based on all the data it has accumulated.
- **Fine-tune retraining** refers to using the parameters of the old model that were optimized by the old data to initialize the new model and train it with the new data. It reduces the time and storage overhead of retraining, making life-long updating feasible.

Numerous deep learning-based retraining methods have been proposed. Zhang *et al.* (2020g) propose a sequential meta-learning model (SML), including a meta-learning retraining framework for vanilla matrix factorization models. SML model captures the trend of changes between two distinct retraining phases at adjacent times. Recently, graph convolutional neural network (GCN) has become the cutting-edge technique for recommendation (He *et al.*, 2020; Ying *et al.*, 2018). However, GCN-based recommender models encounter challenges regarding model retraining as GCN-based models take more time to converge. CIGC (Ding *et al.*, 2021) has been proposed with two novel operators: incremental graph convolution and colliding effect distillation. The incremental graph convolution estimates the output of fully retraining the graph convolution using only new data; whereas the colliding effect distillation uses causal inference to retrain the representations of users (or items) that have no new data.

6.4 Re-ranking Strategies

The two-stage recommendation framework faces several problems in e-commerce recommendation: (i) The point-wise objective functions (e.g., log-loss) at the ranking stage often become sub-optimal because of the neglect of mutual influences between items. (ii) Users with different preferences may exhibit different behavior patterns. (iii) Recommendation result diversification has not yet been well addressed during the ranking stage. Therefore, before exposing recommended items to users, most e-commerce recommender systems refine the results through

an additional re-ranking stage. Generally, the goal of re-ranking is to enhance the recommendation results through additional criteria or constraints (Chen *et al.*, 2017c; Zehlike *et al.*, 2017; Abdollahpouri, 2019). Re-ranking methods can be categorized into heuristic strategies and list-wise objective functions. The former type of method is based on determinantal point processes (Wilhelm *et al.*, 2018; Chen *et al.*, 2017c) and maximal marginal relevance (Carbonell and Goldstein, 1998). These methods have been widely applied in e-commerce to avoid the items with the same category being presented consecutively for greater diversity. While for list-wise objective functions, DLCM (Ai *et al.*, 2018) and PRM (Pei *et al.*, 2019) have been proposed with specific list-wise optimization objectives.

6.5 Emerging Directions

We discuss five emerging research directions in e-commerce recommendation: (i) structured recommendations, (ii) conversational recommendation, (iii) reasoning recommendations and explanations, (iv) biases and debiasing in recommendation, and (v) unifying recommendation and search.

6.5.1 Structured recommendations

The task of structured recommendations is to predict the next structured item sets instead of the next item. we discuss three categories of structured recommendations: slate recommendation, playlist recommendation, and next-basket recommendation.

In applications like music or bundle recommendations, the objective is to provide users with a “slate” – a combination of items – to maximize their engagement with the recommended content. This task raises critical questions, including the consideration of metrics such as diversity and the computational challenges posed by the combinatorial nature of slates. Reinforcement learning (RL) is extensively applied in slate recommendation (Ie *et al.*, 2019; Sunehag *et al.*, 2015; Deffayet *et al.*, 2023; Tomasi *et al.*, 2023). However, due to the combinatorial complexity of actions, RL typically necessitates simplifying assump-

tions, such as the user selecting the optimal item (Ie *et al.*, 2019). An alternative approach involves integrating a separate user preference model to optimize slate assembly and subsequently training the RL model (Tomasi *et al.*, 2023). Swaminathan *et al.* (2017) investigate off-policy evaluation and optimization via inverse propensity scores for slate interactions. Mehrotra *et al.* (2019) construct a hierarchical model to assess user satisfaction in slate recommendation systems.

Music playlist recommendation can be considered as a special case of music recommendation, focusing on delivering a curated list of songs to users. The order and characteristics of music tracks significantly influence the playlist's overall quality. An earlier study employs time-series-based machine learning to address the challenge of recommending music playlists (Choi *et al.*, 2016; Irene *et al.*, 2019; Monti *et al.*, 2018; Vall *et al.*, 2018; Kim *et al.*, 2018; Yang *et al.*, 2019). Choi *et al.* (2016) use a recurrent neural network (RNN) for music playlist generation, emphasizing track transition qualities. Monti *et al.* (2018) implement an ensemble of RNNs, using pre-trained embeddings for album and title representation. Irene *et al.* (2019) predict user preferences by analyzing manually created playlists, employing both RNN and convolutional neural network (CNN) models. Later studies have employed reinforcement learning (RL)-based methods to capture users' long-term preferences (Hu *et al.*, 2017; Shih and Chi, 2018; Sakurai *et al.*, 2020; Sakurai *et al.*, 2021; Sakurai *et al.*, 2022; Tomasi *et al.*, 2023; Liebman *et al.*, 2015). Liebman *et al.* (2015) use a novel reinforcement-learning-based music recommendation system that generates playlists by considering both song preferences and transitions. Hu *et al.* (2017) enhance playlist generation performance by integrating user feedback into the recommendation reward function. Shih and Chi (2018) incorporate novelty and popularity indices into the reward function, resulting in playlists with a mix of new and well-known tracks. Sakurai *et al.* (2022) use informative knowledge graphs to enhance reinforcement learning optimization, and allowing users to customize flexible reward functions to discover new music genres. Tomasi *et al.* (2023) present a reinforcement learning framework optimizing directly for user satisfaction via the use of a simulated playlist-generation environment.

Next-basket recommendation (NBR) is a task focused on predicting a user's next shopping basket based on their past purchase history, aiming to enhance user experience and satisfaction. There are mainly three families of NBR methods. First are conventional NBR methods, such as those employing pattern mining (Guidotti *et al.*, 2017), KNN models (Hu *et al.*, 2020a; Faggioli *et al.*, 2020), and Markov chain models (Rendle *et al.*, 2010). Hu *et al.* (2020a) and Faggioli *et al.* (2020) model temporal patterns across frequency data and integrate this with neighbor information or user-wise collaborative filtering. Rendle *et al.* (2010) use matrix factorization and Markov chains to model users' general interest and basket transition relations. Second are latent representation methods, which use representation learning techniques to capture implicit patterns in data. For instance, Wang *et al.* (2015) apply aggregation operations to learn a hierarchical representation of user's last basket to predict the next basket. Third are deep learning-based method. Recurrent neural networks (RNNs) have been extensively applied in next-basket recommendation, demonstrating their efficacy in learning long-term trends by modeling the whole basket sequence. For instance, Yu *et al.* (2016) use max/avg pooling to encode baskets and Hu and He (2019) adapt an attention mechanism and integrate frequency information to improve performance. Some methods (Le *et al.*, 2019; Wang *et al.*, 2020c) use item relations to enhance representation. Yu *et al.* (2020) employ graph neural networks (GNNs) to model item-item relations between baskets and a self-attention mechanism to discern temporal dynamics. Some methods (Bai *et al.*, 2018; Chen *et al.*, 2021e; Leng *et al.*, 2020; Sun *et al.*, 2020a; Wang *et al.*, 2019a) use auxiliary information, including product categories, amounts, prices, and explicit timestamps.

6.5.2 Conversational recommendation

The task of a conversational recommender system (CRS) is to provide recommendations to users through conversational interactions. CRSs are increasingly attracting attention (see, e.g., Zhao *et al.*, 2013; Christakopoulou *et al.*, 2016; Yu *et al.*, 2019; Zou *et al.*, 2020b; Mangili *et al.*, 2020; Sun and Zhang, 2018; Lei *et al.*, 2020a; Lei *et al.*, 2020b; Zhou

et al., 2020b; Liu *et al.*, 2020g; Zhou *et al.*, 2020c; Zhang *et al.*, 2020f; Li *et al.*, 2021d). According to Gao *et al.* (2021a), the task of CRS is formally defined as follows:

A recommendation system that can elicit the dynamic preferences of users and take actions based on their current needs through real-time multi-turn interactions.

Building on advances in interactive recommendation (Christakopoulou *et al.*, 2018; Wang *et al.*, 2017a; Liu *et al.*, 2020e), early studies on CRSs formulate the task as a specific application of task-oriented multi-turn dialogue systems (TDS) (Le *et al.*, 2018; Dhingra *et al.*, 2017; Wen *et al.*, 2017b; Zhang *et al.*, 2019e). Studies into CRSs follow one of two main types of strategy: attribute-aware and topic-guided.

Attribute-aware CRSs aim to answer three main research questions: “whether to ask or recommend,” “which attributes to ask” or “which items to recommend.” Early work on attribute-aware CRSs obtains user preferences based on asking about items directly (Zhao *et al.*, 2013; Wang *et al.*, 2018c; Christakopoulou *et al.*, 2016; Zou *et al.*, 2020b; Vendrov *et al.*, 2020), or asking attributes through a heuristic method (Christakopoulou *et al.*, 2018; Zhang *et al.*, 2018c; Luo *et al.*, 2020). There are two main kinds of attribute-aware CRSs solutions. One kind asks a fixed number of questions and makes a recommendation at the last turn (Lei *et al.*, 2020a; Lei *et al.*, 2020b); whereas the other predicts a specific turn to recommend items. Reinforcement learning strategies have successfully been applied to attribute-aware CRSs. Liu *et al.* (2020g) and Li *et al.* (2021d) focus on cold-start users in conversational recommendation and extend bandit-based algorithms to balance the trade-off between exploration and exploitation. Zou *et al.* (2022b) propose TSCR, a transformer-based sequential conversational recommendation method that captures the sequential dependencies in dialogues to enhance recommendation accuracy. Deng *et al.* (2023) propose a novel unified multi-goal conversational recommender system, named UniMIND, which unifies these goals into a single sequence-to-sequence (Seq2Seq) paradigm and employs prompt-based learning strategies to facilitate multi-task learning.

Topic-guided CRSs interact with users through natural language conversations with fluent responses and precise recommendations (Li *et al.*, 2018c; Zhou *et al.*, 2020c; Chen *et al.*, 2019c; Liu *et al.*, 2020g; Zhou *et al.*, 2020b; Ma *et al.*, 2021; Zhou *et al.*, 2022a). Unlike attribute-aware CRSs, topic-guided CRSs focus on making recommendations using free text, which creates considerable flexibility to influence how a dialogue continues. External knowledge has been applied in topic-guided CRSs (Ma *et al.*, 2021; Zhou *et al.*, 2020b; Chen *et al.*, 2019c). Chen *et al.* (2019c) integrate a recommendation system and a dialogue system via an end-to-end framework to bridge the gap between the two systems. Li *et al.* (2018c) use an auto-encoder for recommendation and a hierarchical RNN for response generation. Zhou *et al.* (2020c) propose a topic-guided CRSs method that incorporates topic threads to enforce transitions actively toward a final recommendation. More recently, external knowledge graphs have been shown to be effective in improving the performance of topic-guided conversational recommendation systems. Chen *et al.* (2019c) apply knowledge graphs to enhance the semantics of contextual items for recommendation. Zhou *et al.* (2020b) incorporate both word-oriented and entity-oriented knowledge graphs. Ma *et al.* (2021) perform tree-structured reasoning on a knowledge graph for recommendation. Zhang *et al.* (2022) focus on user reformulation behaviors to improve the robustness of conversational agents. Ren *et al.* (2022) explore user preferences in conversational recommendation and propose a variational reasoning mechanism to jointly track both short-term and long-term user behaviors. Zhang *et al.* (2023) present the first attempt to explicitly address the problem of dynamic reasoning over incomplete knowledge graphs.

However, no study is capable of fusing recommendation and response generation in an end-to-end manner, which limits the potential for mutual reinforcement between these two tasks. Additionally, a lack of interpretability in current conversational recommendation system (CRS) models further hinders their ability to fully align with user needs. Models are typically trained on conversational recommendation datasets, but the assumption that the standard items and responses in these benchmark datasets are optimal leads to a tendency for CRSs to replicate the logic of the recommenders found in the data, rather than

truly addressing the evolving needs of the users. This misalignment remains a significant challenge in advancing more user-centric and adaptable conversational recommendation systems.

Although CRSs have many merits, their evaluation is still a thorny issue. Recent studies have evaluated CRSs either through offline evaluation or human evaluation (Lamel *et al.*, 2000; Li *et al.*, 2015). Offline evaluation evaluates a dialogue system based on test sets, whereas human evaluation reflects the overall performance of the agent through in-field experiments (Black *et al.*, 2011; Gilotte *et al.*, 2018) or crowdsourcing (Zhou *et al.*, 2020c; Li *et al.*, 2018c). However, offline evaluation is often limited to single turn assessments, while human evaluation is intrusive, time-intensive, and is not scalable (Zhao *et al.*, 2019b; Siro *et al.*, 2022). As an alternative, user simulators that mimic user behavior are able to provide broad insights to generate human-like conversations for assessing CRSs (Afzali *et al.*, 2023).

6.5.3 Explainable e-commerce recommendation

Although recommendation models can generate relevant items for users in many e-commerce applications, it is often ambiguous to understand why an item is recommended to a user. Hence it is necessary to develop explainable recommendation strategies to generate not only high-quality recommendations but also intuitive explanations. Recent years have witnessed a growth in the number of publications on explainable recommendation. Zhang *et al.* (2014) generated textual sentences as recommendation explanation to help users understand each recommendation result. Chen *et al.* (2018c) propose visually explainable recommendations where particular regions of a recommended image are highlighted as the visual explanations for users. Sharma and Cosley (2013) and Quijano-Sanchez *et al.* (2017) generate a list of social friends who also like the recommended product as social explanations for target user, whereas Gao *et al.* (2019a) generate the recommendation described by a set of topics.

Several researchers have started to generate explanations for deep recommendation models. For example, several studies use knowledge graphs for interpretation. They construct multi-hop paths from users

to items along the knowledge graph, which indicates a specific explainable user-item relation (Hu *et al.*, 2018a; Wang *et al.*, 2019d; Xian *et al.*, 2019). Besides, Chen *et al.* (2021c) propose a neural collaborative reasoning system integrating the power of representation learning and logical reasoning. However, research on explainable deep recommendation models is relatively new and deserves to be further explored in e-commerce.

6.5.4 Biases and debiasing in recommendations

Many recommendation solutions about fitting user behavior may deteriorate owing to biases in behavior inherent in e-commerce recommendation (He *et al.*, 2020; Sun *et al.*, 2019a). In e-commerce scenarios, user behavior is observational rather than experimental, which is often affected by many factors, e.g., self-selection of the user (selection bias) (Marlin *et al.*, 2007), systematic exposure mechanisms (exposure bias) (Ovaisi *et al.*, 2020), public opinions (conformity bias) (Krishnan *et al.*, 2014; Liang *et al.*, 2016) and the display position (position bias) (Joachims *et al.*, 2007). These biases make the data deviate from reflecting true preferences of users in recommender systems. Efforts to debias recommendation can be divided into three major categories: (i) data imputation, which assigns pseudo-labels to missing data to reduce variance (Steck, 2013), (ii) inverse propensity scoring (IPS), which reweighs the collected data for an expectation-unbiased learning (Sun *et al.*, 2019b; Wang *et al.*, 2016b), and (iii) generative modeling, which assumes the generation process of data and reduces biases (Liang *et al.*, 2016). Most approaches lack the universal capacity to account for mixed or even unknown biases. To bridge the gap, Chen *et al.* (2021d) propose a universal debiasing framework that not only account for multiple biases and their combinations, but also frees human efforts to identify biases. Huang *et al.* (2022) introduce DANCER, a debiasing method that accounts for dynamic selection bias and user preferences, demonstrating its improved rating prediction performance over static bias methods. Heuss *et al.* (2023) explore the use of uncertainty estimates in ranking scores to reduce societal biases in retrieved documents while minimizing utility loss. They propose an uncertainty-aware, post hoc bias mitigation method

that outperforms baselines in terms of utility-fairness trade-offs, controllability, and computational costs, without requiring additional training. Although recent years have seen a surge in research efforts devoted to recommendation biases, biases are still an important problem in e-commerce recommender systems. Sophisticated meta models to capture complex patterns and exploration of dynamic biases in recommendation should provide helpful insights.

6.5.5 Unifying recommendation and search

Search and recommendation in e-commerce have similar characteristics, except for the different representation of “contexts” – search aims at retrieving relevant items for matching a query while recommendation aims at finding items for matching a user’s preferences. However, researchers usually conduct separate studies on them and use different techniques and training data for the two tasks. Thus, building a unified model for search and recommendation has the potential to improve both tasks as more comprehensive user behavior data can be used. One practical way to unify the two tasks is as part of the aforementioned conversational recommendation scenario, and the other is personalized search, which we detail next.

Early search engines, like Google and AltaVista, retrieved personalized results based on keywords. Personalized search has become far more complex with the goal to “understand exactly what you mean and give you exactly what you want.” Concretely, a personalized search engine not only focuses on retrieving items that satisfy the user’s current information needs, which is usually related to the query topic, but also considers user personality and aims at retrieving items that meet user preference. To achieve both goals, it is critical to model interactions between users, items and queries. Ai *et al.* (2017) use a hierarchical embedding model to linearly combine the item-query matching scores with item-user preference scores; Guo *et al.* (2019c) explore long and short term user preference learning model for personalized search; Yao *et al.* (2021) integrate user behavior in search and recommendation into a heterogeneous behavior sequence and use a joint model to handle both tasks based on this unified sequence; Si *et al.* (2023) use users’

search interests for recommendations; they separately learn similar and dissimilar representations from search and recommendation behaviors using transformer encoders. Liu *et al.* (2020c) construct a specific user-item-query graph and conduct node representation learning on the graph. Zhao *et al.* (2022a) propose a method that jointly predicts user clicks for both search and recommendation scenarios by constructing a unified graph to share user and item representations uniformly. Such graph embedding techniques open the potential to integrate both node information and topological structure information, which can capture high-order user-item-query interactions.

6.5.6 Large language models in recommendation

Large language models (LLMs) have exhibited strong capabilities in understanding and processing text. Their application to recommendation systems is actively being explored. The main benefit of using LLMs in recommendation systems is their ability to produce high-quality representations of text features and make use of the wide range of knowledge they hold (Liu *et al.*, 2023a). LLM-based models can capture context more accurately, allowing them to better understand user questions, product descriptions, and other textual information. Studies that apply LLMs to recommendation systems can be divided into two categories: discriminative strategies and generative strategies.

For studies into discriminative strategies, to improve the quality of vector representations for queries and products, and fully use the external knowledge stored in LLMs, a common approach is to fine-tune the original models, adapting them to recommendation tasks in order to obtain high-quality representations. Qiu *et al.* (2021) propose a novel U-BERT approach that utilizes a pre-training and fine-tuning framework to learn user representations. By using content-rich domains, U-BERT compensates for users' features in domains where behavior data is insufficient, improving recommendation performance. Similarly, Wu *et al.* (2021a) use unlabeled user behavior data and incorporate two self-supervised tasks: masked behavior prediction and behavior sequence matching for user model training.

Compared to discriminative models, generative models have better natural language generation capabilities. Therefore, most generative models typically translate recommendation tasks into natural language tasks, allowing the model to directly output recommendation results through fine-tuning or in-context learning. Sun *et al.* (2023b) introduce a sliding window prompt strategy for ranking candidates. This strategy ranks items within a window at each step, sliding the window from back to front multiple times to generate the final ranking results. This approach helps improve ranking performance by iteratively refining the candidate list. Kang *et al.* (2023) investigate the ability of LLMs to predict user ratings based on past behavior, comparing them with traditional collaborative filtering methods.

The application of LLMs to e-commerce recommender systems is still in its early days. Many challenges remain, in terms of evaluation, effectiveness, efficiency, and transparency.

7

E-commerce QA and Conversations

Section 3 provides insights on how natural language processing technologies have been widely applied in e-commerce platform interfaces to help consumers better communicate with those platforms. This section zooms in on question answering (QA) services and dialogue systems on e-commerce platforms. We divide this section into three parts: e-commerce question answering, e-commerce dialogue systems, and emerging directions. We first detail characteristics and approaches to e-commerce question answering (Section 7.1). Then, we demonstrate recent studies on dialogue systems applicable in e-commerce customer services (Section 7.2). Lastly, we describe emerging directions in e-commerce question answering and dialogue systems (Section 7.3).

7.1 Question Answering in E-commerce

In this section, we describe related work on e-commerce question answering. We divide this section into three parts: we first introduce studies on question answering in Section 7.1.1, then in Section 7.1.2 we formulate characteristics of product-aware question answering; finally, we detail approaches to e-commerce question answering in Section 7.1.3 and 7.1.4.

7.1.1 Introduction to question answering

QA systems (Simmons, 1965) are meant to facilitate users' access to information. For many web-based applications QA services provide a proper answer to a given question from the user (Heilman and Smith, 2010; Li and Roth, 2002). Question answering research has received much attention in the past decades, including approaches to question classification, answer selection, answer generation, and answer summarization (Li and Roth, 2002; Heilman and Smith, 2010; Liu *et al.*, 2016b; Geigle and Zhai, 2016; Song *et al.*, 2017). QA systems have various classifications. QA systems can be divided into open-domain and domain-specific QA systems (Chen and Yih, 2020). Open-domain QA focuses on answering questions relying on knowledge and ontologies (Ferrucci *et al.*, 2010), whereas domain-specific QA focuses on providing proper answers in a specific scenario, e.g., customer service, hotel booking, etc. QA systems can also be divided into retrieval-based and generation-based QA systems according to how they generate answers (Yang *et al.*, 2015a). The former searches and extracts potential answers via search engines, whereas the latter applies generation-based methods to give proper answers to the questions. And finally, according to the answers, QA systems can be divided into factoid QA and non-factoid QA (Song *et al.*, 2017). Factoid QA systems return a concise answer to the given question, whereas non-factoid QA systems provide more subjective answers to the given questions.

Early work on QA distinguishes between four categories of QA system: list-structured database systems, graphical database systems, text-based systems, and logical inference systems (Simmons, 1965). All these systems have a limited scope with their rule-based strategies. Search engines remain integral components of QA systems. With the development of information retrieval, research “re-discovered” QA systems in the late 1990s (Jurafsky, 2000). TREC has launched dedicated QA tracks in 1999, with the purpose of advancing research into QA systems (Srihari and Li, 1999; Voorhees, Tice, *et al.*, 1999). A typical retrieval-based TREC QA system has three main components: question processing, passage retrieval, and answer processing (Jurafsky, 2000). For each step, sub-tasks must be considered, e.g., query formulation

or answer type detection. Based on this framework, several approaches have been proposed to address research tasks in each component (Brill *et al.*, 2001; Li and Roth, 2002). Early studies on QA focus on factoid QA systems that generate concise answers (Srihari and Li, 1999; Jurafsky, 2000; Brill *et al.*, 2001). Retrieval-based methods effectively answer these concise and simple questions (Jurafsky, 2000; Ahn *et al.*, 2004). However, complicated questions are found difficult to be addressed by pure retrieval-based methods (Lin, 2006). Therefore, integrating natural language understanding and knowledge-based reasoning techniques is essential for retrieval-based QA strategies in answering complicated questions.

In TREC-QA 2004, questions are grouped into topics, which motivates research on fact identification from reference knowledge resources, e.g., Wikipedia (Ahn *et al.*, 2004). Wikipedia can be considered a generic collection of articles with real-world facts for open-domain QA systems. With the development of knowledge bases, innovations have occurred in the context of QA from knowledge bases with the creation of resources like web questions and short questions (Berant *et al.*, 2013; Bordes *et al.*, 2015). However, inherent limitations such as incompleteness and fixed schemas have persisted in traditional knowledge-based QA systems. Thus, in the 2000s QA work increasingly on systems that are able to generate answers from raw text explored, especially using Wikipedia (Ahn *et al.*, 2004; Buscaldi and Rosso, 2006; Ferrucci *et al.*, 2010; Ryu *et al.*, 2014). As far as we know, Ahn *et al.* (2004) are the first to combine Wikipedia as a text resource with other resources in QA. Similarly, Ryu *et al.* (2014) perform QA using a Wikipedia-based knowledge model by combining articles with other answer-matching components. Ferrucci *et al.* (2010) and Baudiš (2015) integrate web-based and Wikipedia-based articles as knowledge resources into highly developed full-pipeline QA platforms.

In more recent years, QA models increasingly apply deep neural networks to understand questions and generate answers. Yin *et al.* (2016) present an end-to-end neural network model, neural generative question answering (GENQA), that can generate answers to simple factoid questions. Subsequently, a bi-directional attention flow mechanism has been proposed to obtain query-aware passage representations (Seo *et al.*,

2017). Chen *et al.* (2017a) develop a system for question answering from Wikipedia, DrQA, that is composed of a two-stage retrieval-reader QA framework. DrQA includes a document retriever module based on bigram hashing and TF-IDF matching. It also contains a document reader module where a multi-layer recurrent neural network is trained to detect answer spans in those few returned documents.

Following Chen *et al.* (2017a), most open-domain QA systems apply a two-stage retrieval-reader framework in their QA mechanisms (Wang *et al.*, 2018d; Sun *et al.*, 2018b; Lin *et al.*, 2018; Pang *et al.*, 2019; Lee *et al.*, 2019; Guu *et al.*, 2020; Karpukhin *et al.*, 2020; Izacard and Grave, 2021; Mao *et al.*, 2021; Sachan *et al.*, 2021; Singh *et al.*, 2021; Yu *et al.*, 2022a; Kedia *et al.*, 2022; Ju *et al.*, 2022; Wang *et al.*, 2023b). These studies employ a determinate retrieval function and treat each passage independently in the retrieval stage. Wang *et al.* (2018d) propose a reinforcement learning-based ranking strategy in the retrieval stage. Sun *et al.* (2018b) offer a graph convolution-based neural network by operating over heterogeneous graphs of knowledge base facts and text sentences. In contrast with previous knowledge-based open-domain QA systems, the authors propose heterogeneous update rules that handle knowledge base nodes differently from the text nodes. Lin *et al.* (2018) design a coarse-to-fine denoising model to extract correct answers from multiple paragraphs in the noisy data. Their model employs a paragraph selector to filter out those noisy paragraphs and keep informative paragraphs. Similarly, Pang *et al.* (2019) describe a three-level probabilistic formulation model for open-domain QA. Word-level matching strategies are usually applied in the retrieval stage to match keywords represented in high-dimensional and sparse vectors. Dense passage retrieval has successfully been applied to open-domain QA to improve the matching performance as it is complementary to sparse representations in the retrieval stage.

Karpukhin *et al.* (2020) train a dense embedding model using only pairs of questions and passages. Izacard and Grave (2021) detail an effective two-step dense passage retrieval method; the authors retrieve supporting passages using either sparse or dense embeddings and then employ a sequence-to-sequence model to generate the answer. Zhu *et al.* (2021c) use a partially observed Markov decision process (POMDP) to

re-formulate the QA problem using a reinforcement learning method to optimize the interactions between different components. Yu *et al.* (2022a) use a knowledge graph to establish relational dependencies among retrieved passages and employ a graph neural network to re-rank retrieved passages for each query. More recent work has proposed to improve reader performance and thereby improve QA performance. Kedia *et al.* (2022) introduce a method for fusing information across multiple passages within a transformer encoder using global representation tokens. Ju *et al.* (2022) design a knowledge graph enhanced passage reader that fuses graph and contextual representations into the hidden states of the reader model. Wang *et al.* (2023b) enhance the fusion-in-decoder (FiD) framework by incorporating a process to distinguish between relevant and spurious passages, thereby improving the model's reasoning and performance in open-domain QA.

A model that matches the question with a passage using gated attention-based recurrent networks has been shown to be effective on QA benchmark datasets (Wang *et al.*, 2017d). QANet combines local convolution with global self-attention for reading comprehension, which improved the reading comprehension performances (Yu *et al.*, 2018a). More recent studies have shown that pre-trained language models effectively understand questions and answers in QA systems (Guu *et al.*, 2020; Mao *et al.*, 2021; Sachan *et al.*, 2021; Singh *et al.*, 2021). Mao *et al.* (2021) augment a query in open-domain QA using text generation of a pre-trained language model. Sachan *et al.* (2021) propose a QA method with an unsupervised pre-training of the retriever with a supervised fine-tune procedure.

Many benchmark QA datasets have been proposed. Several QA benchmark datasets, such as SQuAD (Rajpurkar *et al.*, 2016), TriviaQA (Joshi *et al.*, 2017), and SearchQA (Dunn *et al.*, 2017), only evaluate the reasoning ability within a single paragraph, whereas the other relevant documents or paragraphs are neglected. These datasets employ knowledge bases for multi-hop reasoning, and are therefore constrained by the schema of knowledge bases. Yang *et al.* (2018b) introduce an open-domain QA benchmark dataset, HotpotQA, which requires reasoning over multiple documents without constraining itself to a knowledge base. To understand how the questions and answers are

distributed in open-domain QA, Lewis *et al.* (2021) perform a large-scale analysis on open-domain QA benchmark datasets, and provide annotated subsets of test sets indicating whether test-time questions are duplicates of training time questions.

7.1.2 Characteristics of e-commerce question answering

E-commerce QA services focus on answering product-aware questions asked by e-commerce users. Early studies on e-commerce QA focus on providing answers automatically from reviews by heuristic methods (Li *et al.*, 2009; Moghaddam and Ester, 2011; Yu *et al.*, 2012). With the development of both QA techniques and e-commerce services, e-commerce QA has received increasing attention in recent years (McAuley and Yang, 2016; Yu *et al.*, 2018b; Yu and Lam, 2018; Fan *et al.*, 2019b; Zhang *et al.*, 2020c; Gao *et al.*, 2019b; Gao *et al.*, 2021b; Feng *et al.*, 2021; Deng *et al.*, 2022).

Distinct characteristics of e-commerce QA, as opposed to open-domain QA, are: (i) Domain-specific aspects are the first e-commerce QA characteristic. E-commerce QA systems rely on exploiting domain-specific information from product descriptions. Different products make different product-related aspects relevant or popular (McAuley and Yang, 2016). These product-aware aspects can help distinguish products and answer questions. (ii) There is a large number of consumer reviews, which can be used as a data source to help people form opinions and decisions (Liu *et al.*, 2016b; McAuley and Yang, 2016). With the growth of those opinionated reviews, e-commerce users rely on advice from reviews before making purchase decisions. Reviews have been used as supporting data and candidate answers to supervise QA prediction models (Yu and Lam, 2018). (iii) There is a variety of answer sources. Most e-commerce QA services focus on extracting answers from reviews (McAuley and Yang, 2016), and many e-commerce sites provide question answer pairs as knowledge bases for QA.

Text generation approaches have been studied to generate answers to given questions and reviews. Question reranking and answer reranking also have been studied (Yu *et al.*, 2018b; Zhang *et al.*, 2020c). E-commerce QA research can be divided into two directions: extrac-

tive product-aware QA and generative product-aware QA. The former focuses on extracting sentences or passages from reviews to answer questions, whereas the latter applies textual generation approaches to generate answers. We detail each type of e-commerce QA study in Section 7.1.3 and 7.1.4, respectively.

7.1.3 Extractive product-aware QA

Most e-commerce QA systems extract relevant sentences or fragments from the input text to answer the question given by the consumer. Early studies automatically extracted answers from reviews by heuristic unsupervised methods (Li *et al.*, 2009; Moghaddam and Ester, 2011). Follow-up work mainly focuses on the matching between questions and reviews or candidate answers (McAuley and Yang, 2016; Yu and Lam, 2018). Yu *et al.* (2012) proposed a framework for opinionated QA, which organizes reviews into a hierarchical structure and retrieves review sentences as the answer. The authors then use such a hierarchical structure to help retrieve questions and relevant review fragments. A joint optimization approach is proposed by simultaneously considering review salience, coherence, and diversity to rank fragments. Liu *et al.* (2016b) find a concise set of questions addressed by a given review and cover its main points to help the user quickly comprehend the reviews. The authors propose a two-stage framework, where a probabilistic retrieval model is used to retrieve candidate questions and a matching procedure between answers and questions is used to bridge the vocabulary gap between reviews and questions.

Some products, such as clothes and paintings, may not have proper names. Different strategies have been considered to replace the external knowledge of e-commerce to address this problem. McAuley and Yang (2016) propose an answer prediction model by incorporating an aspect analytic model to learn latent aspect-specific review representation for predicting the answer. Wan and McAuley (2016) address ambiguity, subjectivity, and diversity problems in consumer reviews. By using multiple answers in a supervised framework, the authors provide more accurate answers to objective and subjective questions. The authors also release a large-scale e-commerce QA dataset consisting of 135 thousand

products from Amazon, 808 thousand questions, 3 million answers, and 11 million reviews. Carmel *et al.* (2018) focus on subjective questions from Amazon customers, which can relate to various intent types such as product usage, recommendations, and opinions. The authors apply automatic QA methods, enhanced with community QA approaches to retrieve the most relevant answer found in reviews and QAs to address this problem. Yu and Lam (2018) propose an answer prediction model by incorporating an aspect analytic model to learn latent aspect-specific review representation for predicting the answer.

The authors establish the advantage of generating aspect-specific representations for new questions, which they use to develop a predictive answer model to capture intricate relationships among question texts and review texts. The proposed model uses reviews as a knowledge source to predict the answer by classifying answers into two types, binary (i.e. “yes” or “no”) and open-ended responses. As the amount of labeled data is limited in customer reviews, Das *et al.* (2019) propose an adversarial review-based approach to answer subjective and specific product-aware questions in a weakly supervised setting. Reading comprehension has been found to be useful to help extract relevant answers from e-commerce reviews (Fan *et al.*, 2019b; Xu *et al.*, 2019; Zhang *et al.*, 2020e; Chen *et al.*, 2019a). Using the raw text of product-aware questions and customer reviews, Fan *et al.* (2019b) introduce an end-to-end neural network model to synthesize multiple review representations. Chen *et al.* (2019a) design a multi-task attentive model, namely QAR-Net, to identify plausible answers from product reviews for user questions. QAR-Net can use generated question answer pairs to help question-review matching.

Pre-trained language models help understand the content of questions and reviews. Xu *et al.* (2019) apply a BERT-based fine-tuning approach to extract answers from reviews. Mittal *et al.* (2021) use a pre-trained language model to learn a relevance function by jointly learning unified syntactic and semantic representations of questions and reviews. A QA dataset for review comprehension with subjectivity labels for questions and answers has also been exploited (Bjerva *et al.*, 2020). Besides user reviews, another type of information, namely product details provided by the manufacturer, has been considered an auxiliary information source for addressing product-related questions (Zhang

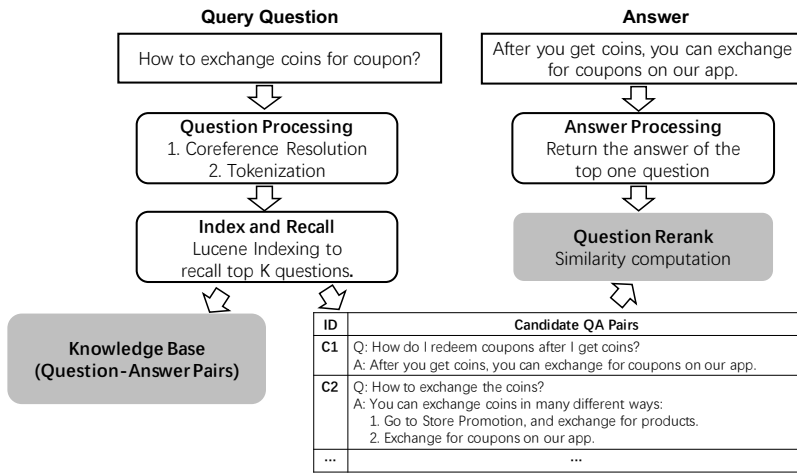


Figure 7.1: An overview of the retrieval-based QA system in Alibaba. Image source: Yu *et al.* (2018b).

et al., 2020e). In order to alleviate the unavailability of labeled data, Jain *et al.* (2023) introduce a distant supervision based model to prepare training data without manual effort.

Review-based QA approaches extract answers from customer reviews, which can partially address users' questions. However, there are many products with few or no reviews available. By collecting question answer pairs from real users, many e-commerce platforms develop retrieval-based QA systems for automatically answering frequently asked questions (FAQs) in the e-commerce industry (Yu *et al.*, 2018b; Song *et al.*, 2020b; Song, 2021; Zhang *et al.*, 2020c). In Figure 7.1, we see the retrieval-based QA framework applied by Alibaba (Yu *et al.*, 2018b). Given a collection of question answer pairs (i.e., the knowledge base in Figure 7.1), a key component is the question rerank module, which reranks candidate questions in a question answering knowledge base to find the best match given a question from a user. Based on such a framework, Yu *et al.* (2018b) formulate e-commerce QA as a paraphrase identification problem, where the target is to identify semantic relations of the given sentence pairs. The authors describe a transfer learning QA strategy to adapt the shared knowledge learned from a resource-rich

source domain to a resource-poor target domain. Amazon has presented a large review-based QA dataset, namely AmazonQA, based on their real-world community QA platform (Gupta *et al.*, 2019). AmazonQA uses consumer reviews as the data resource and extracts snippets to answer questions. Song *et al.* (2020b) improve the matching performance in retrieval-based e-commerce QA by introducing a multi-layer triple convolutional neural network model. Also, a sub-graph searching mechanism is shown to improve the efficiency of retrieval-based e-commerce QA (Song, 2021). Zhang *et al.* (2020c) focus on answer selection in retrieval-based e-commerce QA. Using graph neural networks, the authors jointly model multiple semantic relations, including semantic relevance between the question and answers, textual similarity among answers, and textual entailment between answers and reviews. Rozen *et al.* (2021) detail an answer prediction approach that uses similar questions about other products. The authors calculate contextual product similarity to determine whether two products are similar in the context of a specific question. Two large-scale datasets, including a question-to-question similarity dataset from Amazon and a corpus of question answer pairs from Amazon, have been released with the publication.

7.1.4 Generative product-aware QA

Many e-commerce portals have provided question answering services that assist users in posing product-aware questions to other consumers who have purchased the same product before. Users must read the product's reviews to find the answer themselves. Given product attributes and reviews, following a cascading procedure, an answer is manually generated: (i) a user skims reviews and finds relevant sentences; (ii) they extract functional semantic units; and (iii) and the user jointly combines these semantic units with attributes and writes an appropriate answer. With a rapidly increasing number of reviews this process needs support (Gao *et al.*, 2019b). Several strategies have been proposed to automatically generate answers using the product's reviews to alleviate the burdens of customers (Gao *et al.*, 2019b; Chen *et al.*, 2019e; Deng *et al.*, 2020; Lu *et al.*, 2020; Feng *et al.*, 2021; Deng *et al.*, 2022). The

task on which these approaches focus is *generative product-aware QA* given reviews and product attributes.

In first attempts, Gao *et al.* (2019b) and Chen *et al.* (2019e) propose the task of *product-aware answer generation*, where a product-related question answering model is applied to incorporate customer reviews with product attributes. The authors formulate the research problem in generative e-commerce QA: for a product, there is a question $X^q = \{x_1^q, x_2^q, \dots, x_{T_q}^q\}$, T_r reviews $X^r = \{x_1^r, x_2^r, \dots, x_{T_r}^r\}$ and T_a key-value pairs of attributes $A = \{(a_1^k, a_1^v), (a_2^k, a_2^v), \dots, (a_{T_a}^k, a_{T_a}^v)\}$, where a_i^k is the name of i -th attribute and a_i^v is the attribute content. Each attribute, including key a_i^k and value a_i^v , is represented as a single word in the generation task. Given a question X^q , an answer generator reads the reviews X^r and attributes A , then generates an answer $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{T_y}\}$. The goal is to generate an answer \hat{Y} that is grammatically correct and consistent with product attributes and opinions in the reviews.

Figure 7.2 provides an overview of the product-aware answer generator, the PAAG model, proposed by Gao *et al.* (2019b). PAAG has four parts: (i) a *review reader* reads the review to extract relevant semantic parts; (ii) an *attribute encoder* encodes the attribute key-value pairs using a key-value memory network; (iii) a *facts decoder* generates the final answer according to the facts learned by the two modules introduced before; and (iv) a *consistency discriminator* distinguishes whether the generated answer matches the extracted facts, and we also use the result of the discriminator as another training signal. A generative e-commerce QA dataset extracted from JD.com is released with the publication. Similarly, Chen *et al.* (2019e) formulate a noise-tolerant solution based on convolutional neural networks to generate natural answers. Deng *et al.* (2020) exploited opinion information reflected in the reviews. The authors generated opinion-aware natural answers using multi-task learning to integrate opinion detection and answer generation simultaneously.

It is necessary to consider the text information from different reviews and attributes to answer specific questions in the wild. In Figure 7.3, Feng *et al.* (2021) provide examples to demonstrate the multi-type text

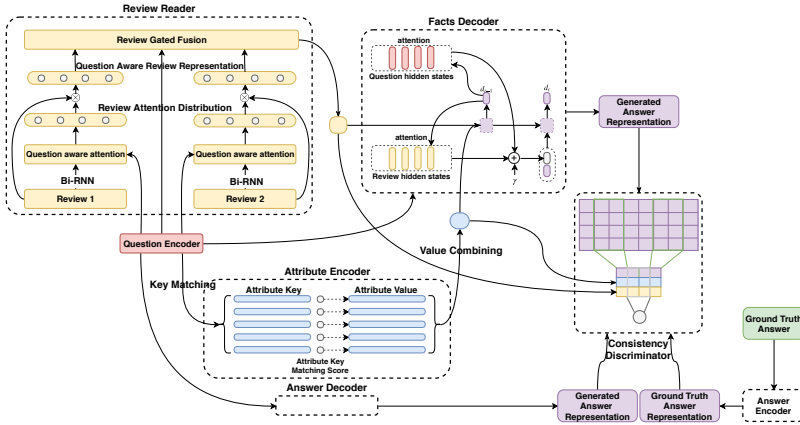


Figure 7.2: Overview of the product-aware answer generator model. Image source: Gao *et al.* (2019b).



Women's Knit Tunic
Tops Loose Long Sleeve Button Up V Neck Shirts

★★★★☆ > 21
\$15⁹⁹

Product Attributes

- ***Size:** Fits true to size.
- ***Type:** Tops/Tunics
- ***Style:** Make you beautiful, fashionable, sexy and elegant.
- ***Occasion:** Summer beach/casual/party/evening/wedding/holiday.
- ***Garment Care:** Hand wash, dry clean
- ***Collar:** V-Neck.

Questions



Q1: Dose The **design** of this tops looks baggy?

Q2: What are the **garment care** instructions for this top?

Reviews



R1: The clothes is so beautiful, it **designs** with **bat sleeve**, button down style, and v-neck.

R2: Bat sleeve looks **baggy**. So I send it back to get a small one.

R3: I **hand wash** it **in cold water** and the shape hold up well. But when I use hot water, it becomes shrink and fade.

Answers



A1: The design of this tops looks **baggy**.

A2: Hand wash **in cold water**, or dry clean.

Figure 7.3: Examples of the multi-type text relation for product-aware question answering. Image source: Feng *et al.* (2021).

relation for product-aware question answering. As an example, *Q1* asks “Does the design of this top look baggy?” *R1* and *R2* do not answer this question directly. But they provide a common entity “bat-like sleeve.” If we transfer the information provided by *R1* and *R2* to answer *Q1* indirectly, it is easy to generate the answer that “The design of this tops looks baggy.” By integrating, understanding, and reasoning over the information of reviews and product attributes we may generate more accurate and pleasing answers to complex questions. A major limitation of most generative QA approaches is that they analyze each review and the corresponding attribute of the product individually, i.e., they neglect the relationship between different reviews/attributes of the product. Feng *et al.* (2021) propose a review-attribute heterogeneous graph neural network, RAHGNN, for product-aware answer generation to sufficiently understand and reason about the related information and its inner logic in multiple types of text. Most generative product-aware QA methods neglect personalization as it is insufficient to provide the same “completely summarized” answer to all customers. As an exception, Deng *et al.* (2022) describe a personalized answer generation method, PAGE, to model multi-perspective user preferences in personalized product question answering.

7.2 Dialogue Systems in E-commerce

Dialogue systems have increasingly attracted attention in e-commerce. This section introduces studies on dialogue systems that can be applied to e-commerce platforms. Following previous work investigating this problem (Chen *et al.*, 2017b; Chen *et al.*, 2018b), we divide this section into three parts. We introduce recent studies on dialogue systems in Section 7.2.1, then detail task-oriented dialogue systems in e-commerce in Section 7.2.2, and discuss knowledge-grounded conversational agents in Section 7.2.3.

7.2.1 Introduction to dialogue systems

Dialogue systems are being considered in numerous applications, from e-commerce technical support to personal assistant tools (Song *et al.*,

2017; Chen *et al.*, 2017b; Chen *et al.*, 2018b; Sun *et al.*, 2016; Zhang *et al.*, 2018c; Lei *et al.*, 2018; Liu *et al.*, 2018b; Meng *et al.*, 2020b; Sun *et al.*, 2021a; Shen *et al.*, 2021; Zhao *et al.*, 2021a; Liu *et al.*, 2021a; Ren *et al.*, 2022; Li *et al.*, 2022c; Yu *et al.*, 2022c). The goal of creating an automatic human-computer conversational system as an assistant or chat companion is no longer an illusion now that two important factors have been seen progress. First, many conversation logs are now accessible, making it possible for machines to learn how to respond to input utterances. Second, deep generative neural network models, such as sequence-to-sequence and generative adversarial networks, are now able to capture complex patterns in large volumes of data (Chen *et al.*, 2017b). Based on these two factors, studies on dialogue systems focuses on methods to provide a natural and coherent response given an utterance from a user (Young *et al.*, 2013; Ritter *et al.*, 2011; Banchs and Li, 2013; Ameixa *et al.*, 2014).

Dialogue systems can be divided into chitchat systems, task-oriented dialogue systems, and knowledge-grounded conversations (Chen *et al.*, 2017b). Chitchat agents are applied widely in open-domain dialogue systems, where dialogue systems interact with humans to provide reasonable and natural responses for open-domain dialogues (Chen *et al.*, 2018b; Yan *et al.*, 2017). Chitchat messages usually represent user experiences and preferences, playing an essential role in many real-world applications. Yan *et al.* (2017) reveal that most utterances in the online shopping scenario are chitchat messages.

Task-oriented dialogue systems aim to complete a specific task, e.g., restaurant reservation, along with a response generation process. Figure 7.4 shows the four individual modules on which traditional task-oriented dialogue systems are based: natural language understanding, dialogue state tracking, policy learning, and natural language generation (Wen *et al.*, 2017a; Mrkšić *et al.*, 2015). Given an utterance from a user, the system generates a proper response to address the user's intention. In recent years, end-to-end task-oriented dialogue generation methods have been proposed to address the overall purpose more efficiently (Wen *et al.*, 2017b; Lei *et al.*, 2018; Wu *et al.*, 2019a).

Knowledge-grounded conversations focus on generating a response with the correct knowledge to address the user's utterance (Meng *et*

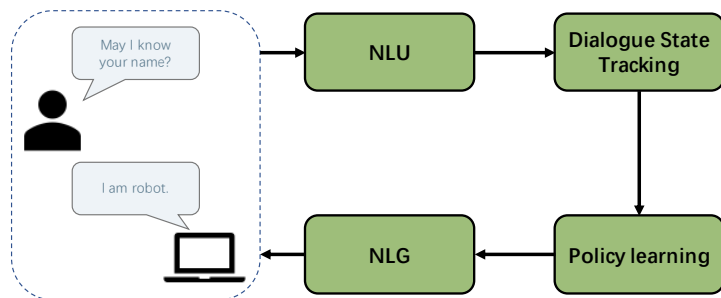


Figure 7.4: Traditional pipeline for task-oriented dialogue systems. Image source: (Chen *et al.*, 2017b).

al., 2020b). Work on knowledge-grounded conversations can be categorized into two groups. Methods in the first group use *structured knowledge* (given knowledge graphs) (Zhou *et al.*, 2018c; Liu *et al.*, 2018b; Tuan *et al.*, 2019; Wu *et al.*, 2019b; Moon *et al.*, 2019; Wu *et al.*, 2020c; Zhou *et al.*, 2020a; Wang *et al.*, 2020a; Wu *et al.*, 2020b; Xu *et al.*, 2020b; Xu *et al.*, 2020a; Jung *et al.*, 2020; Xu *et al.*, 2021b). Those in the second group use *unstructured knowledge*, such as *document-based unstructured knowledge* (given a whole document, e.g., a Wikipedia article) (Meng *et al.*, 2020a; Ma *et al.*, 2020a; Ma *et al.*, 2020b; Tian *et al.*, 2020; Ren *et al.*, 2020; Gopalakrishnan *et al.*, 2020; Li *et al.*, 2019d; Qin *et al.*, 2019; Moghe *et al.*, 2018; Zhou *et al.*, 2018d) or *piece-based unstructured knowledge* (given some separate pieces of knowledge, e.g., Foursquare tips) (Ghazvininejad *et al.*, 2018; Dinan *et al.*, 2019; Meng *et al.*, 2020b; Kim *et al.*, 2020a; Lian *et al.*, 2019; Zheng and Zhou, 2019; Zheng *et al.*, 2020a; Chen *et al.*, 2020e; Zheng *et al.*, 2020b; Zhao *et al.*, 2020b; Yu *et al.*, 2022c).

With respect to generating responses, dialogue systems can be divided into retrieval-based and generation-based dialogue systems. The former retrieves several response candidates from a prebuilt index and then selects an appropriate one as a response. In contrast, the latter directly synthesizes a reply via natural language generation techniques (Serban *et al.*, 2017a; Tao *et al.*, 2021b).

Retrieval-based dialogue systems retrieve several response candidates from a prebuilt index and then select an appropriate one as a

response. Social networks have accumulated a significant amount of conversational data among humans on the web, motivating researchers to investigate data-driven approaches to re-use human conversations and select a response for new input from candidates (Tao *et al.*, 2021b; Xu *et al.*, 2021c). Retrieval-based dialogue generation methods outperform their generation-based counterparts in response fluency and informativeness. They power a series of real-world applications, e.g., XiaoIce from Microsoft (Zhou *et al.*, 2020d). Learning to rank and matching approaches have been widely applied in the retrieval process (Yan *et al.*, 2016; Wu *et al.*, 2017b; Zhou *et al.*, 2018f; Yang *et al.*, 2018a; Yuan *et al.*, 2019; Su *et al.*, 2020a; Tao *et al.*, 2021b; Lin *et al.*, 2021b). A core task in retrieval-based dialogue systems is response selection. Studies into retrieval-based response selection can be divided into three types: representation-based, interaction-based, and pre-trained language model-based methods. Representation-based methods are composed of a representation-matching paradigm and consist of a representation layer and a matching layer (Yan *et al.*, 2016; Wu *et al.*, 2018b; Wang *et al.*, 2017c; Zhou *et al.*, 2018d; Zhou *et al.*, 2018f; Yan *et al.*, 2018; Xu *et al.*, 2021e). Interaction-based methods use context-response interactions to match potential responses (Tao *et al.*, 2021b). These methods follow a representation-matching-aggregation paradigm, formulating an interaction function to calculate the interaction between the two representation matrices of input utterances. The interaction function has two main types of definition: similarity-based and attention-based methods (Tao *et al.*, 2021b). Similarity-based methods calculate the similarity of each word pair between the context message and the response candidate (Wu *et al.*, 2017b; Zhang *et al.*, 2018d; Zhou *et al.*, 2018f). Attention-based methods, however, use an attention mechanism to match the context message and the candidate's response (Chen and Wang, 2019; Humeau *et al.*, 2019; Yuan *et al.*, 2019). In recent years, pre-trained language models have been applied in retrieval-based dialogue systems due to their strong ability for language representation and understanding. These approaches employ an attention-based strategy to unify the representation, interaction, and aggregation operations by feeding the concatenation of context utterances and the candidate responses into a pre-trained multi-layer self-attention network (Whang

et al., 2020; Gu *et al.*, 2020a; Xu *et al.*, 2021d; Han *et al.*, 2021; Tao *et al.*, 2021a; Feng *et al.*, 2022a; Li *et al.*, 2022a).

Generation-based dialogue systems generate natural-sounding replies automatically to exchange information (e.g., knowledge, sentiments, etc.) and complete a variety of specific tasks in a conversational interaction process (Young *et al.*, 2013; Shawar and Atwell, 2007). End-to-end textual generation models (Shang *et al.*, 2015; Vinyals and Le, 2015; Sordoni *et al.*, 2015; Li *et al.*, 2016a; Li *et al.*, 2016b; Serban *et al.*, 2016) have proved capable in multiple dialogue systems applications with promising performance. Most end-to-end neural generation models apply an encoder-decoder architecture based on a recurrent neural network, which directly maps an input context to the output response. Several approaches have been proposed to softly model language patterns, such as word alignment and repeating into sequence-to-sequence structure (Bahdanau *et al.*, 2015; Gu *et al.*, 2016; Serban *et al.*, 2017b; Cao and Clark, 2017). Gu *et al.* (2016) propose a copy mechanism to consider additional copying probabilities for contextual words in forum conversations. Serban *et al.* (2017b) decode coarse tokens before generating the complete response. Variational neural networks perform efficient inference and learning in models with directed probabilities on a large-scale dataset (Kingma and Welling, 2014; Kingma and Ba, 2015).

Cao and Clark (2017) tackle the boring output issue of deterministic dialogue models by introducing a latent variable model for a one-shot dialogue response. Serban *et al.* (2017a) propose HVRED to use the latent variable at the sub-sequence level in a hierarchical setting, whereas Chen *et al.* (2018b) add a hierarchical structure and a variational memory module into a neural encoder-decoder network. In e-commerce platforms, after-sale customer service is the main application scenario for dialogue systems. E-commerce dialogues need to address three targets: (i) task completion, such as changing the order address, providing the receipt, and returning the order; (ii) knowledge-based response selection and generation, such as checking the status of the delivery, answering the request about the refund period; and (iii) empathetic response generation, such as satisfying the consumers' request and replying to consumers' complaints. The JDDC datasets have been

q1	可以帮我改下订单的地址吗？(Could you help me change the address of the order?)
r1	同一市内可以联系配送员直接修改的哦。(You can contact the delivery staff directly if the two addresses are in the same city.)
q2	不在同一个城市，现在地址是上海，但是我明天要回安徽。(Not the same city. The current address is Shanghai, but I am going to Anhui tomorrow.)
r2	抱歉，地址在不同城市不能操作的，只能建议您重新下单哦。(Sorry, you cannot change the address to a different city. In this case, we suggest you place a new order.)
q3	那我取消订单的话退款多久到账呢？(How long does it take for the refund to arrive if I cancel the order?)
r3	微信零钱1个工作日内到账，储蓄卡1-7个工作日内到账，信用卡1-15个工作日内到账的哦！(For Wechat change, it arrives in 1 working day. For debit card, it arrives in 1-7 working days. And for credit card, it arrives in 1-15 working days.)
q4	为什么不能改地址，你们这也太不方便了。(Why can't I change my address? That is too inconvenient.)
r4	非常抱歉，我们物流还有待完善呢。(I'm sorry. Our logistics system needs to be improved.)
q5	这也太麻烦了，我还急着用呢。(That is too troublesome, I'm in a hurry.)
r5	非常抱歉！如果是我的话我也会很着急的，我们会改进的！(I'm so sorry! If I were you, I would feel the same. We will do our best to improve it!)
q6	行吧。(Fine.)
r6	感谢您的理解！还有什么能帮助到您的吗？(Thanks for your understanding! What else can I do for you?)

Figure 7.5: An example e-commerce dialogue in the JDDC 1.0 dataset. Image source: Chen *et al.* (2020c).

collected from JD.com, one of the largest e-commerce platforms in China. Figure 7.5 shows a typical example dialogue from JDDC 1.0 dataset. The blue text shows the target completion task, the red text indicates the knowledge-based response generation task, and the purple text shows the empathetic response generation task. Another characteristic of e-commerce dialogue systems is the phenomenon of multiple modalities. Text and images are often used in customer service dialogues (Zhao *et al.*, 2021a; Yuan *et al.*, 2022b).

Figure 7.6 shows another example from the JDDC 2.0 dataset to demonstrate the multi-modality characteristic of e-commerce. In this figure, three dialogue segments show images that are used to distinguish different product models for the same brand or used for identifying the location and cause of product failures. According to the above three targets in e-commerce dialogues, we detail e-commerce dialogue systems from three angles: task-oriented dialogue systems, knowledge-grounded dialogue systems, and empathetic dialogue systems in Sections 7.2.2, 7.2.3, and 7.2.4, respectively.

7.2.2 Task-oriented dialogue systems

Methods underlying task-oriented dialogue systems can be divided into pipeline methods and end-to-end methods. As shown in Figure 7.4, the pipeline of task-oriented dialogue systems can be divided into natural language understanding (NLU), dialogue state tracking (DST),

<p>Customer:  这款有货吗？ (Is this one available?)</p> <p>Agent: 现在白色是有货的，黑白的目前没有货。 (White is in stock now, black and white is currently out of stock.)</p>	<p>Customer: 我插上20分钟了，锅不热啊。(I have plugged it in for 20 minutes, the pot is not hot.)</p> <p>Agent: 调节大档了吗？灯亮吗？(Have you adjusted the big gear? Is the light on?)</p> <p>Customer: </p> <p>Agent: 插头没有插进去哦。(The plug is not inserted)</p>
<p>Customer:  没货啊？我下不了单的。(Out of stock? I can't place an order.)</p> <p>Agent: 您选白色的看看，黑白色的目前41.5码的是缺货。(Please choose the white one. The black and white one is out of stock at size 41.5.)</p> <p>Customer: 不好意思，我看错了啊。(Sorry, I made a mistake.)</p> <p>Agent: 😊👍</p>	<p>Customer: 货我打开了里面少了滤网。(I opened it and there is no filter inside.)</p> <p>Agent: 麻烦提供一下收到的产品图片啊。(Could you please provide us with pictures of the received products?)</p> <p>Customer: </p> <p>Agent: 只有部分机型有滤网的，这个没有的啊。(Only some models have a filter, this one does not.)</p> <p>Customer: 明白了，谢谢！(Thanks, I see.)</p>

Figure 7.6: Three segments of dialogue sampled from our JDDC 2.0 corpus. Image source: Zhao *et al.* (2021a).

and dialogue policy learning (DPL), and natural language generation (NLG) (Chen *et al.*, 2017b). For each stage, a number of pipeline-based approaches have been proposed, even though a lot of domain-specific handcrafting in traditional task-oriented dialogue systems is required, which, moreover, is difficult to adapt to new domains (Bordes and Weston, 2017). Recently, end-to-end neural network solutions have been widely applied to the task (Bordes and Weston, 2017; Zhao and Eskenazi, 2016; Wen *et al.*, 2017b; Jin *et al.*, 2018; Gao *et al.*, 2018; Madotto *et al.*, 2018).

In NLU, the dialogue system maps the input utterance into semantic slots. These semantic slots are pre-defined in different application scenarios. NLU includes two challenging problems: intent detection and slot filling. Intent detection methods for language understanding are performed to detect the user's intent (Deng *et al.*, 2012; Tur *et al.*, 2012). Deep neural networks have also been applied to detect the user's intent. Huang *et al.* (2013) use convolutional neural networks (CNN) to detect the user's intent; see also Shen *et al.* (2014). Slot filling is

usually set as a sequence labeling problem, where words are assigned with semantic labels (Chen *et al.*, 2017b). Deep belief networks have been successfully applied to address the filling problem (Deng *et al.*, 2012; Deoras and Sarikaya, 2013). Subsequently, recurrent neural networks have been shown to be effective in slot filling (Mesnil *et al.*, 2013; Sarikaya *et al.*, 2011; Yao *et al.*, 2013; Yao *et al.*, 2014). Unlike other NLU approaches, Liang *et al.* (2020) jointly formulate intent detection and slot filling as a sequence generation problem. Rastogi *et al.* (2020) provide a schema-guided paradigm for NLU. Recently, pre-trained language models have been applied to enhance NLU in task-oriented dialogue systems. Wu *et al.* (2020a) propose a self-supervised language model trained on multiple task-oriented dialogue system benchmarks. Zhang *et al.* (2021a) design a pre-trained model for few-shot NLU by fine-tuning BERT on a small set of labeled utterances. He *et al.* (2022a) explore tree-induced semi-supervised contrastive pre-training for NLU in task-oriented dialogue systems. The authors improve the NLU performance by injecting structural-semantic information to enhance the representation of dialogues. To explore more knowledge from long sequences in dialogue context, Zhong *et al.* (2022) formulate a window-based pre-trained model for NLU based on the sequence-to-sequence model architecture.

In a conversation, a *dialogue state* refers to a full and temporal representation of each participant's intention (Goddeau *et al.*, 1996). In task-oriented dialogue generation, dynamically tracking dialogue states is the key to generating coherent and context-sensitive responses. In DST, we use a dialogue state H_t to denote the representation of the dialogue till time t . Traditional approaches to DST focus on searching hand-crafted rules to select the most likely results (Wang and Lemon, 2013; Young *et al.*, 2010; Williams, 2012), where the dialogue state tracking is transferred to a slot filling problem. This type of slot filling problem has also been addressed by approaches using conditional random fields (Lee, 2013; Lee and Eskenazi, 2013; Ren *et al.*, 2013) and maximum entropy (Williams, 2013). However, relying on the most likely results from an NLU module (Perez and Liu, 2017), these rule-based systems hardly model uncertainty, which is prone to frequent errors (Williams, 2014; Perez and Liu, 2017).

Unlike rule-based state tracking methods, Young *et al.* (2010) propose a distributional dialogue state for statistical dialog systems and maintain a distribution over multiple hypotheses facing noisy conditions and ambiguity. Neural networks have been successfully applied to dialogue state tracking (Henderson *et al.*, 2013; Mrkšić *et al.*, 2015; Mrkšić *et al.*, 2017). In task-oriented dialogue systems, end-to-end neural networks are employed for tracking dialogue states via interacting with an external knowledge base (Wen *et al.*, 2017b; Eric *et al.*, 2017; Bordes and Weston, 2017; Williams *et al.*, 2017). Wen *et al.* (2017b) divide the training procedure into two phases: dialogue state tracker training and complete model training. Mrkšić *et al.* (2017) propose a dialogue state tracker based on word embedding similarities. Eric *et al.* (2017) implicitly model a dialogue state through an attention-based retrieval mechanism to reason over a key-value representation of the underlying knowledge base. Bordes and Weston (2017) track the dialogue context in a memory module and repeatedly search this context to select an adequate system response. Instead of employing symbolic knowledge queries, Dhingra *et al.* (2017) propose an induced “soft” posterior distribution over the knowledge base to search for matching entities. Williams *et al.* (2017) combine an RNN with domain-specific knowledge encoded as software and system action templates. The copying mechanism is shown to be effective as generative dialogue state tracking. Lei *et al.* (2018) propose an extendable framework to track dialogue states with a text span, including the constraints for a knowledge base query. The limited amount of labeled data is a severe challenge for DST. Jin *et al.* (2018) introduce a semi-supervised way to integrate a copy procedure with the dialogue state tracking.

While early studies on DST methods focused on a single-domain scenario, more recent studies have turned their attention to multi-domain DST with the release of a multi-domain DST benchmark dataset, MultiWoZ (Budzianowski *et al.*, 2018). Ramadan *et al.* (2018) jointly identify the domain and tracks the belief states corresponding to that domain to address the multi-domain DST problem. Zhou and Small (2019) formulate multi-domain DST as a question-answering task and used reading comprehension techniques to generate the answers. Similarly, Gao *et al.* (2019c) also formulate DST as a reading comprehension task

and propose an attention-based neural network to find the state answer as a span over tokens. The DSTC challenges have provided a series of popular experimentation frameworks and dialogue datasets collected through human-machine interactions for benchmarking (Henderson *et al.*, 2014a; Henderson *et al.*, 2014b; Williams *et al.*, 2014; Williams *et al.*, 2016).

In real-world scenarios, it is often not practical to enumerate all possible slot value pairs and perform scoring from a large, dynamically changing knowledge base (Xu and Hu, 2018). Wu *et al.* (2019a) propose a method to generate dialogue states from utterances using a copy mechanism, where tracking knowledge across domains is shared. To alleviate data sparsity in DST, Yin *et al.* (2020) propose a reinforced data augmentation framework to increase both the amount and diversity of the training data. Chen *et al.* (2020b) incorporate slot relations and model slot interactions in multi-domain dialogue state tracking to enhance the slot interrelation between disciplines. Feng *et al.* (2022b) extend this method by dynamically updating slot relations in the schema graph. Heck *et al.* (2020) maintain two memories in DST: one for system inform slots and one for the previously seen slots. Li *et al.* (2021e) combine a generation and extraction method with hierarchical ontology integration for DST. To tackle the understanding of ellipsis and reference expressions in open vocabulary-based methods, Ouyang *et al.* (2020) propose a copy-augmented encoder-decoder model by connecting the target slot and its source slot explicitly. Liao *et al.* (2021) formulate multi-domain DST as a recursive inference mechanism to improve the generation performance. Most DST models are trained offline, which requires a fixed dataset prepared in advance. Given a new domain in multi-domain DST, Campagna *et al.* (2020) propose a zero-shot transfer learning method to handle new domains without incurring the high cost of data acquisition. More recently, the granularity of dialogue history has been proposed to mitigate the sparseness in DST (Yang *et al.*, 2021b). Guo *et al.* (2022a) propose a multi-perspective dialogue collaborative selector module to dynamically select the granularity of dialogue history in DST.

Pre-trained language models have been shown to be effective in dialogue state tracking. Lin *et al.* (2021d) apply a pre-trained model for

DST to exploit external knowledge from reading comprehension data. Similarly, Zhong *et al.* (2022) verify that document summarization can provide helpful signals to improve DST. Liu *et al.* (2021d) introduce domain-lifelong learning into DST. The authors propose a knowledge preservation network that includes a multi-prototype enhanced retro-spection component and a multi-strategy knowledge distillation component. Lin *et al.* (2021e) successfully apply T5 to improve zero-shot cross-domain DST. Lee *et al.* (2021) introduce a solution for multi-domain DST by prompting knowledge from a large-scale pre-trained language model. Lin *et al.* (2021c) detail a hybrid method to integrate GPT-2 with graph attention networks to enhance the DST performance. To mitigate the problem of incorrect in DST, Wang *et al.* (2022a) design a BERT-based method by explicitly aligning each slot with its most relevant utterance.

Scalability, robustness, and efficiency in DST have also been addressed recently. Lei *et al.* (2018) formulate a two-stage copy-aware network demonstrating good scalability. Ren *et al.* (2019b) consider the DST task as a sequence generation problem and design a scalable hierarchical encoder-decoder neural network with constant inference time complexity. Kumar *et al.* (2020) extend this to improve the encoding of dialogue context and slot semantics for DST to robustly capture critical dependencies between slots and the conversation history. Kim *et al.* (2020b) focus on an open vocabulary-based setting and consider the dialogue state as a memory that can be selectively overwritten to improve the efficiency in multi-domain DST. Zhu *et al.* (2020b) introduce an efficient multi-domain dialogue state tracker by jointly encoding the previous dialogue state, the current turn dialogue, and the schema graph by internal and external attention mechanisms.

The policy learning module in task-oriented dialogue systems is meant to generate the following available system action given the state generation result (Cuayáhuitl *et al.*, 2015). Traditional rule-based methods are first applied in the policy learning procedure (Cuayáhuitl *et al.*, 2015). Supervised and reinforcement learning have also proven to be effective in policy learning (Su *et al.*, 2016; Yan *et al.*, 2017). Su *et al.* (2016) propose a two-stage framework for policy learning, i.e., a supervised learning stage and a reinforcement learning stage. Chen *et al.*

(2019b) propose a structured deep reinforcement learning approach for policy learning based on graph neural networks. The dialogue policy can be further trained in an end-to-end way with reinforcement learning to lead the system in making policies toward the final performance (Yan *et al.*, 2017; Chen *et al.*, 2019b). In an e-commerce scenario, the policy learning component needs to trigger the “recommendation” or a concrete service provided by the customer service (Sun *et al.*, 2016; Sun and Zhang, 2018; Zhao *et al.*, 2021a). Most task-oriented dialogue datasets, including WoZ and MultiWoZ, focus on language understanding and dialogue state tracking. However, selecting actions in real life requires obeying user requests and following practical policy limitations. Accordingly, Chen *et al.* (2021b) present an action-based conversations dataset consisting of 10042 conversations containing numerous actions with precise procedural requirements. He *et al.* (2022b) utilized semi-supervised pre-training to model explicit dialogue policy in task-oriented dialogue systems.

The natural language generation (NLG) component transfers a dialogue action into a natural language response (Chen *et al.*, 2017b). Neural network-based NLG approaches have been proposed for task-oriented dialogues (Wen *et al.*, 2015a; Wen *et al.*, 2015b; Tran and Nguyen, 2017; Zhou, Huang, *et al.*, 2016). Wen *et al.* (2015a) apply an RNN-based generator module and a CNN-based module to rerank candidate utterances. Wen *et al.* (2015b) use an additional control cell to gate the dialogue act to address the slot information omitting and duplicating problems in surface realization. Tran and Nguyen (2017) extend this approach by gating the input token vector of an LSTM with the dialogue act. A sequence-to-sequence approach is applied to produce natural language output and deep syntax dependency trees from input dialogue acts (Dušek and Jurcicek, 2016). Zhou, Huang, *et al.* (2016) propose an encoder-decoder LSTM-based method to jointly incorporate the request information, semantic slot values, and dialogue act type to generate correct answers. The copy mechanism (Vinyals *et al.*, 2015; Gu *et al.*, 2016) has been successfully applied to the task-oriented dialogue systems to enhance the performance of NLG (Eric and Manning, 2017; Lei *et al.*, 2018; Jin *et al.*, 2018). Aiming to augment dialogue datasets through paraphrasing, Gao *et al.* (2020) jointly optimize dia-

logue paraphrasing and dialogue response generation via a paraphrase augmented response generation approach. Pre-trained language models have shown supreme performance in text generation tasks (Li *et al.*, 2021b). In recent years, more and more studies have applied pre-trained language models to enhance the performance of NLG in task-oriented dialogue systems (Zhang *et al.*, 2020h; Peng *et al.*, 2020; Wu *et al.*, 2020a; Hosseini-Asl *et al.*, 2020; Zhong *et al.*, 2022).

End-to-end methods have been proposed for task-oriented dialogue systems. Wen *et al.* (2017b) propose an end-to-end trainable goal-oriented dialogue system with a new way of collecting dialogue data based on a pipeline framework toward end-to-end learning for DST and policy learning (Zhao and Eskenazi, 2016). The pipeline-aware method can also be implemented and trained end-to-end using the copy mechanism (Lei *et al.*, 2018; Jin *et al.*, 2018; Liao *et al.*, 2021). A copy-augmented sequence-to-sequence architecture has been proposed to provide better performance in task-oriented dialogues (Eric and Manning, 2017), while Eric *et al.* (2017) propose a key-value retrieval network for task-oriented dialogue response generation. Using the copy mechanism, Lei *et al.* (2018) formulate a theoretical framework that is end-to-end trainable using only one sequence-to-sequence model. Jin *et al.* (2018) propose a semi-supervised copy flow neural network to train the end-to-end dialogue generation model. Madotto *et al.* (2018) formulate a memory-to-sequence neural network that combines the multi-hop attention over memories with the idea of a pointer network. Xu and Hu (2018) also apply a pointer network to handle unknown slot values in the absence of a predefined ontology. Hosseini-Asl *et al.* (2020) enable modeling of the inherent dependencies between the sub-tasks of task-oriented dialogue by optimizing for all tasks in an end-to-end manner, recasting task-oriented dialogues as a simple and casual language modeling task. Liao *et al.* (2021) propose a recursive inference mechanism to resolve multi-domain DST in an end-to-end way. More recently, pre-trained language models have also been applied to end-to-end solutions for task-oriented dialogue systems (Wu *et al.*, 2020a; Lin *et al.*, 2021c; He *et al.*, 2022a).

In contrast to other types of dialogue systems, evaluation metrics in task-oriented dialogue systems need to consider specific metrics. *Entity*

match rate evaluates task completion (Wen *et al.*, 2017b); it determines if a system can generate all correct constraints to search the indicated entities of the user. This metric is either 0 or 1 for each dialogue. The original *success rate* metric measures if the system answered all the requested information (e.g., address, phone number) (Wen *et al.*, 2015a; Mrkšić *et al.*, 2015). However, this metric only evaluates recall. As a variant, *Success F1* evaluates task completion and is modified from the success rate by balancing both recall and precision (Lei *et al.*, 2018). Automatic user satisfaction has received much attention in task-oriented dialogues. User simulation is a promising approach to evaluate dialogue systems at scale in task-oriented dialogue scenarios (Zhang and Balog, 2020). Sun *et al.* (2021b) formulate the task of simulating user satisfaction for evaluating task-oriented dialogue systems to enhance the evaluation of dialogue systems. The authors also share a dataset about user satisfaction simulation. Kim *et al.* (2022b) propose the relative slot accuracy metric in DST evaluation, which is not affected by unseen slots in the current dialogue turn.

7.2.3 Knowledge-grounded dialogue systems

Although answering inquiries is essential for dialogue systems, especially for task-oriented dialogue systems, it is still far behind a natural knowledge-grounded dialogue system, which should be able to understand the facts involved in the current dialogue session (so-called fact matching) and diffuse them to other similar entities for knowledge-based dialogues (i.e., entity diffusion):

- *Fact matching*: In dialogue systems, matching utterances to exact facts is much harder than answering explicit factoid inquiries. Though some utterances, whose subjects and relations can be easily recognized, are fact-related inquiries, the subjects and relations are often elusive, leading to challenges when matching exact facts. Table 7.1 shows an example, with items 1 and 2 talking about the film “Titanic.” Unlike item 1, which is a typical question-answering conversation, item 2 is a knowledge-related chat without any explicit relation. It is difficult to define the exact fact match for item 2.

ID	Dialogue
1	<p>A: Who is the director of the <u>Titanic</u>? 泰坦尼克号的导演是谁?</p> <p>B: <u>James Cameron</u>. 詹姆斯卡梅隆。</p>
2	<p>A: <u>Titanic</u> is my favorite film! 泰坦尼克号是我最爱的电影!</p> <p>B: The love inside it is so touching. 里面的爱情太感人了。</p>
3	<p>A: Is there anything like the <u>Titanic</u>? 有什么像泰坦尼克号一样的电影吗?</p> <p>B: I think the love story in film <u>Waterloo Bridge</u> is beautiful, too. 我觉得魂断蓝桥中的爱情故事也很美。</p>
4	<p>A: Is there anything like the <u>Titanic</u>? 有什么像泰坦尼克号一样的电影吗?</p> <p>B: <u>Poseidon</u> is also a classic marine film. 海神号也是一部经典的海难电影。</p>

Table 7.1: Examples of knowledge grounded conversations. Knowledge entities are underlined. Image source: Liu *et al.* (2018b).

- *Entity diffusion:* Conversations usually drift from one entity to another. In Table 7.1, the utterances in items 3 and 4 are about the entity “Titanic.” However, the entities in the responses are other similar films. Current knowledge triplets rarely capture such entity diffusion relations. The response in item 3 shows that the two entities, “Titanic” and “Waterloo Bridge,” are relevant through “love stories.” Item 4 suggests another similar shipwreck film “Titanic.”

Knowledge-grounded dialogue systems address the aforementioned challenges. Work on knowledge-grounded dialogue systems can be categorized into two groups. Methods in the first group use *structured knowledge* (given knowledge graphs) (Wu *et al.*, 2020c; Zhou *et al.*, 2020a; Wang *et al.*, 2020a; Wu *et al.*, 2020b; Xu *et al.*, 2020b; Xu *et al.*,

2020a; Jung *et al.*, 2020; Tuan *et al.*, 2019; Wu *et al.*, 2019b; Moon *et al.*, 2019; Zhou *et al.*, 2018c; Liu *et al.*, 2018b). Methods in the second group focus on using *unstructured knowledge*, such as *document-based unstructured knowledge* (given a whole document, e.g., a Wikipedia article) (Meng *et al.*, 2020a; Ma *et al.*, 2020a; Ma *et al.*, 2020b; Tian *et al.*, 2020; Ren *et al.*, 2020; Gopalakrishnan *et al.*, 2020; Li *et al.*, 2019d; Qin *et al.*, 2019; Moghe *et al.*, 2018; Zhou *et al.*, 2018d; Parthasarathi and Pineau, 2018) or *piece-based unstructured knowledge* (given some separate pieces of knowledge, e.g., Foursquare tips) (Ghazvininejad *et al.*, 2018; Dinan *et al.*, 2019; Meng *et al.*, 2020b; Kim *et al.*, 2020a; Lian *et al.*, 2019; Zheng *et al.*, 2020a; Chen *et al.*, 2020e; Zheng *et al.*, 2020b; Zhao *et al.*, 2020b; Zheng and Zhou, 2019; Lin *et al.*, 2020a). There are key research directions for both groups: (i) improving knowledge selection (Kim *et al.*, 2020a); (ii) improving knowledge-aware response generation (Zhao *et al.*, 2020b) or response selection (Young *et al.*, 2018; Zhao *et al.*, 2019c; Hua *et al.*, 2020; Sun *et al.*, 2020b); that is, given the chosen knowledge, how to better generate a response token by token or select a response from pre-defined response candidates; (iii) using multiple knowledge modalities (Liu *et al.*, 2019b; Zhao *et al.*, 2020a); that is, how to use structured, unstructured, and even other types of knowledge simultaneously; and (iv) overcoming data scarcity (Zhao *et al.*, 2019d; Li *et al.*, 2020b).

The neural knowledge diffusion model introduces knowledge into the dialogue generation. This method can match the relevant facts for the input utterance and diffuse them to similar entities (Liu *et al.*, 2018b). Early studies on knowledge selection in knowledge-grounded dialogue systems calculate the weight of each piece of knowledge and obtain a weighted sum of their representations (Ghazvininejad *et al.*, 2018; Zheng and Zhou, 2019; Lin *et al.*, 2020a; Zheng *et al.*, 2020b). Bordes and Weston (2017) employ memory networks to address restaurant reservations, using a small number of keywords to handle entity types in a knowledge base (cuisine type, location, price range, party size, rating, phone number, and address). Ghazvininejad *et al.* (2018) adapt it to memorize relevant, grounded facts for a neural conversation model. The hierarchical variational memory network (HVMN) adds hierarchical structure and a variational memory network into a neural encoder-

decoder network for non-task-oriented dialogue generation (Chen *et al.*, 2018b).

Several recent studies on knowledge selection focus on calculating a weight on each piece of knowledge and then directly sampling the amount of knowledge with the highest weight. Specifically, Dinan *et al.* (2019) design the TMemNet model that uses context to predict a distribution over pieces of knowledge and then only sample one of them into a decoder. They also introduce a knowledge selection loss to supervise knowledge selection during training. Lian *et al.* (2019) describe PostKS, which uses a context to predict a prior distribution over pieces of knowledge. During training, the prior distribution is supervised by a posterior distribution predicted by the context and the corresponding response. Similar to PostKS, Zheng *et al.* (2020b) use a context and a piece of knowledge retrieved by the context to predict a distribution over fragments of knowledge, where the probability of the amount of knowledge retrieved by the corresponding response is maximized during training. The former distills a context containing multiple utterances at different turns into a vector that is used to match with the representation of a piece of knowledge to get a score, while the latter matches every utterance in a context with a piece of knowledge to get matching features that are aggregated to get a score. A piece of knowledge is chosen based on the score list for all pieces of knowledge.

Kim *et al.* (2020a) propose a sequential knowledge transformer (SKT), which jointly uses previously selected knowledge and context to facilitate knowledge selection. Chen *et al.* (2020e) upgrade SKT by adding the knowledge distillation-based training strategy to improve knowledge selection. Zheng *et al.* (2020a) detail a method that introduces the difference information between the previously selected knowledge and the current pieces of candidate knowledge to facilitate knowledge selection. Meng *et al.* (2020b) design DukeNet, which regards tracking the previously selected knowledge and selecting the current knowledge as dual tasks within a dual learning paradigm (Qin, 2020). Zhao *et al.* (2020b) describe a method, RLKS, where the selected knowledge is sent to a decoder to generate a response that would be compared with the ground truth response to give feedback to further supervised knowledge selection. Meng *et al.* (2021a) introduce a mixed-initiative knowledge

selection method for knowledge-grounded conversations that explicitly distinguishes between user-initiative and system-initiative knowledge selection at each conversation turn to improve the performance of knowledge selection. Sun *et al.* (2021a) find that the amount of knowledge available in different languages is highly unbalanced. Hence, the authors address cross-lingual knowledge grounded conversations with a self-distillation knowledge selection and curriculum learning.

More recent years have witnessed the rapid development of pre-trained language models in open-domain dialogue systems. Large pre-trained language models can store knowledge into their parameters during pre-training and can generate informative responses in conversations (Zhao *et al.*, 2020c). Petroni *et al.* (2019) have shown that pre-trained language models can serve as knowledge bases for downstream tasks (e.g., question-answering, Roberts *et al.*, 2020). On this basis, Zhao *et al.* (2020c) have shown that pre-trained language models can ground open-domain dialogues using their implicit knowledge. Madotto *et al.* (2020) embed knowledge bases into model's parameters for end-to-end task-oriented dialogues. Roller *et al.* (2021) fine-tune pre-trained language models on knowledge-grounded conversational data. Cui *et al.* (2021) describe knowledge-enhanced fine-tuning methods to handle unseen entities. Xu *et al.* (2022b) propose a topic-aware adapter to adapt pre-trained language models in knowledge-grounded dialogues. Liu *et al.* (2022e) introduce a multi-stage prompting approach for triggering knowledge in pre-trained language models. Wu *et al.* (2022b) design lexical knowledge internalization to integrate token-level knowledge into the model's parameters. The problem of hallucination is becoming more and more challenging. Sun *et al.* (2023a) optimize an implicit knowledge eliciting process, i.e., they reduce hallucination of pre-trained language models in knowledge-grounded dialogues through a contrastive learning framework.

7.2.4 Empathetic dialogue generation

Several approaches to data-driven open-domain dialogue generation generate emotional responses based on a manually specified label to control the dynamic content of the target output (Zhou *et al.*, 2018b; Li

and Sun, 2018; Zhou and Wang, 2018; Huang *et al.*, 2018a; Wei *et al.*, 2019; Colombo *et al.*, 2019; Shen and Feng, 2020).

Unlike emotional dialogue generation, the study of empathetic dialogue generation avoids an additional step of determining which emotion type to respond to explicitly (Skowron *et al.*, 2013). Several studies (Rashkin *et al.*, 2018; Zhong *et al.*, 2019; Shin *et al.*, 2019; Rashkin *et al.*, 2019; Santhanam and Shaikh, 2019; Lin *et al.*, 2019; Lin *et al.*, 2020b; Zhong *et al.*, 2020; Majumder *et al.*, 2020; Li *et al.*, 2020c) have attempted to make dialogue models more empathetic. Rashkin *et al.* (2019) combine models in different ways to produce empathetic responses. Lin *et al.* (2019) softly combine the possible emotional responses from several separate experts. Majumder *et al.* (2020) consider polarity-based emotion clusters and emotional mimicry. Li *et al.* (2020c) propose a multi-resolution adversarial framework that considers multi-granularity emotion factors and users' feedback. Li *et al.* (2022b) investigate how to use external knowledge to explicitly improve the emotional understanding and expression in the task of empathetic dialogue generation. Sabour *et al.* (2022) focus on two aspects in empathetic dialogue generation: affection and cognition. The authors propose a method with various commonsense reasoning to improve understanding of interlocutors' situations and feelings.

Besides advances in empathetic dialogue models, the emergence of new emotion-labeled dialogue corpora has also contributed to this research field (Li *et al.*, 2017c; Hsu *et al.*, 2018; Rashkin *et al.*, 2019). Rashkin *et al.* consider a rich and well-balanced set of emotions and release a dataset, EMPATHETICDIALOGUES, where a listener responds to a speaker in an emotional situation in an empathetic way.

7.3 Emerging Directions

This section describes recent emerging directions on question-answering and dialogue systems in e-commerce. These emerging directions in QA and dialogue systems can be divided into five perspectives: safety, ethics, interpretability, privacy, and evaluation.

As discussed in Section 7.1, e-commerce QA agents aim to answer questions based on large volumes of reviews. However, reviews may not

answer these questions as they may not contain any relevant answers for the question, or a query may be poorly phrased and therefore require additional clarification. Moreover, untruthful comments and spam are widely observed in e-commerce reviews (Carmel *et al.*, 2018). Mihaylova *et al.* (2019) investigate the fact-checking problem in a QA scenario with a system to classify the veracity of answers. Zhang *et al.* (2020d) release a large-scale fact-checking dataset called AnswerFact for investigating the answer veracity in e-commerce QA. Estes *et al.* (2022) develop a high-speed fact-verification system that has a very high false statement recall and very high true statement precision to product question-answering. However, the authors still apply a rule-based method in their system, and find that pre-trained language models are unable to perform fact-checking well on structured catalog data. As limitations such as poor generalization exist in rule-based methods, how to optimize pre-trained language models in fact-checking for product question-answering still needs more attention in future research.

Ethical challenges in dialogue systems are attracting significant amounts of attention in recent years. Currently, most dialogue systems are developed from scratch with large corpora or fine-tuned through pre-trained language models. Large-scale datasets collected from the open internet have been applied during model training. However, offensive and malevolent content can be observed in the data (Si *et al.*, 2022). To avoid being unintentionally offensive or harming the user, studies have been performed to detect toxic speech around, e.g., religion, race, and violence (Tripathi *et al.*, 2019; Dinan *et al.*, 2020; Zhang *et al.*, 2021d; Kann *et al.*, 2022; Si *et al.*, 2022). Zhang *et al.* (2021d) propose a human-machine collaborative evaluation framework for reliable toxic speech detection in dialogue systems. Si *et al.* (2022) study toxic speech in open-domain dialogue systems to reveal that specific kinds of “non-toxic” queries are able to trigger an open-domain conversational assistant to output toxic responses. However, how to respond when these malevolence topics are being identified is still an open question (Kann *et al.*, 2022). As more and more e-commerce dialogue systems have also been trained based on pre-trained language models, ethical challenges will need to be tackled in future research.

Poor explainability is a challenging problem for most e-commerce QA approaches. Most e-commerce QA approaches apply end-to-end semantic matching methodologies, which tend to be black-box and directly output a matching score for each question answer pair. Zhao *et al.* (2019a) address the explainable QA problem through a hybrid retrieval-based framework. The authors employ a bidirectional recurrent neural network in the internal word representation stage and apply a keyword-aware retrieval method during the second stage. In contrast, the tf-idf ranking function naturally exhibits much better interpretability owing to its transparency and intuitiveness. Conversational recommendation is another typical application of e-commerce dialogue systems (Sun *et al.*, 2016; Mangili *et al.*, 2020). Most approaches neglect explainability when learning recommendation actions. However, Chen *et al.* (2020f) propose an incremental multi-task learning framework using user feedback for the task of explainable conversational recommendation. By considering user preferences as latent variables in a variational Bayesian manner, Ren *et al.* (2022) employ a method to estimate explicit user preferences during the dialogue.

Privacy protection has received more and more attention in dialogue systems (Papernot *et al.*, 2016; Henderson *et al.*, 2018). For many task-oriented dialogue systems, it is necessary to notice that we are using the same dialogue assistant. Recent studies on membership inference attacks have confirmed that privacy information in training data for sequence-to-sequence generative models and pre-trained language models can be attacked (Hisamoto *et al.*, 2020; Liu *et al.*, 2021b). As we discussed in Section 7.2, most e-commerce dialogue system models are designed based on sequence-to-sequence generative neural networks and pre-trained language models. By learning through interactions and communications, a dialogue assistant can inadvertently and implicitly store sensitive information. Hence, consumers' privacy information may get obtained by attackers through membership inference attacks. To address this problem, developing privacy-aware dialogue systems is likely to attract increased attention in the future.

The evaluation of e-commerce dialogue systems is a crucial part of the development process. Recent studies on evaluating dialogue systems are either through offline evaluation or human evaluation (Lamel *et al.*,

2000; Jurcicek *et al.*, 2011). Offline evaluation is often limited to single-turn assessments, while human evaluation is intrusive, time-intensive, and does not scale (Deriu *et al.*, 2020). User simulators have been applied to exhaustively enumerate user goals to generate human-like conversations for simulated evaluation (Zhang *et al.*, 2020i). However, user simulators as evaluation methods for e-commerce dialogue systems are still under-explored. Today’s simulators suffer from limited realism and evaluation capabilities (Balog *et al.*, 2021). Moreover, evaluation metrics that specifically target e-commerce aspects are still underexplored and underexploited in today’s e-commerce dialogue agents. Dedicated evaluation metrics in e-commerce search and recommendation scenarios will likely help to advance progress.

8

Conclusion and Outlook

We summarize the main topics presented in this survey in Section 8.1. In Section 8.2, we describe our outlook on future developments.

8.1 Conclusion

From the large number of user engagements on e-commerce platforms a large amount of information can be inferred. The aim of this survey has been to give a broad overview of information discovery in e-commerce portals. Our overview has included methods about user behavior modeling, search, recommendation, question answering, and dialogue systems in e-commerce.

Our strategy with this survey has been to provide a broad coverage of research directions about information discovery in e-commerce. Although we have tried our best to provide all key approaches in each direction as much as possible, the amount of technical details is limited. For areas that are broad enough to have their own survey, we have only focused on key publications and provide structure and pipelines for each direction. Additionally, we only focus on areas that are relevant to information retrieval research; studies in other areas relevant to e-commerce, such as supply chains and computational advertising, are ignored in this survey.

In our introduction, we summarized the outline and topics covered in this survey, followed by a description of basic concepts and key definitions in Section 2. Then we introduced preliminaries about e-commerce interfaces and users in Section 3. We formulated concepts of e-commerce infrastructures and summarized studies about information seeking via e-commerce interfaces. We investigated e-commerce information components, i.e., titles, product descriptions, and reviews, and detailed characteristics of consumer behaviors in e-commerce portals. We introduced studies into e-commerce user analyses concerning multiple behaviors, including clicks, purchases, engagements, and post-clicks, on e-commerce platforms.

The core of this survey is organized around five directions: e-commerce user modeling, e-commerce search, e-commerce recommendation, e-commerce QA, and e-commerce dialogue systems. We have detailed each of these in four sections (i.e., Section 4, 5, 6, and 7). Each section starts with an overview of the main direction discussed in the section, with characteristics and subtasks. After that, key research studies of each subtasks were demonstrated with some level of detail. We discussed emerging research directions at the end of each key component.

In particular, in Section 4, we introduced approaches to user modeling and profiling for e-commerce applications. We divided this section into two main components: user behavior modeling and user profiling. We provided a summary of studies on modeling these e-commerce user behavior, analyzed research on user profiling in e-commerce, and discussed emerging directions on user modeling in e-commerce.

In Section 5, we focused on search technologies in e-commerce platforms. We provided the characteristics of e-commerce search and divided research studies based on matching strategies and ranking technologies for e-commerce search scenarios, respectively. We presented approaches aiming for studying matching strategies for e-commerce search. And we studied research approaches on ranking technologies for e-commerce search. Emerging research directions were discussed at the end of the section.

In Section 6, we introduced the most prominent approaches to e-commerce recommendation methods. We summarized the key char-

acteristics of e-commerce recommendation, towards which a two-stage framework was developed that contains candidate retrieval and candidate ranking, forming the mainstream solution for e-commerce recommender systems. We reviewed models developed for the two stages and detailed mainstream learning methods for optimizing model parameters to provide a complete view of e-commerce recommender systems.

Section 7 detailed research methods for question answering and dialogue systems in e-commerce. We addressed question answering and dialogue systems in a single section as most research background and approaches are shared between these two directions. We reviewed previous work on question answering and then demonstrated the characteristics of e-commerce QA. For e-commerce question answering (QA), we described studies both on extractive QA and generative QA. For e-commerce dialogue systems, we demonstrated the patterns of e-commerce dialogue systems, especially about task-oriented dialogue systems, knowledge-grounded conversations, and empathetic dialogue systems. We discussed emerging research directions around QA and dialogue agents in Section 7.3.

8.2 Outlook

Information discovery is increasingly a mixed initiative scenario, where users and e-commerce platforms take turns. As described in the previous sections, the research presented on information discovery in e-commerce has been addressed from six angles: infrastructures, user modeling, search, recommendation, QA, and dialogue systems. As we summarized at the end of each section, a broad variety of emerging research has also been motivated following each angle. For user modeling, we consider three research topics as key emerging directions: graph learning for user behavior modeling, dynamic user behavior modeling and profiling, and multi-modal user profiling. For emerging directions in e-commerce search, we focus on applications for multi-modal e-commerce search and ranking. Online learning to rank technologies also provide key insights. We also foresee the development of new learning theories that will improve e-commerce search and ranking performance in the future. We divide emerging directions on e-commerce recommendation into three

directions: (i) reasoning, recommendation, and explanations; (ii) conversational recommendation; and (iii) unifying recommendation and search. In our view, future work on e-commerce language processing should include generating explainable reasons for search and recommendation, improving the robustness of e-commerce question answering, and improving conversational e-commerce search and recommendation.

8.2.1 Four directions

Among these emerging research approaches, we have identified potential directions of future work that are encountered across multiple angles. In particular, we list future research directions in four bigger themes: conversational search, conversational recommendation, multi-modal information discovery, and generative information discovery.

Conversational search refers to a novel search paradigm using multiple interactions between users and search engines. As we have discussed in Section 5, conversational search is increasingly receiving more attention in the IR community. Different from the traditional query-aware search paradigm, conversational search allows users to express their information need by directly conducting conversations with search engines. More recent studies have begun to apply conversational search to online shopping scenarios as it is able to provide a natural, adaptive and interactive shopping experience for consumers (Xiao *et al.*, 2021). In e-commerce, conversational search faces two challenges: imperfect product attribute schemas and product knowledge. The former exists as product attributes link lengthy multi-turn utterances with products in conversational search systems, whereas the latter derives from the lack of manually labelling in benchmark datasets. Core tasks in conversational search, e.g., search intent detection, action prediction, query selection, passage selection, and response generation (Ren *et al.*, 2021), also provide insights in future work about e-commerce conversational search systems.

Conversational recommendation refers to recommendation systems that can elicit the dynamic preferences of users and take actions based on their current needs through real-time multi-turn interactions. As we have discussed in Section 6, conversational recommendation is an

emerging direction in e-commerce recommendation. Integrating more accurate domain-specific knowledge to promote the recommendation and conversation is a challenging problem in conversational recommendation (Chen *et al.*, 2020f). Moreover, as with conversational search, current studies on conversational recommendation suffer from a lack of manually labelled data in benchmark datasets. Recent studies on empathetic dialogue systems reveal that there exist some kind of dependency between commonsense knowledge and emotional preference (Li *et al.*, 2022b; Siro *et al.*, 2022). Hence, conversational recommender systems that jointly combine emotion detection and knowledge exploration are worth studying in future.

Multi-modal information can be widely observed in many e-commerce scenarios, e.g., user behavior, search, recommendation, and dialogue systems. Most previous e-commerce information discovery approaches are constructed only based on text understanding and retrieval; addressing multi-modal information, e.g., images and videos, still appears to be difficult. In future work, it is more and more important to tackle challenges about multi-modality in e-commerce scenarios. Multimedia technologies focusing on integrating various types of modalities are expected to help to understand those multi-modal information for various types of e-commerce applications. It is interesting to explore multi-modal generation through powerful generative deep neural networks in e-commerce review generation, question answering, and dialogue systems.

Large-scale generative models have the potential to significantly enhance various e-commerce information discovery applications, such as search, recommendation, and conversational AI. Transformer-based pre-trained language models like BERT have already proven effective in both search and recommendation tasks in e-commerce. More recently, large language models (LLMs) based on auto-regressive mechanisms, such as T5 and GPT, have demonstrated promising capabilities in understanding and generating human-like information, making them valuable for e-commerce contexts. Moreover, while traditional two-stage paradigms (i.e., retrieval followed by re-ranking) have been widely used in e-commerce search and recommendation scenarios. They face two limitations: (i) heterogeneous modules with different optimization objectives

may lead to sub-optimal performance; and (ii) a large document index is needed which may come with substantial memory and computational requirements. This has motivated research into end-to-end solutions using generative models. Recent studies on generative retrieval, such as DSI (Tay *et al.*, 2022), have shown encouraging performance on several information retrieval benchmarks, suggesting that exploring generative models for end-to-end e-commerce search and recommendation could be a promising direction for future research.

8.2.2 Beyond accuracy

Besides the directions for future work listed above, we also consider the following important issues when it comes to information discovery tasks in e-commerce: *fairness*, *trustworthiness*, and *explainability* (Roegiest *et al.*, 2019; de Rijke, 2023).

Recently, the problem of bias has attracted considerable attention in the IR community, in multiple contexts, e.g., for user behavior modeling, profiling, ranking, and recommendation. To address the bias problem, *fairness* is considered as a significant metric during the optimization procedure. Early on, fairness was studied from the perspective of information exposure regarding sensitive attributes such as gender and race (Singh and Joachims, 2018). Fairness in IR also focuses on how to let different items receive equal exposure, or exposure proportional to their utility or impact, depending on which exposure distribution is considered to be fair by the system (Morik *et al.*, 2020; Chen *et al.*, 2020a). In e-commerce, ranking-based interfaces are quite common in various scenarios; hence, fairness is a matter of great importance to information discovery in e-commerce. Future work on interactive fairness-aware reranking can be helpful for debiasing user modeling, search, recommendation, and answer generation in e-commerce platforms (Sarvi *et al.*, 2022). Also, knowledge-based and dynamic fairness-aware methods are able to address more real-world challenges. The trade-off between accuracy and fairness is of importance in e-commerce search and recommendation scenarios, where equally treating different groups has been shown to sacrifice the performance (Ariannezhad *et al.*, 2023). To address this problem, an important research direction is to understand the dimensions of causality and design fairness-aware algorithms.

Fake news and fake information are increasingly widespread. It is now viewed as one of the greatest threats on the web (Zhou and Zafarani, 2020). As we have discussed in Section 7, spam and fake reviews and answers are widely observed in many e-commerce platforms. Therefore, pursuing *trustworthiness* has become an important issue in e-commerce question answering and dialogue systems. Several studies distinguish spam or fake reviews in online review systems via graph neural networks (Kaghazgaran *et al.*, 2018; Rao *et al.*, 2020; Liu *et al.*, 2020h; Dou *et al.*, 2020). Textual generation methods based on deep neural networks have been applied to fake reviews or spam generation by camouflaged fraudsters. Thus, distinguishing the authenticity of e-commerce information is a challenging task. Other patterns in the reviews, e.g., sentiments and emotions, can be applied to improve the detection. Also, investigating inconsistency problems under multiple domains provides new avenues of research.

Explainability in e-commerce aims to answer the question about why we receive a specific ranking, recommendation, or answer result. The task of explainability can be divided into explainability of the learning models and explainability of the results. The former aims to provide more transparent learning details for the proposed methods, whereas the latter focuses on provide more explainable results to various application scenarios, e.g., search, recommendation, and question answering, etc. Explainability methods have been shown to be effective for enhancing the e-commerce search and recommendation (Zhang *et al.*, 2018c; Liu *et al.*, 2020f). For future work, it is important to evaluate if and how users and other stakeholders are satisfied with the explanations generated from an e-commerce system, especially as these are increasingly conversational in nature (Lucic *et al.*, 2021). Generating coherent, faithful, and naturally-sounding explanations based on a sequence of reasoning steps (including search or recommendation system output) is still difficult.

As we have shown in this survey, the number of research studies in the area of information discovery for e-commerce is increasing rapidly. We believe that this is only the beginning. The recent launch of a dedicated product search track at TREC seems to confirm this.¹ The

¹<https://trec-product-search.github.io>

volume of the work described in this survey and the steady pace of publications in the field, together with the arrival of open research challenges indicate a promising future ahead. A lot remains to be done.

Acknowledgements

This survey grew out of a tutorial “Information Discovery in e-Commerce” taught at SIGIR 2018. We thank the audience for their feedback and questions.

We also thank our colleagues Mozhdeh Ariannezhad, Hongshen Chen, Jiawei Chen, Zhumin Chen, Songgaojun Deng, Zhuoye Ding, Yue Feng, Stefan Grafberger, Paul Groth, Yulong Gu, Shuyu Guo, Ziyi Guo, Maria Heuss, Mariya Hendriksen, Na Huang, Sami Jullien, Barrie Kersbergen, Jiahuan Lei, Dongdong Li, Ming Li, Xiang Li, Xinyi Li, Zhenyang Li, Xiaozhong Liu, Hengliang Luo, Si Luo, Yougang Lyu, Jun Ma, Yao Ma, Pengjie Ren, Emma de Rijke, Fatemeh Sarvi, Sebastian Schelter, Xinlei Shi, Clemencia Siro, Hongye Song, Olivier Sprangers, Changlong Sun, Fei Sun, Weiwei Sun, Jiliang Tang, Bart Voorn, Jingang Wang, Shuaiqiang Wang, Xuepeng Wang, Zihan Wang, Long Xia, Xin Xin, Zhen Zhang, Jiashu Zhao, Xiangyu Zhao, and Yihong Zhao for help, feedback, and inspiration.

We thank our editors Yiqun Liu and Mark Sanderson for support, patience, and valuable feedback.

This research was supported by Alibaba DAMO Academy, Baidu.com, JD.com, and Meituan, as well as AIRLab, a collaboration between Ahold Delhaize and the University of Amsterdam, the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation

for Scientific Research, <https://hybrid-intelligence-centre.nl>, project nr. 024.004.022, project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO), project ROBUST with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO) and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023, and the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Appendix

Datasets

In this appendix, we list benchmark datasets that are relevant for studying information discovery in e-commerce. We follow the topical organization of our sections, and divide the datasets into five types: e-commerce infrastructures, e-commerce user modeling, e-commerce search, e-commerce recommendation, and e-commerce QA & dialogues.

A.1 Datasets for E-commerce Infrastructures

To begin, we list benchmark datasets about e-commerce interfaces and users:

- **Taobao short title dataset** (Sun *et al.*, 2018a): This dataset contains 411,246 title-product pairs in 94 categories. Each item in the dataset is represented as a triple $\langle Q, K, S \rangle$, where Q denotes the products' original titles, K refers to the background knowledge about the products, and S represents the human-written short titles.
- **eCOM-C2C dataset** about product categories and titles (Wang *et al.*, 2018a): This dataset takes advantage of realistic data from a well-known C2C website in China. The dataset contains 185,386 triplets in the Women's Clothes category. Each item in the dataset is represented as a triple $\langle S, T, Q \rangle$, where S refers to a product's original title, T denotes a handcrafted short title, and Q is a successful transaction-leading search queries.

- **Walmart product summarization dataset** (Mukherjee *et al.*, 2020): The dataset includes 40,445 top-selling Walmart grocery products during the calendar year 2018, together with their product titles and corresponding human-generated summaries. There are also descriptions, brand names, and category information of the products.
- **Taobao multi-modal title dataset** (Miao *et al.*, 2020): The dataset contains 114,278 original titles with corresponding short titles and product images. The short titles are manually written by professional editors, whereas the images are selected by the seller.
- **Walmart e-commerce product dataset** (Mukherjee, 2021): The dataset contains five parts: D-search includes the top 12 million product search queries on Walmart.com and their frequencies over a one year period. D-product includes 250,000 top-selling Walmart products over a six month period. D-com-human includes 40,445 human-generated title compressions from the Walmart catalog across eight different product categories. D-meta-auto contains 40,000 meta-training examples. And D-meta-human is a dataset consisting of 16,000 human-generated 1-shot title compression examples.
- **LESD4EC dataset** (Gong *et al.*, 2019): The dataset consists of 6,481,623 pairs of original and short product titles in a module in Taobao named “Youhuashuo.” Each product in this dataset includes a long product title and a short title summary written by professional writers, along with a high-quality image and attributes tags.

Table A.1 summarizes the key statistics of the datasets listed above.

A.2 Datasets for E-commerce User Modeling

Next, we list benchmark datasets about e-commerce user modeling:

- **Taobao Tianchi consumer dataset** (Kim *et al.*, 2021): The dataset includes responses of users to advertisements of inventory in the user profile and advertising information. The time length of the data is eight days, and the dataset is divided into four ta-

Table A.1: Statistics of datasets about e-commerce infrastructures.

Datasets	Statistics					References
	#Dataset size	#Number of category	#Avg.length of original titles	#Avg.length of short titles	#Avg.length of background knowledge	
Taobao short title dataset	453,138	94	25.34	7.73	5.92	(Sun <i>et al.</i> , 2018a)
eCOM-C2C dataset	185,386	1	25.1	7.5	8.3	(Wang <i>et al.</i> , 2018a)
Walmart product summarization dataset	40,445		4/10/35	1/2/5		(Mukherjee <i>et al.</i> , 2020)
Taobao multi-modal title dataset	114,278					(Miao <i>et al.</i> , 2020)
Walmart e-commerce product dataset	40,000 + 16,000	4				(Mukherjee, 2021)
LESDEEC dataset	6,481,623		12	5		(Gong <i>et al.</i> , 2019)

bles: advertisement features, user profiles, past shopping behavior that users engaged in, and who received the advertisement with responses.²

- **Instacart.MB dataset** (Sheng *et al.*, 2021): The Instacart Market Basket (Instacart.MB) dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user in the dataset, there are between 4 and 100 of their orders, with the sequence of products purchased in each order.³
- **Bing advertising service dataset** (Lian *et al.*, 2021): The dataset contains user click logs within a two week period from the Bing Native Advertising service. It also includes users' online behavior history before their corresponding clicks. The user behavior sequences are truncated to 100 in the dataset.
- **Feeds user dataset** (Yi *et al.*, 2021): The feeds dataset is collected on Microsoft News App from August 1, 2020 to September 1, 2020. It contains 643,177 news items, over 10,000 users, 320,925 impressions, and 970,846 clicks.
- **JD user profiling dataset** (Chen *et al.*, 2019f): This dataset is collected from one of the largest e-commerce platforms in China. In this dataset, users, items, and attributes reflect real-world e-commerce consumers, products, and words in the titles of the products respectively. The profiles of users are the age and gender labels.
- **Twitter user behavior dataset** (Al Zamal *et al.*, 2012): Each attribute dataset consists of approximately 400 labeled Twitter users, 200 with one label (e.g., "female") and 200 with a second label (e.g., "male"). In addition, all of the friends of these labeled users are identified; for each of these labeled and neighbor users, the most recent 1,000 tweets generated by the user were collected.
- **UCL social media user profiling dataset** (Liang *et al.*, 2017): This dataset was collected by UCL's Big Data Institute. The data set includes 1,375 active Twitter users chosen randomly and their

²<https://tianchi.aliyun.com/dataset/dataDetail?dataId=56>

³<https://www.kaggle.com/c/instacart-market-basket-analysis/data>

tweets from the time they registered until May 31, 2015. The dataset has 3.78 million tweets in total. The length of a tweet is 12 words on average.

- **CALL dataset** (Dong *et al.*, 2014): The dataset is extracted from a collection of more than 1 billion (i.e., 1,000,229,603) call and text-message events from an anonymous country, which spans from August 2008 to September 2008. The data does not contain any communication content.
- **W-NUT dataset** (Han *et al.*, 2016): This is a user-level dataset of the geolocation prediction shared task released at the W-NUT workshop in 2016. The dataset consists of over 1 million training users, 10,000 development users, and 10,000 test users. The ground truth location of a user is decided by majority voting of the closest city center.
- **Facebook user profiling dataset** (Farnadi *et al.*, 2018): This is a re-collected dataset based on Facebook’s MyPersonality project dataset.⁴ The dataset includes information about each user’s demographics, friendship links, Facebook activities (e.g., number of group affiliations, page likes, education, and work history), status updates, profile picture, and Big Five Personality scores (ranging from 1 to 5).

Table A.2 summarizes the key statistics of the datasets listed above.

A.3 Datasets for E-commerce Search

We list benchmark datasets about e-commerce search as follows:

- **QUARTS e-commerce search dataset** (Nguyen *et al.*, 2020): This is a human-labeled dataset of query-item pairs, obtained from an e-commerce search platform. There are in total 3.2 million pairs of which only a small fraction are mismatches. About 100,000 labeled pairs are used as a separate test set. Another 3 million query-item pairs are deemed “matched” by considering items that are purchased frequently in response to those queries from the search logs.

⁴<http://www.mypersonality.org>

Table A.2: Statistics of datasets about e-commerce user modeling.

Datasets	Statistics					References
	#Users	#Items	#Interactions	#Avg.seq.len	TimeSpan	
Taobao Tianchi consumer dataset	1,140,000			26,000,000	20170506-20170513	(Kim <i>et al.</i> , 2021)
Instacart.MB dataset	11,464	42,207	7,764,043	677.25		(Sheng <i>et al.</i> , 2021)
Bing advertising service dataset	748,000	409,000		74		(Lian <i>et al.</i> , 2021)
Feeds user dataset	10,000	643,177	970,846			(Yi <i>et al.</i> , 2021)
JD user profiling dataset	54,161	203,712				(Chen <i>et al.</i> , 2019f)
Twitter user behavior dataset	400	400,000		1,000		(Al Zamal <i>et al.</i> , 2012)
UCL social media user profiling dataset	1,375	3,780,000		12	time of registration-20150531	(Liang <i>et al.</i> , 2017)
CALL dataset	1,090,000				200808-200809	(Dong <i>et al.</i> , 2014)
W-NUT dataset	1,020,000	13,000,000				(Han <i>et al.</i> , 2016)
Facebook user profiling dataset	5,670	49,372				(Farnadi <i>et al.</i> , 2018)

- **SCEM product search dataset** (Bi *et al.*, 2020b): The dataset contains three category-specific datasets, namely, “Toys & Games,” “Garden & Outdoor,” and “Cell Phones & Accessories,” from the logs of a commercial product search engine spanning ten months between years 2017 and 2018. The datasets include up to a few million query sessions containing several hundred thousand unique queries.
- **Walmart product search dataset** (Karmaker Santu *et al.*, 2017): This is a subset obtained from Walmart’s online product catalog. The dataset consists of more than 2,800 randomly selected product search queries and a catalog of around 5 million products. For each query, the top 120 products are retrieved.
- **Walmart query log dataset** (Magnani *et al.*, 2019): This is a large query log dataset on shoe segments during a six-month window from May 2018 to October 2018 on Walmart.com. Historical data of the extra features such as clicks and orders are collected from the query log six months before May 2018. The dataset is composed of more than 100 million query and product pairs, of which there are more than 1 million unique queries and more than 1 million unique item titles.
- **Bestbuy dataset** (Duan *et al.*, 2013b): The dataset consists of a full crawl of the “Laptop & Netbook Computers” category of Bestbuy.com. In total, there are 864 laptops in the database, each entity has 44 specifications on average. And 260 laptops have user reviews. The annotated datasets contain 40 queries, on average, there are 2.8 keywords per query and 3.8 keywords per query for the hard queries.
- **Amazon product dataset** (Bi *et al.*, 2021; McAuley *et al.*, 2015): The Amazon product dataset is a well-known benchmark for product search and recommendation. It contains information for millions of customers, products and associated metadata, including descriptions, reviews, brands, and categories.⁵
- **Etsy product search dataset** (Wu *et al.*, 2018a): The dataset contains 4 weeks worth of search log data with clicks and purchases

⁵<http://jmcauley.ucsd.edu/data/amazon>

from Etsy.⁶ In total, there are 334,931 search sessions with 239,928 queries and 6,347,251 items. In total, 270,239 buyers and 550,025 sellers are involved in the transactions, whereas 631,778 keywords are used by sellers to describe their items.

Table A.3 summarizes the key statistics of the datasets listed above.

A.4 Datasets for E-commerce Recommendations

Next, we list benchmark datasets about e-commerce recommender systems:

- **Amazon product dataset** (He and McAuley, 2016b; McAuley *et al.*, 2015): For e-commerce recommendations, the Amazon product dataset is split by top-level product categories in amazon and is notable for its high sparsity and variability. This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996–July 2014. This dataset includes reviews (i.e., ratings, text, helpfulness votes), product metadata (i.e., descriptions, category information, price, brand, and image features), and links (i.e., substitutive/complementary relations).
- **Amazon soc dataset** (McAuley *et al.*, 2015): A large-scale database of 230,000 users; each data sample includes a user’s profile, user feedback on a product, and social relationship among users. More specifically, the user’s profile includes gender, income, age, and hobby. User feedback includes the user’s comments and browsing history.
- **AliExpress dataset** (Ahmed *et al.*, 2021): This dataset is collected from an online retailer service owned by the Alibaba group. There are about 2,260,923 records from AliExpress, the data for about fourteen months from January 1, 2019 to February 23, 2020. The dataset contains 1,506,850 users that submitted reviews against 49,221 items in 205 different categories, such as electronics, entertainment, education, house, and garden, etc., and the items are rated from 1 to 5 scale.

⁶<https://www.etsy.com>

Table A.3: Statistics of datasets about e-commerce search.

Datasets	Statistics				References
	#Queries	#Products	#Pairs	Product title length	
QUARTS e-commerce search dataset			3,200,000		(Nguyen <i>et al.</i> , 2020)
SCEM product search dataset					(Bi <i>et al.</i> , 2020b)
SCEM-Toys&Games				13.14±6.46	381,620
SCEM-Garden&Outdoor				16.39±7.38	1,054,980
SCEM-CellPhones&Accessories				22.02±7.34	194,022
Walmart product search dataset	2,800	5,000,000			(Karmaker Santu <i>et al.</i> , 2017)
Walmart query log dataset	1,000,000+	1,000,000+	100,000,000+		(Magnani <i>et al.</i> , 2019)
Bestbuy dataset	40	864			(Duan <i>et al.</i> , 2013b)
Amazon product dataset					(Bi <i>et al.</i> , 2021; McAuley <i>et al.</i> , 2015)
Etsy product search dataset	239,928	6,347,251		26.5	(Wu <i>et al.</i> , 2018a)

- **Instacart orders dataset:** This is an anonymized dataset collected from the Instacart site.⁷ It contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, 4 and 100 of his/her orders are provided, with the sequence of products purchased in each order. There are also the week and hour of the day the order was placed and a relative measure of time between orders.⁸
- **MovieLens dataset** (Harper and Konstan, 2015): This is a widely used benchmark dataset collected from <https://movielens.org>. The dataset contains user ratings and timestamps for the movie. There is side-info of users and movies. According to the year and the size of the dataset, there are multiple specific versions.⁹
- **Yoochoose dataset** (Ben-Shimon *et al.*, 2015): This dataset is collected from the 2015 recommender systems challenge (RecSys Challenge 2015). The dataset includes six months of user activities for a large European e-commerce business that sells various consumer goods, including garden tools, toys, clothes, electronics, and more. There are 33,040,175 records in the click file and 1,177,769 records in the buys file. The training set consists of 9,512,786 unique sessions, and the test file consists of 2,312,432 click sessions.
- **Alibaba Cloud/TIANCHI dataset** (Zhu *et al.*, 2018): The dataset was randomly selected from Taobao; it contains about 1 million users with their behavior, which includes clicks, purchases, adding items to the shopping cart, and item favoring from November 25 to December 3, 2017. The dataset is organized in a very similar form to MovieLens-20M, i.e., each line represents a specific user-item interaction, which consists of user ID, item ID, item's category ID, behavior type, and timestamp, separated by commas.¹⁰

Table A.4 summarizes the key statistics of the datasets listed above.

⁷<https://www.instacart.com>

⁸<https://www.instacart.com/datasets/grocery-shopping-2017>

⁹<https://grouplens.org/datasets/movielens/>

¹⁰<https://tianchi.aliyun.com/dataset/dataDetail?dataId=649&userId=1&lang=en-us>

Table A.4: Statistics of datasets about e-commerce recommendations.

Datasets	Statistics					References
	#Users	#Items	#Records	#Categories	TimeSpan	
Amazon product dataset	20,980,320	5,933,184	143,663,229	11		(He and McAuley, 2016; McAuley <i>et al.</i> , 2015)
Amazon soc dataset	230,000					(McAuley <i>et al.</i> , 2015)
AliExpress dataset	1,506,850	49,221	2,260,923	205	20190101-20200223	(Ahmed <i>et al.</i> , 2021)
Instacart orders dataset	200,000+	3,000,000				Instacart dataset ^a
Movielens-ML100K	943	1,682	100,000		199709-199804	
Movielens-ML1M	6,040	3,706	1,000,209		200004-200302	(Harper and Konstan, 2015)
Movielens-ML10M	69,878	10,681	10,000,054		199501-200901	
Movielens-ML20M	138,493	27,278	20,000,263		199501-201503	
Yoochoose dataset	9,249,729	52,739	34,154,697			(Ben-Shimon <i>et al.</i> , 2015)
Alibaba Cloud/TIANCHI dataset	1,000,000	4,023,451	100,934,102	9,378	20171125-20171203	(Zhu <i>et al.</i> , 2018)

^a<https://www.instacart.com/datasets/grocery-shopping-2017>

A.5 Datasets for E-commerce QA and Dialogues

Next, we list benchmark datasets about e-commerce question answering and dialogue systems:

- **JD product question answering** (Gao *et al.*, 2019b): This dataset consists of online product-aware QA pairs. Each QA pair is associated with the reviews and attributes of the corresponding product. The corpus covers 469,953 products and 38 product categories. The average length of the question is 9.03 words, and the ground truth answer is 10.3 words. The average number of attributes is 9.0 key-value pairs.
- **Taobao question answering dataset** (Chen *et al.*, 2019e): This dataset is collected on Taobao. The dataset includes 4,457 and 47,979 products under the category Cellphone and Household Electrics, respectively. For each product, the associated question-answering pairs and user reviews are included. After pre-processing, Cellphone/Household Electrics products have 356,842 and 798,688 QA-pairs in two subsets, respectively.
- **Amazon complex question/answer dataset** (McAuley and Yang, 2016): This dataset was collected from Amazon, including reviews and descriptions of products and QA data. This dataset contains 1.4 million answered questions on 191 thousand products and 13 million related reviews.
- **Hierarchical product review corpus** (Yu *et al.*, 2011): This corpus contains consumer reviews on 11 popular products in four domains. These reviews were crawled from several prevalent forum websites, including cnet.com, viewpoints.com, reevoo.com, and gsmarena.com. All of the reviews were posted between June 2009 and September 2010. The aspects of the reviews, as well as the opinions on the aspects, were manually annotated.
- **Amazon question answering dataset** (Deng *et al.*, 2020): This dataset is constructed by combining Amazon Question Answering Dataset (McAuley and Yang, 2016) and Amazon Product Review Dataset (He and McAuley, 2016b) by matching the product ID. In this dataset, each QA sample contains a question, a reference answer, the answer opinion type label, and a set of relevant re-

view snippets with corresponding ratings. After collecting the final dataset, each QA sample contains a question, a reference answer, the answer opinion type label, and a set of relevant review snippets with corresponding ratings. There are three categories, namely Electronics, Home & Kitchen, and Sports & Outdoors, with 193,960 (Electronics), 90,269 (Home & Kitchen), and 50,020 pairs (Sports & Outdoors).

- **JDDC e-commerce dialogue dataset** (Chen *et al.*, 2020c): JDDC is a large-scale real scenario Chinese E-commerce conversation corpus, with more than one million multi-turn dialogues, 20 million utterances, and 150 million words, which contains conversations about after-sales topics between users and customer service staffs in an e-commerce scenario. JDDC was updated with multi-modal customer service information in 2021 (Zhao *et al.*, 2021a).
- **E-commerce dialogue corpus dataset** (Zhang *et al.*, 2018d): The dataset is collected from the real-world conversations between customers and customer service staff on Taobao. It contains over five types of conversations (i.e., commodity consultation, logistics express, recommendation, negotiation, and chitchat) based on over 20 commodities.¹¹

Table A.5 summarizes the key statistics for the datasets listed above.

¹¹<https://drive.google.com/file/d/154J-neBo20ABtSmJDvm7DK0eTuieAuvw/view?usp=sharing>

Table A.5: Statistics of datasets about e-commerce question answering and dialogues.

Datasets	Statistics					References
	#Products	#Q-A pairs	#Categories	#Avg.length of questions	#Avg.length of ground truth	
JD product question answering	469,953		38	9.03	10.3	(Gao <i>et al.</i> , 2019b)
Taobao question answering dataset	4,457/47,979	356,842/798,688	2	9/8	13/13	(Chen <i>et al.</i> , 2019e)
Amazon complex question/answer dataset	191,185	1,447,173	8			(McAuley and Yang, 2016)
Hierarchical product review corpus	11		4			(Yu <i>et al.</i> , 2011)
Amazon question answering dataset		334,249	3			(Deng <i>et al.</i> , 2020)
JDDC e-commerce dialogue dataset	1,024,196	20,451,337	150,716,172	7.4	20	(Chen <i>et al.</i> , 2020c)
E-commerce Dialogue Corpus dataset	1,000,000 (Train) 10,000 (Valid) 10,000 (Test)			7.02 (Train) 6.99 (Valid) 7.11 (Test)	5.51 (Train) 5.48 (Valid) 5.64 (Test)	(Zhang <i>et al.</i> , 2018d)

References

- Abdollahpouri, H. (2019). “Popularity Bias in Ranking and Recommendation”. In: *Proceedings of AIES*. 529–530.
- Adomavicius, G. and A. Tuzhilin. (1999). “User Profiling in Personalization Applications through Rule Discovery and Validation”. In: *Proceedings of KDD*. 377–381.
- Adomavicius, G. and A. Tuzhilin. (2005). “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions”. *IEEE Transactions on Knowledge & Data Engineering*. 17(6): 734–749.
- Afzali, J., A. M. Drzewiecki, K. Balog, and S. Zhang. (2023). “User-SimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems”. In: *Proceedings of WSDM*. 1160–1163.
- Agichtein, E., E. Brill, and S. Dumais. (2006). “Improving Web Search Ranking by Incorporating User Behavior Information”. In: *Proceedings of SIGIR*. 19–26.
- Ahmed, A., K. Saleem, O. Khalid, and U. Rashid. (2021). “On Deep Neural Network for Trust Aware Cross Domain Recommendations in E-commerce”. *Expert Systems with Applications*. 174: 114757.
- Ahn, D., V. Jijkoun, G. Mishne, K. Müller, M. de Rijke, S. Schlobach, M. Voorhees, and L. Buckland. (2004). “Using Wikipedia at the TREC QA Track”. In: *Proceedings of TREC*.

- Ai, Q., K. Bi, J. Guo, and W. B. Croft. (2018). “Learning a Deep Listwise Context Model for Ranking Refinement”. In: *Proceedings of SIGIR*. 135–144.
- Ai, Q., D. N. Hill, S. Vishwanathan, and W. B. Croft. (2019a). “A Zero Attention Model for Personalized Product Search”. In: *Proceedings of CIKM*. 379–388.
- Ai, Q., Y. Zhang, K. Bi, X. Chen, and W. B. Croft. (2017). “Learning a Hierarchical Embedding Model for Personalized Product Search”. In: *Proceedings of SIGIR*. 645–654.
- Ai, Q., Y. Zhang, K. Bi, and W. B. Croft. (2019b). “Explainable Product Search with a Dynamic Relation Embedding Model”. *ACM Transactions on Information Systems (TOIS)*. 38(1): 1–29.
- Al Zamal, F., W. Liu, and D. Ruths. (2012). “Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors”. In: *Proceedings of ICWSM*.
- Alonso, O. and S. Mizzaro. (2009). “Relevance Criteria for E-Commerce: A Crowdsourcing-Based Experimental Analysis”. In: *Proceedings of SIGIR*. 760–761.
- Ameixa, D., L. Coheur, P. Fialho, and P. Quaresma. (2014). “Luke, I am Your Father: Dealing with Out-of-Domain Requests by Using Movies Subtitles”. In: *Proceedings of IVA*. Springer International Publishing. 13–21.
- Anderson, P. and E. Anderson. (2002). “The New E-commerce Intermediaries”. *MIT Sloan Management Review*. 43(4): 53.
- Andoni, A. and P. Indyk. (2006). “Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions”. In: *Proceedings of FOCS*. 459–468.
- Angelidis, S., R. K. Amplayo, Y. Suhara, X. Wang, and M. Lapata. (2021). “Extractive Opinion Summarization in Quantized Transformer Spaces”. *Transactions of the Association for Computational Linguistics*. 9: 277–293.
- Angelidis, S. and M. Lapata. (2018). “Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised”. In: *Proceedings of EMNLP*. 3675–3686.

- Anwaar, M. U., D. Rybalko, and M. Kleinstauber. (2020). “Mend the Learning Approach, Not the Data: Insights for Ranking E-Commerce Products”. In: *Proceedings of ECML-PKDD*. 257–272.
- Ariannezhad, M., S. Jullien, M. Li, M. Fang, S. Schelter, and M. de Rijke. (2022). “ReCANet: A Repeat Consumption-Aware Neural Network for Next Basket Recommendation in Grocery Shopping”. In: *Proceedings of SIGIR*. ACM. 1240–1250.
- Ariannezhad, M., S. Jullien, P. Nauts, M. Fang, S. Schelter, and M. de Rijke. (2021). “Understanding Multi-channel Customer Behavior in Retail”. In: *Proceedings of CIKM*. ACM.
- Ariannezhad, M., M. Li, S. Schelter, and M. de Rijke. (2023). “A Personalized Neighborhood-based Model for Within-basket Recommendation in Grocery Shopping”. In: *Proceedings of WSDM*. ACM.
- Ariannezhad, M., S. Schelter, and M. de Rijke. (2020). “Demand Forecasting in the Presence of Privileged Information”. In: *Proceedings of AALTD. LNCS 12588*. Springer. 46–62.
- Aryafar, K., D. Guillory, and L. Hong. (2017). “An Ensemble-based Approach to Click-Through Rate Prediction for Promoted Listings at Etsy”. In: *Proceedings of ADKDD*. 10.
- Azzopardi, L., M. Dubiel, M. Halvey, and J. Dalton. (2018). “Conceptualizing Agent-human Interactions During the Conversational Search Process”. In: *Proceedings of CAIR*. ACM.
- Bahdanau, D., K. Cho, and Y. Bengio. (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *Proceedings of ICLR*.
- Bai, T., J.-Y. Nie, W. X. Zhao, Y. Zhu, P. Du, and J.-R. Wen. (2018). “An Attribute-aware Neural Attentive Model for Next Basket Recommendation”. In: *Proceedings of SIGIR*. ACM. 1201–1204.
- Balaneshin-kordan, S. and A. Kotov. (2018). “Deep Neural Architecture for Multi-Modal Retrieval Based on Joint Embedding Space for Text and Images”. In: *Proceedings of WSDM*. 28–36.
- Balog, K. (2011). “On the Investigation of Similarity Measures for Product Resolution”. In: *Proceedings of LDH*. 49–54.
- Balog, K., D. Maxwell, P. Thomas, and S. Zhang. (2021). “Sim4IR: The SIGIR 2021 Workshop on Simulation for Information Retrieval Evaluation”. In: *Proceedings of SIGIR*. 2697–2698.

- Balog, K., E. Meij, and M. de Rijke. (2010). “Entity Search: Building Bridges between Two Worlds”. In: *Proceedings of 3rd international semantic search workshop*. 1–5.
- Banchs, R. E. and H. Li. (2013). “IRIS: A Chat-oriented Dialogue System based on the Vector Space Model”. In: *Proceedings of ACL*. 37–42.
- Bao, W., H. Wen, S. Li, X.-Y. Liu, Q. Lin, and K. Yang. (2020). “GMCM: Graph-based Micro-behavior Conversion Model for Post-click Conversion Rate Estimation”. In: *Proceedings of SIGIR*. 2201–2210.
- Battaglia, P., R. Pascanu, M. Lai, D. J. Rezende, *et al.* (2016). “Interaction Networks for Learning about Objects, Relations and Physics”. In: *Proceedings of NIPS*. 4502–4510.
- Baudiš, P. (2015). “YodaQA: A Modular Question Answering System Pipeline”. In: *Proceedings of POSTER*. 1156–1165.
- Bearden, W. O. and R. G. Netemeyer. (1999). *Handbook of Marketing Scales: Multi-item Measures for Marketing and Consumer Behavior Research*. Sage.
- Belkin, N. J., C. Cool, A. Stein, and U. Thiel. (1995). “Cases, Scripts, and Information-seeking Strategies: On the Design of Interactive Information Retrieval Systems”. *Expert Systems with Applications*. 9(3): 379–395.
- Bell, R. M. and Y. Koren. (2007). “Lessons from the Netflix Prize Challenge”. *Acm SIGKDD Explorations Newsletter*. 9(2): 75–79.
- Bellman, S., G. L. Lohse, and E. J. Johnson. (1999). “Predictors of Online Buying Behavior”. *Communications of the ACM*. 42(12): 32–38.
- Ben-Shimon, D., A. Tsikinovsky, M. Friedmann, B. Shapira, L. Rokach, and J. Hoerle. (2015). “RecSys Challenge 2015 and the YOO-CHOOSE Dataset”. In: *Proceedings of RecSys*. 357–358.
- Bendersky, M. and W. B. Croft. (2009). “Analysis of Long Queries in a Large Scale Search Log”. In: *Proceedings of WSCD*. 8–14.
- Berant, J., A. Chou, R. Frostig, and P. Liang. (2013). “Semantic Parsing on Freebase from Question-Answer Pairs”. In: *Proceedings of EMNLP*. 1533–1544.

- Bernardi, L., J. Kamps, J. Kiseleva, and M. J. Müller. (2015). “The Continuous Cold Start Problem in e-Commerce Recommender Systems”. *arXiv preprint arXiv:1508.01177*.
- Bevilacqua, M., G. Ottaviano, P. Lewis, S. Yih, S. Riedel, and F. Petroni. (2022). “Autoregressive Search Engines: Generating Substrings as Document Identifiers”. *Proceedings of NIPS*. 35: 31668–31683.
- Bhatt, R., V. Chaoji, and R. Parekh. (2010). “Predicting Product Adoption in Large-Scale Social Networks”. In: *Proceedings of CIKM*. 1039–1048.
- Bi, K., Q. Ai, and W. B. Croft. (2020a). “A Transformer-based Embedding Model for Personalized Product Search”. In: *Proceedings of SIGIR*. 1521–1524.
- Bi, K., Q. Ai, and W. B. Croft. (2021). “Learning a Fine-Grained Review-based Transformer Model for Personalized Product Search”. In: *Proceedings of SIGIR*. 123–132.
- Bi, K., Q. Ai, Y. Zhang, and W. B. Croft. (2019a). “Conversational Product Search Based on Negative Feedback”. In: *Proceedings of CIKM*. 359–368.
- Bi, K., C. H. Teo, Y. Dattatreya, V. Mohan, and W. B. Croft. (2019b). “A Study of Context Dependencies in Multi-page Product Search”. In: *Proceedings of CIKM*. 2333–2336.
- Bi, K., C. H. Teo, Y. Dattatreya, V. Mohan, and W. B. Croft. (2020b). “Leverage Implicit Feedback for Context-aware Product Search”. In: *Proceedings of SIGIR workshop on e-commerce*.
- Bian, S., W. X. Zhao, K. Zhou, J. Cai, Y. He, C. Yin, and J.-R. Wen. (2021). “Contrastive Curriculum Learning for Sequential User Behavior Modeling via Data Augmentation”. In: *Proceedings of CIKM*. 3737–3746.
- Bjerva, J., N. Bhutani, B. Golshan, W.-C. Tan, and I. Augenstein. (2020). “SubjQA: A Dataset for Subjectivity and Review Comprehension”. *arXiv preprint arXiv:2004.14283*.
- Black, A. W., S. Burger, A. Conkie, H. Hastie, S. Keizer, O. Lemon, N. Merigaud, G. Parent, G. Schubiner, B. Thomson, J. Williams, K. Yu, S. J. Young, and M. Eskénazi. (2011). “Spoken Dialog Challenge 2010: Comparison of Live and Control Test Results”. In: *Proceedings of SIGDIAL*.

- Blake, T., C. Nosko, and S. Tadelis. (2016). “Returns to Consumer Search: Evidence from eBay”. In: *Proceedings of EC*. 531–545.
- Bogina, V. and T. Kuflik. (2017). “Incorporating Dwell Time in Session-Based Recommendations with Recurrent Neural Networks”. In: *Proceedings of CEUR*. Vol. 1922. 57–59.
- Bordes, A., N. Usunier, S. Chopra, and J. Weston. (2015). “Large-scale Simple Question Answering with Memory Networks”. *arXiv preprint arXiv:1506.02075*.
- Bordes, A. and J. Weston. (2017). “Learning End-to-End Goal-Oriented Dialog”. In: *Proceedings of ICLR*.
- Borisov, A., I. Markov, M. de Rijke, and P. Serdyukov. (2016). “A Neural Click Model for Web Search”. In: *Proceedings of Web Conference*. 531–541.
- Braynov, S. (2003). “Personalization and Customization Technologies”. *The Internet Encyclopedia*.
- Bražinskis, A., M. Lapata, and I. Titov. (2020). “Few-Shot Learning for Opinion Summarization”. In: *Proceedings of EMNLP*. 4119–4135.
- Breiman, L. (2001). “Random Forests”. *Machine Learning*. 45(1): 5–32.
- Brill, E., J. J. Lin, M. Banko, S. T. Dumais, A. Y. Ng, *et al.* (2001). “Data-Intensive Question Answering.” In: *Proceedings of TREC*. 90.
- Broder, A. (2002). “A Taxonomy of Web Search”. In: *ACM SIGIR Forum*. ACM New York. 3–10.
- Brown, M., N. Pope, and K. Voges. (2003). “Buying or Browsing? An Exploration of Shopping Orientations and Online Purchase Intention”. *European Journal of Marketing*. 37(11/12): 1666–1684.
- Brown, T. B. (2020). “Language Models are Few-shot Learners”. *arXiv preprint arXiv:2005.14165*.
- Bruna, J., W. Zaremba, A. Szlam, and Y. LeCun. (2014). “Spectral Networks and Locally Connected Networks on Graphs”. In: *Proceedings of ICLR*.
- Budzianowski, P., T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gasic. (2018). “MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”. In: *Proceedings of EMNLP*. 5016–5026.

- Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. (2005). “Learning to Rank Using Gradient Descent”. In: *Proceedings of SIGIR*. 89–96.
- Burges, C. J. (2010). “From RankNet to LambdaRank to LambdaMART: An Overview”. *Learning*. 11(23-581): 81.
- Buscaldi, D. and P. Rosso. (2006). “Mining Knowledge from Wikipedia for the Question Answering task”. In: *Proceedings of LREC*. 727–730.
- Campagna, G., A. Foryciarz, M. Moradshahi, and M. Lam. (2020). “Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking”. In: *Proceedings of ACL*. 122–132.
- Cao, C., H. Ge, H. Lu, X. Hu, and J. Caverlee. (2017a). “What Are You Known For?: Learning User Topical Profiles with Implicit and Explicit Footprints”. In: *Proceedings of SIGIR*. 743–752.
- Cao, K. and S. Clark. (2017). “Latent Variable Dialogue Models and their Diversity”. In: *Proceedings of EACL*. 182–187.
- Cao, Z., M. Long, J. Wang, and Q. Yang. (2017b). “Transitive Hashing Network for Heterogeneous Multimedia Retrieval”. In: *Proceedings of AAAI*. 81–87.
- Cao, Z., T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. (2007). “Learning to Rank: From Pairwise Approach to Listwise Approach”. In: *Proceedings of ICML*. 129–136.
- Carbonell, J. G. and J. Goldstein. (1998). “The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries”. In: *Proceedings of SIGIR*. 335–336.
- Carmel, D., L. Lewin-Eytan, and Y. Maarek. (2018). “Product Question Answering Using Customer Generated Content - Research Challenges”. In: *Proceedings of SIGIR*. 1349–1350.
- Cen, Y., J. Zhang, X. Zou, C. Zhou, H. Yang, and J. Tang. (2020). “Controllable Multi-interest Framework for Recommendation”. In: *Proceedings of KDD*. 2942–2951.
- Chan, P. P., X. Hu, L. Zhao, D. S. Yeung, D. Liu, and L. Xiao. (2018). “Convolutional Neural Networks based Click-Through Rate Prediction with Multiple Feature Sequences.” In: *Proceedings of IJCAI*. 2007–2013.

- Chan, Z., X. Chen, Y. Wang, J. Li, Z. Zhang, K. Gai, D. Zhao, and R. Yan. (2019). “Stick to Facts: Towards Fidelity-oriented Product Description Generation”. In: *Proceedings of EMNLP*. 4959–4968.
- Chang, J., C. Zhang, Z. Fu, X. Zang, L. Guan, J. Lu, Y. Hui, D. Leng, Y. Niu, Y. Song, *et al.* (2023). “TWIN: Two-stage Interest Network for Lifelong User Behavior Modeling in CTR Prediction at Kuaishou”. In: *Proceedings of KDD*. 3785–3794.
- Chang, W.-C., D. Jiang, H.-F. Yu, C. H. Teo, J. Zhang, K. Zhong, K. Kolluri, Q. Hu, N. Shandilya, V. Ievgrafov, *et al.* (2021). “Extreme Multi-label Learning for Semantic Matching in Product Search”. In: *Proceedings of KDD*. 2643–2651.
- Chapelle, O. (2014). “Modeling Delayed Feedback in Display Advertising”. In: *Proceedings of KDD*. 1097–1105.
- Chapelle, O., E. Manavoglu, and R. Rosales. (2015). “Simple and Scalable Response Prediction for Display Advertising”. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 5(4): 61.
- Chen, B., Y. Wang, Z. Liu, R. Tang, W. Guo, H. Zheng, W. Yao, M. Zhang, and X. He. (2021a). “Enhancing Explicit and Implicit Feature Interactions via Information Sharing for Parallel Deep CTR Models”. In: *Proceedings of CIKM*. 3757–3766.
- Chen, C., Y. Yang, J. Zhou, X. Li, and F. Bao. (2018a). “Cross-Domain Review Helpfulness Prediction Based on Convolutional Neural Networks with Auxiliary Domain Discriminators”. In: *Proceedings of NAACL*. 602–607.
- Chen, D., A. Fisch, J. Weston, and A. Bordes. (2017a). “Reading Wikipedia to Answer Open-Domain Questions”. In: *Proceedings of ACL*. 1870–1879.
- Chen, D. and W.-t. Yih. (2020). “Open-Domain Question Answering”. In: *Proceedings of ACL*. 34–37.
- Chen, D., H. Chen, Y. Yang, A. Lin, and Z. Yu. (2021b). “Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems”. In: *Proceedings of NAACL*. 3002–3017.
- Chen, H., S. Shi, Y. Li, and Y. Zhang. (2021c). “Neural Collaborative Reasoning”. In: *Proceedings of Web Conference*. 1516–1527.

- Chen, H., X. Liu, D. Yin, and J. Tang. (2017b). “A Survey on Dialogue Systems: Recent Advances and New Frontiers”. *ACM SIGKDD Explorations Newsletter*. 19(2).
- Chen, H., Z. Ren, J. Tang, Y. E. Zhao, and D. Yin. (2018b). “Hierarchical Variational Memory Network for Dialogue Generation”. In: *Proceedings of Web Conference*.
- Chen, H., M. Sun, C. Tu, Y. Lin, and Z. Liu. (2016a). “Neural Sentiment Classification with User and Product Attention”. In: *Proceedings of EMNLP*. 1650–1659.
- Chen, J., H. Dong, Y. Qiu, X. He, X. Xin, L. Chen, G. Lin, and K. Yang. (2021d). “AutoDebias: Learning to Debias for Recommendation”. In: *Proceedings of SIGIR*. 21–30.
- Chen, J., H. Dong, X. Wang, F. Feng, M. Wang, and X. He. (2020a). “Bias and Debias in Recommender System: A Survey and Future Directions”. *arXiv preprint arXiv:2010.03240*.
- Chen, J., B. Sun, H. Li, H. Lu, and X.-S. Hua. (2016b). “Deep CTR Prediction in Display Advertising”. In: *Proceedings of MM*. 811–820.
- Chen, L., G. Zhang, and H. Zhou. (2017c). “Improving the Diversity of Top-N Recommendation via Determinantal Point Process”. *CoRR*. abs/1709.05135.
- Chen, L., Z. Guan, W. Zhao, W. Zhao, X. Wang, Z. Zhao, and H. Sun. (2019a). “Answer Identification from Product Reviews for User Questions by Multi-Task Attentive Networks”. In: *Proceedings of AAAI*. 45–52.
- Chen, L., Z. Chen, B. Tan, S. Long, M. Gašić, and K. Yu. (2019b). “AgentGraph: Toward Universal Dialogue Management With Structured Deep Reinforcement Learning”. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 27(9): 1378–1391.
- Chen, L., B. Lv, C. Wang, S. Zhu, B. Tan, and K. Yu. (2020b). “Schema-Guided Multi-Domain Dialogue State Tracking with Graph Attention Neural Networks”. In: *Proceedings of AAAI*. 7521–7528.
- Chen, M., R. Liu, L. Shen, S. Yuan, J. Zhou, Y. Wu, X. He, and B. Zhou. (2020c). “The JDDC Corpus: A Large-Scale Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service”. In: *Proceedings of LREC*. 459–466.

- Chen, Q. and W. Wang. (2019). “Sequential Matching Model for End-to-end Multi-turn Response Selection”. In: *Proceedings of ICASSP*. IEEE. 7350–7354.
- Chen, Q., J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. (2019c). “Towards Knowledge-Based Recommender Dialog System”. In: *Proceedings of EMNLP*. 1803–1813.
- Chen, Q., J. Lin, Y. Zhang, H. Yang, J. Zhou, and J. Tang. (2019d). “Towards Knowledge-Based Personalized Product Description Generation in E-commerce”. In: *Proceedings of KDD*. ACM. 3040–3050.
- Chen, S., C. Li, F. Ji, W. Zhou, and H. Chen. (2019e). “Driven Answer Generation for Product-Related Questions in E-Commerce”. In: *Proceedings of WSDM*. 411–419.
- Chen, T. and C. Guestrin. (2016). “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of KDD*. 785–794.
- Chen, T., H. Yin, G. Ye, Z. Huang, Y. Wang, and M. Wang. (2020d). “Try This Instead: Personalized and Interpretable Substitute Recommendation”. In: *Proceedings of SIGIR*. 891–900.
- Chen, W., Y. Gu, Z. Ren, X. He, H. Xie, T. Guo, D. Yin, and Y. Zhang. (2019f). “Semi-supervised User Profiling with Heterogeneous Graph Attention Networks.” In: *Proceedings of IJCAI*. 2116–2122.
- Chen, X., F. Meng, P. Li, F. Chen, S. Xu, B. Xu, and J. Zhou. (2020e). “Bridging the Gap between Prior and Posterior Knowledge Selection for Knowledge-Grounded Dialogue Generation”. In: *Proceedings of EMNLP*. 3426–3437.
- Chen, X., Y. Zhang, H. Xu, Y. Cao, Z. Qin, and H. Zha. (2018c). “Visually Explainable Recommendation”. *arXiv preprint arXiv:1801.10288*.
- Chen, Y., X. Zhao, and M. de Rijke. (2017d). “Top-N Recommendation with High-dimensional Side Information via Locality Preserving Projection”. In: *Proceedings of SIGIR*. ACM. 985–988.
- Chen, Y., J. Li, C. Liu, C. Li, M. Anderle, J. J. McAuley, and C. Xiong. (2021e). “Modeling Dynamic Attributes for Next Basket Recommendation”. *CoRR*. abs/2109.11654.

- Chen, Y., J. Jin, H. Zhao, P. Wang, G. Liu, J. Xu, and B. Zheng. (2022). “Asymptotically Unbiased Estimation for Delayed Feedback Modeling via Label Correction”. In: *Proceedings of Web Conference*. 369–379.
- Chen, Z., A. Mukherjee, and B. Liu. (2014). “Aspect Extraction with Automated Prior Knowledge Learning”. In: *Proceedings of ACL*. 347–358.
- Chen, Z., X. Wang, X. Xie, M. Parsana, A. Soni, X. Ao, and E. Chen. (2020f). “Towards Explainable Conversational Recommendation.” In: *Proceedings of IJCAI*. 2994–3000.
- Cheng, D., J. Chen, W. Peng, W. Ye, F. Lv, T. Zhuang, X. Zeng, and X. He. (2022). “IHGNN: Interactive Hypergraph Neural Network for Personalized Product Search”. In: *Proceedings of Web Conference*. 256–265.
- Cheng, H.-T., L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, *et al.* (2016). “Wide & Deep Learning for Recommender Systems”. In: *Proceedings of Workshop on Deep Learning for Recommender Systems*. 7–10.
- Cheng, Y. (2022). “Dynamic Explicit Embedding Representation for Numerical Features in Deep CTR Prediction”. In: *Proceedings of CIKM*. 3888–3892.
- Cheng, Y. and Y. Xue. (2021). “Looking at CTR Prediction Again: Is Attention All You Need?” In: *Proceedings of SIGIR*. 1279–1287.
- Choi, K., G. Fazekas, and M. B. Sandler. (2016). “Towards Playlist Generation Algorithms Using RNNs Trained on Within-Track Transitions”. *ArXiv*.
- Christakopoulou, K., A. Beutel, R. Li, S. Jain, and E. H. Chi. (2018). “Q&R: A Two-stage Approach toward Interactive Recommendation”. In: *Proceedings of KDD*. 139–148.
- Christakopoulou, K., F. Radlinski, and K. Hofmann. (2016). “Towards Conversational Recommender Systems”. In: *Proceedings of KDD*. 815–824.
- Chu, X., J. Zhao, L. Zou, and D. Yin. (2022). “H-ERNIE: A Multi-Granularity Pre-Trained Language Model for Web Search”. In: *Proceedings of SIGIR*. 1478–1489.

- Chuklin, A., I. Markov, and M. de Rijke. (2015). *Click Models for Web Search. Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool Publishers.
- Colombo, P., W. Witon, A. Modi, J. Kennedy, and M. Kapadia. (2019). “Affect-Driven Dialog Generation”. In: *Proceedings of NAACL*. 3734–3743.
- Covington, P., J. Adams, and E. Sargin. (2016). “Deep Neural Networks for YouTube Recommendations”. In: *Proceedings of RecSys*. ACM. 191–198.
- Croft, W. B. and R. H. Thompson. (1987). “I3R: A new approach to the design of document retrieval systems”. *Journal of the American Society for Information Science*. 38(6): 389–404.
- Cuayáhuitl, H., S. Keizer, and O. Lemon. (2015). “Strategic Dialogue Management via Deep Reinforcement Learning”. *arXiv preprint arXiv:1511.08099*.
- Cufoglu, A. (2014). “User Profiling-A Short Review”. *International Journal of Computer Applications*. 108(3).
- Cui, L., Y. Wu, S. Liu, and Y. Zhang. (2021). “Knowledge Enhanced Fine-Tuning for Better Handling Unseen Entities in Dialogue Generation”. In: *Proceedings of EMNLP*. 2328–2337.
- Dagan, A., I. Guy, and S. Novgorodov. (2021). “An Image is Worth a Thousand Terms? Analysis of Visual E-Commerce Search”. In: *Proceedings of SIGIR*. 102–112.
- Dai, Q., H. Li, P. Wu, Z. Dong, X.-H. Zhou, R. Zhang, R. Zhang, and J. Sun. (2022). “A Generalized Doubly Robust Learning Framework for Debiasing Post-Click Conversion Rate Prediction”. In: *Proceedings of KDD*. 252–262.
- Dai, S., J. Liu, Z. Dou, H. Wang, L. Liu, B. Long, and J.-R. Wen. (2023). “Contrastive Learning for User Sequence Representation in Personalized Product Search”. In: *Proceedings of KDD*. 380–389.
- Das, M., Z. Wang, E. Jaffe, M. Chattopadhyay, E. Fosler-Lussier, and R. Ramnath. (2019). “Learning to Answer Subjective, Specific Product-Related Queries using Customer Reviews by Adversarial Domain Adaptation”. *arXiv preprint arXiv:1910.08270*.
- De Cao, N., G. Izacard, S. Riedel, and F. Petroni. (2020). “Autoregressive Entity Retrieval”. *arXiv preprint arXiv:2010.00904*.

- de Rijke, M. (2023). “Beyond Accuracy Goals, Again”. In: *Proceedings of WSDM*. ACM. 2–3.
- de Vries, A. P., A.-M. Vercoustre, J. A. Thom, N. Craswell, and M. Lalmas. (2007). “Overview of the INEX 2007 Entity Ranking Track”. In: *Proceedings of INEX*. 245–251.
- Deffayet, R., T. Thonet, J.-M. Renders, and M. de Rijke. (2023). “Generative Slate Recommendation with Reinforcement Learning”. In: *Proceedings of WSDM*. ACM. 580–588.
- Defferrard, M., X. Bresson, and P. Vandergheynst. (2016). “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering”. In: *Proceedings of NIPS*. 3844–3852.
- Degenhardt, J., S. Kallumadi, M. de Rijke, L. Si, A. Trotman, and X. Yinghui. (2017). “eCom: The SIGIR 2017 Workshop on eCommerce”. In: *Proceedings of SIGIR*.
- Demartini, G., J. Gaugaz, and W. Nejdl. (2009). “A Vector Space Model for Ranking Entities and Its Application to Expert Search”. In: *Proceedings of ECIR*. 189–201.
- Deng, L., G. Tur, X. He, and D. Hakkani-Tur. (2012). “Use of Kernel Deep Convex Networks and End-to-end Learning for Spoken Language Understanding”. In: *Proceedings of SLT*. 210–215.
- Deng, Y., Y. Li, W. Zhang, B. Ding, and W. Lam. (2022). “Toward Personalized Answer Generation in E-Commerce via Multi-perspective Preference Modeling”. *ACM Transactions on Information Systems*. 40(4): 1–28.
- Deng, Y., W. Zhang, and W. Lam. (2020). “Opinion-Aware Answer Generation for Review-Driven Question Answering in E-Commerce”. In: *Proceedings of CIKM*. 255–264.
- Deng, Y., W. Zhang, W. Xu, W. Lei, T.-S. Chua, and W. Lam. (2023). “A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems”. *ACM Transactions on Information Systems*. 41(3): 1–25.
- Deoras, A. and R. Sarikaya. (2013). “Deep Belief Network based Semantic Taggers for Spoken Language Understanding”. In: *Proceedings of Interspeech*. 2713–2717.

- Deriu, J., Á. Rodrigo, A. Otegi, G. Echevoyen, S. Rosset, E. Agirre, and M. Cieliebak. (2020). “Survey on Evaluation Methods for Dialogue Systems”. *Artificial Intelligence Review*. 54: 755–810.
- Dhingra, B., L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng. (2017). “Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access”. In: *Proceedings of ACL*.
- Dinan, E., A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston. (2020). “Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation”. In: *Proceedings of EMNLP*. 8173–8188.
- Dinan, E., S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. (2019). “Wizard of Wikipedia: Knowledge-Powered Conversational Agents”. In: *Proceedings of ICLR*.
- Ding, S., F. Feng, X. He, Y. Liao, J. Shi, and Y. Zhang. (2021). “Causal Incremental Graph Convolution for Recommender System Retraining”. *arXiv preprint arXiv:2108.06889*.
- Ding, X., T. Liu, J. Duan, and J.-Y. Nie. (2015). “Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network.” In: *Proceedings of AAAI*. 2389–2395.
- Dong, L., S. Huang, F. Wei, M. Lapata, M. Zhou, and K. Xu. (2017). “Learning to Generate Product Reviews from Attributes”. In: *Proceedings of EACL*. 623–632.
- Dong, Y., Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. (2014). “Inferring User Demographics and Social Strategies in Mobile Social Networks”. In: *Proceedings of KDD*. 15–24.
- Dou, Y., Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu. (2020). “Enhancing Graph Neural Network-Based Fraud Detectors against Camouflaged Fraudsters”. In: *Proceedings of CIKM*. 315–324.
- Duan, H., C. Zhai, J. Cheng, and A. Gattani. (2013a). “A Probabilistic Mixture Model for Mining and Analyzing Product Search Log”. In: *Proceedings of CIKM*. 2179–2188.
- Duan, H., C. Zhai, J. Cheng, and A. Gattani. (2013b). “Supporting Keyword Search in Product Database: A Probabilistic Approach”. *Proceedings of VLDB*. 6(14): 1786–1797.

- Dunn, M., L. Sagun, M. Higgins, V. U. Guney, V. Cirik, and K. Cho. (2017). “SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine”. *arXiv preprint arXiv:1704.05179*.
- Dušek, O. and F. Jurčiček. (2016). “Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings”. In: *Proceedings of ACL*. 45.
- Eickhoff, C., J. Teevan, R. White, and S. Dumais. (2014). “Lessons from the Journey: A Query Log Analysis of within-Session Learning”. In: *Proceedings of WSDM*. 223–232.
- Eksombatchai, C., P. Jindal, J. Z. Liu, Y. Liu, R. Sharma, C. Sugnet, M. Ulrich, and J. Leskovec. (2018). “Pixie: A System for Recommending 3+ Billion Items to 200+ Million Users in Real-Time”. In: *Proceedings of Web Conference*. 1775–1784.
- Eric, M., L. Krishnan, F. Charette, and C. D. Manning. (2017). “Key-Value Retrieval Networks for Task-Oriented Dialogue”. In: *Proceedings of SIGDIAL*. 37–49.
- Eric, M. and C. D. Manning. (2017). “A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue”. In: *Proceedings of EACL*. 468–473.
- Estes, A., N. Vedula, M. D. Collins, M. Cecil, and O. Rokhlenko. (2022). “Fact Checking Machine Generated Text with Dependency Trees”. In: *Proceedings of EMNLP*.
- Faggioli, G., M. Polato, and F. Aiolli. (2020). “Recency Aware Collaborative Filtering for Next Basket Recommendation”. In: *Proceedings of UMAP*. ACM. 80–87.
- Fan, L., Q. Li, B. Liu, X.-M. Wu, X. Zhang, F. Lv, G. Lin, S. Li, T. Jin, and K. Yang. (2022). “Modeling User Behavior with Graph Convolution for Personalized Product Search”. In: *Proceedings of Web Conference*.
- Fan, M., C. Feng, L. Guo, M. Sun, and P. Li. (2019a). “Product-Aware Helpfulness Prediction of Online Reviews”. In: *Proceedings of Web Conference*. 2715–2721.
- Fan, M., C. Feng, M. Sun, P. Li, and H. Wang. (2019b). “Reading Customer Reviews to Answer Product-related Questions”. In: *Proceedings of SDM*. 567–575.

- Fan, M., Y. Feng, M. Sun, P. Li, H. Wang, and J. Wang. (2018). “Multi-Task Neural Learning Architecture for End-to-End Identification of Helpful Reviews”. In: *2Proceedings of ASONAM*. 343–350.
- Fan, M., J. Guo, S. Zhu, S. Miao, M. Sun, and P. Li. (2019c). “MOBIUS: Towards the Next Generation of Query-Ad Matching in Baidu’s Sponsored Search”. In: *Proceedings of KDD*. 2509–2517.
- Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. (2008). “LIBLINEAR: A Library for Large Linear Classification”. *Journal of machine Learning research*. 9: 1871–1874.
- Farnadi, G., J. Tang, M. De Cock, and M.-F. Moens. (2018). “User Profiling through Deep Multimodal Fusion”. In: *Proceedings of WSDM*. 171–179.
- Fawcett, T. and F. J. Provost. (1996). “Combining Data Mining and Machine Learning for Effective User Profiling.” In: *Proceedings of KDD*. 8–13.
- Fei, H., Y. Ren, S. Wu, B. Li, and D. Ji. (2021). “Latent Target-Opinion as Prior for Document-Level Sentiment Classification: A Variational Approach from Fine-Grained Perspective”. In: *Proceedings of Web Conference*. 553–564.
- Feng, J., C. Tao, Z. Li, C. Liu, T. Shen, and D. Zhao. (2022a). “Reciprocal Learning of Knowledge Retriever and Response Ranker for Knowledge-Grounded Conversations”. In: *Proceedings of COLING*. 389–399.
- Feng, Y., A. Lipani, F. Ye, Q. Zhang, and E. Yilmaz. (2022b). “Dynamic Schema Graph Fusion Network for Multi-Domain Dialogue State Tracking”. In: *Proceedings of ACL*. 115–126.
- Feng, Y., Z. Ren, W. Zhao, M. Sun, and P. Li. (2021). “Multi-Type Textual Reasoning for Product-Aware Answer Generation”. In: *Proceedings of SIGIR*. 1135–1145.
- Feng, Y., F. Lv, W. Shen, M. Wang, F. Sun, Y. Zhu, and K. Yang. (2019). “Deep Session Interest Network for Click-Through Rate Prediction”. In: *Proceedings of IJCAI*. 2301–2307.
- Ferro, N., C. Lucchese, M. Maistro, and R. Perego. (2017). “On Including the User Dynamic in Learning to Rank”. In: *Proceedings of SIGIR*. 1041–1044.

- Ferro, N., C. Lucchese, M. Maistro, and R. Perego. (2019). “Boosting Learning to Rank with User dynamics and Continuation Methods”. *Information Retrieval Journal*. 23(6): 1–27.
- Ferrucci, D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, *et al.* (2010). “Building Watson: An overview of the DeepQA project”. *AI magazine*. 31(3): 59–79.
- Fetahu, B., Z. Chen, O. Rokhlenko, and S. Malmasi. (2023). “InstructPTS: Instruction-Tuning LLMs for Product Title Summarization”. In: *Proceedings of EMNLP*. 663–674.
- Freund, Y., R. Iyer, R. E. Schapire, and Y. Singer. (2003). “An Efficient Boosting Algorithm for Combining Preferences”. *Journal of Machine Learning Research*. 4(Nov): 933–969.
- Gäde, M., M. Hall, H. Huurdeman, J. Kamps, M. Koolen, M. Skov, T. Bogers, and D. Walsh. (2016). “Overview of the SBS 2016 Interactive Track”. In: *Proceedings of CLEF*. 1024–1038.
- Gao, C., X. Wang, X. He, and Y. Li. (2022). “Graph Neural Networks for Recommender System”. In: *Proceedings of WSDM*. 1623–1625.
- Gao, C., W. Lei, X. He, M. de Rijke, and T.-S. Chua. (2021a). “Advances and Challenges in Conversational Recommender Systems: A Survey”. *AI Open*. 2: 100–126.
- Gao, C., S. Yuan, Z. Zhang, H. Yin, and J. Shao. (2019a). “BLOMA: Explain Collaborative Filtering via Boosted Local Rank-One Matrix Approximation”. In: *Proceedings of DASFAA*. 487–490.
- Gao, H., J. Tang, X. Hu, and H. Liu. (2013). “Modeling Temporal Effects of Human Mobile Behavior on Location-Based Social Networks”. In: *Proceedings of CIKM*. 1673–1678.
- Gao, J., M. Galley, and L. Li. (2018). “Neural Approaches to Conversational AI”. In: *Proceedings of SIGIR*. 1371–1374.
- Gao, S., X. Chen, Z. Ren, D. Zhao, and R. Yan. (2021b). “Meaningful Answer Generation of E-Commerce Question-Answering”. *ACM Transactions on Information Systems*. 39(2): 1–26.
- Gao, S., Z. Ren, Y. Zhao, D. Zhao, D. Yin, and R. Yan. (2019b). “Product-Aware Answer Generation in E-Commerce Question-Answering”. In: *Proceedings of WSDM*. 429–437.

- Gao, S., A. Sethi, S. Agarwal, T. Chung, and D. Hakkani-Tur. (2019c). “Dialog State Tracking: A Neural Reading Comprehension Approach”. In: *Proceedings of SIGDIAL*. 264–273.
- Gao, S., Y. Zhang, Z. Ou, and Z. Yu. (2020). “Paraphrase Augmented Task-Oriented Dialog Generation”. In: *Proceedings of ACL*.
- Geigle, C. and C. Zhai. (2016). “Scaling up Online Question Answering via Similar Question Retrieval”. In: *Proceedings of ACM Conference on Learning @ Scale*. 257–260.
- Gelli, F., X. He, T. Chen, and T.-S. Chua. (2017). “How Personality Affects our Likes: Towards a Better Understanding of Actionable Images”. In: *Proceedings of MM*. 1828–1837.
- Gers, F. A., J. Schmidhuber, and F. Cummins. (1999). “Learning to Forget: Continual Prediction with LSTM”. In: *Proceedings of ICANN*. 850–855.
- Ghanem, B., L. L. Coleman, J. R. Dexter, S. von der Ohe, and A. Fyshe. (2022). “Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask”. In: *Findings of ACL*. 2131–2146.
- Ghazvininejad, M., C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley. (2018). “A Knowledge-Grounded Neural Conversation Model”. In: *Proceedings of AAAI*. 5110–5117.
- Gilotte, A., C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. (2018). “Offline A/B Testing for Recommender Systems”. In: *Proceedings of WSDM*. 198–206.
- Go, A., R. Bhayani, and L. Huang. (2009). “Twitter Sentiment Classification using Distant Supervision”. *CS224N project report, Stanford*. 1(12): 2009.
- Goddeau, D., H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. (1996). “A Form-based Dialogue Manager for Spoken Language Applications”. In: *Proceedings of ICSLP*. 701–704 vol.2.
- Godoy, D. and A. Amandi. (2005). “User Profiling in Personal Information Agents: A Survey”. *The Knowledge Engineering Review*. 20(4): 329–361.
- Gogna, A. and A. Majumdar. (2017). “Balancing Accuracy and Diversity in Recommendations Using Matrix Completion Framework”. *Knowledge-Based Systems*. 125: 83–95.

- Gong, Y., Z. Jiang, Y. Feng, B. Hu, K. Zhao, Q. Liu, and W. Ou. (2020). “EdgeRec: Recommender System on Edge in Mobile Taobao”. In: *Proceedings of CIKM*. 2477–2484.
- Gong, Y., X. Luo, K. Q. Zhu, W. Ou, Z. Li, and L. Duan. (2019). “Automatic Generation of Chinese Short Product Titles for Mobile Display”. In: *Proceedings of AAAI*. 9460–9465.
- Gopalakrishnan, K., B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, D. Hakkani-Tür, and A. A. AI. (2020). “Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations”. In: *Proceedings of InterSpeech*.
- Gopalan, P., J. M. Hofman, and D. M. Blei. (2015). “Scalable Recommendation with Hierarchical Poisson Factorization”. In: *Proceedings of UAI*. 326–335.
- Graepel, T., J. Q. Candela, T. Borchert, and R. Herbrich. (2010). “Web-Scale Bayesian Click-through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing Search Engine”. In: *Proceedings of ICML*. 13–20.
- Gu, J.-C., T. Li, Q. Liu, Z.-H. Ling, Z. Su, S. Wei, and X. Zhu. (2020a). “Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots”. In: *Proceedings of CIKM*. 2041–2044.
- Gu, J., Z. Lu, H. Li, and V. O. Li. (2016). “Incorporating Copying Mechanism in Sequence-to-Sequence Learning”. In: *Proceedings of ACL*. 1631–1640.
- Gu, Y., Z. Ding, S. Wang, and D. Yin. (2020b). “Hierarchical User Profiling for E-commerce Recommender Systems”. In: *Proceedings of WSDM*. 223–231.
- Guidotti, R., G. Rossetti, L. Pappalardo, F. Giannotti, and D. Pedreschi. (2017). “Market Basket Prediction Using User-Centric Temporal Annotated Recurring Sequences”. In: *Proceedings of ICDM*. IEEE Computer Society. 895–900.
- Gunawan, D. D. and K.-H. Huarng. (2015). “Viral Effects of Social Network and Media on Consumers’ Purchase Intention”. *Journal of Business Research*. 68(11): 2237–2241.
- Guo, F., C. Liu, and Y. M. Wang. (2009). “Efficient Multiple-Click Models in Web Search”. In: *Proceedings of WSDM*. 124–131.

- Guo, H., R. Tang, Y. Ye, Z. Li, and X. He. (2017). “DeepFM: A Factorization-Machine based Neural Network for CTR Prediction”. In: *Proceedings of IJCAI*. 1725–1731.
- Guo, J., Y. Fan, Q. Ai, and W. B. Croft. (2016). “A Deep Relevance Matching Model for Ad-hoc Retrieval”. In: *Proceedings of CIKM*. 55–64.
- Guo, J., Y. Fan, X. Ji, and X. Cheng. (2019a). “Matchzoo: A Learning, Practicing, and Developing System for Neural Text Matching”. In: *Proceedings of SIGIR*. 1297–1300.
- Guo, J., K. Shuang, J. Li, Z. Wang, and Y. Liu. (2022a). “Beyond the Granularity: Multi-Perspective Dialogue Collaborative Selection for Dialogue State Tracking”. In: *Proceedings of ACL*. 2320–2332.
- Guo, S., L. Zou, Y. Liu, W. Ye, S. Cheng, S. Wang, H. Chen, D. Yin, and Y. Chang. (2021). “Enhanced Doubly Robust Learning for Debiasing Post-click Conversion Rate Estimation”. In: *Proceedings of SIGIR*. 275–284.
- Guo, S., M. Wang, and J. Leskovec. (2011). “The Role of Social Networks in Online Shopping: Information Passing, Price of Trust, and Consumer Choice”. In: *Proceedings of EC*. 157–166.
- Guo, W., C. Zhang, Z. He, J. Qin, H. Guo, B. Chen, R. Tang, X. He, and R. Zhang. (2022b). “MISS: Multi-Interest Self-Supervised Learning Framework for Click-Through Rate Prediction”. In: *Proceedings of ICDE*. 727–740.
- Guo, Y., Z. Cheng, L. Nie, Y. Wang, J. Ma, and M. Kankanhalli. (2019b). “Attentive Long Short-Term Preference Modeling for Personalized Product Search”. *ACM Transactions on Information Systems*. 37(2): 1–27.
- Guo, Y., Z. Cheng, L. Nie, Y. Wang, J. Ma, and M. Kankanhalli. (2019c). “Attentive Long Short-Term Preference Modeling for Personalized Product Search”. *ACM Transactions on Information Systems*. 37(2): 1–27.
- Guo, Y., Z. Cheng, L. Nie, X.-S. Xu, and M. Kankanhalli. (2018). “Multi-Modal Preference Modeling for Product Search”. In: *Proceedings of MM*. 1865–1873.

- Gupta, M., N. Kulkarni, R. Chanda, A. Rayasam, and Z. C. Lipton. (2019). “AmazonQA: A Review-Based Question Answering Task”. *arXiv preprint arXiv:1908.04364*.
- Gupta, V., D. Varshney, H. Jhamtani, D. Kedia, and S. Karwa. (2014). “Identifying Purchase Intent from Social Posts”. In: *Proceedings of ICWSM*. 180–186.
- Guu, K., K. Lee, Z. Tung, P. Pasupat, and M. Chang. (2020). “Retrieval Augmented Language Model Pre-Training”. In: *Proceedings of ICML*. 3929–3938.
- Hajli, N., J. Sims, A. H. Zadeh, and M.-O. Richard. (2017). “A Social Commerce Investigation of the Role of Trust in a Social Networking Site on Purchase Intentions”. *Journal of Business Research*. 71: 133–141.
- Hamilton, W., Z. Ying, and J. Leskovec. (2017). “Inductive Representation Learning on Large Graphs”. In: *Proceedings of NIPS*. 1024–1034.
- Han, B., A. Rahimi, L. Derczynski, and T. Baldwin. (2016). “Twitter Geolocation Prediction Shared Task of the 2016 Workshop on Noisy User-generated Text”. In: *Proceedings of Workshop on Noisy User-generated Text*. 213–217.
- Han, J., T. Hong, B. Kim, Y. Ko, and J. Seo. (2021). “Fine-grained Post-training for Improving Retrieval-based Dialogue Systems”. In: *Proceedings of NAACL*. 1549–1558.
- Han, W., H. Chen, Z. Hai, S. Poria, and L. Bing. (2022). “SANCL: Multimodal Review Helpfulness Prediction with Selective Attention and Natural Contrastive Learning”. In: *Proceedings of COLING*. 5666–5677.
- Hansen, T., J. M. Jensen, and H. S. Solgaard. (2004). “Predicting Online Grocery Buying Intention: A Comparison of the Theory of Reasoned Action and the Theory of Planned Behavior”. *International Journal of Information Management*. 24(6): 539–550.
- Harper, F. M. and J. A. Konstan. (2015). “The MovieLens Datasets: History and Context”. *ACM Transactions on Interactive Intelligent Systems*. 5(4): 1–19.

- Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer.
- He, R., W. S. Lee, H. T. Ng, and D. Dahlmeier. (2017a). “An Unsupervised Neural Attention Model for Aspect Extraction”. In: *Proceedings of ACL*. 388–397.
- He, R. and J. McAuley. (2016a). “Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation”. In: *Proceedings of ICDM*. 191–200.
- He, R. and J. McAuley. (2016b). “Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-class Collaborative Filtering”. In: *Proceedings of Web Conference*. 507–517.
- He, W., Y. Dai, M. Yang, J. Sun, F. Huang, L. Si, and Y. Li. (2022a). “Unified Dialog Model Pre-Training for Task-Oriented Dialog Understanding and Generation”. In: *Proceedings of SIGIR*. 187–200.
- He, W., Y. Dai, Y. Zheng, Y. Wu, Z. Cao, D. Liu, P. Jiang, M. Yang, F. Huang, L. Si, *et al.* (2022b). “Galaxy: A Generative Pre-trained Model for Task-oriented Dialog with Semi-supervised Learning and Explicit Policy Injection”. In: *Proceedings of AAAI*. 10749–10757.
- He, X. and T.-S. Chua. (2017). “Neural Factorization Machines for Sparse Predictive Analytics”. In: *Proceedings of SIGIR*. 355–364.
- He, X., K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. (2020). “LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation”. In: *Proceedings of SIGIR*. 639–648.
- He, X., L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. (2017b). “Neural Collaborative Filtering”. In: *Proceedings of Web Conference*. ACM. 173–182.
- He, X., J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, *et al.* (2014). “Practical Lessons from Predicting Clicks on Ads at Facebook”. In: *Proceedings of ADKDD*. 1–9.
- Hearst, M. (2009). *Search User Interfaces*. Cambridge University Press.
- Heck, M., C. van Niekerk, N. Lubis, C. Geishausser, H.-C. Lin, M. Moresi, and M. Gasic. (2020). “TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking”. In: *Proceedings of SIGDIAL*. 35–44.

- Heilman, M. and N. A. Smith. (2010). “Tree Edit Models for Recognizing Textual Entailments, Paraphrases, and Answers to Questions”. In: *Proceedings of NAACL*. 1011–1019.
- Henderson, M., B. Thomson, and J. D. Williams. (2014a). “The Second Dialog State Tracking Challenge”. In: *Proceedings of SIGDIAL*. 263–272.
- Henderson, M., B. Thomson, and J. D. Williams. (2014b). “The Third Dialog State Tracking Challenge”. In: *Proceedings of SLT*. IEEE. 324–329.
- Henderson, M., B. Thomson, and S. Young. (2013). “Deep Neural Network Approach for the Dialog State Tracking Challenge”. In: *Proceedings of SIGDIAL*. 467–471.
- Henderson, P., K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau. (2018). “Ethical Challenges in Data-Driven Dialogue Systems”. In: *Proceedings of 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 123–129.
- Hendriksen, M., E. Kuiper, P. Nauts, S. Schelter, and M. de Rijke. (2020). “Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers”. In: *Proceedings of SIGIR Workshop on eCommerce*. ACM.
- Herlocker, J. L., J. A. Konstan, A. Borchers, and J. Riedl. (1999). “An Algorithmic Framework for Performing Collaborative Filtering”. In: *Proceedings of SIGIR*. 230–237.
- Heuss, M., D. Cohen, M. Mansoury, M. de Rijke, and C. Eickhoff. (2023). “Predictive Uncertainty-based Bias Mitigation in Ranking”. In: *Proceedings of CIKM*. 762–772.
- Hidasi, B., M. Quadrana, A. Karatzoglou, and D. Tikk. (2016). “Parallel Recurrent Neural Network Architectures for Feature-Rich Session-Based Recommendations”. In: *Proceedings of RecSys*. 241–248.
- Hirsch, S., I. Guy, A. Nus, A. Dagan, and O. Kurland. (2020). “Query Reformulation in E-Commerce Search”. In: *Proceedings of SIGIR*. 1319–1328.
- Hisamoto, S., M. Post, and K. Duh. (2020). “Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System?” *Transactions of the Association for Computational Linguistics*. 8(1): 49–63.

- Hofmann, K., L. Li, and F. Radlinski. (2016). “Online Evaluation for Information Retrieval”. *Foundations and Trends in Information Retrieval*. 10(1): 1–117.
- Hofmann, K., S. Whiteson, and M. de Rijke. (2013). “Balancing Exploration and Exploitation in Listwise and Pairwise Online Learning to Rank for Information Retrieval”. *Information Retrieval*. 16(1): 63–90.
- Hollerit, B., M. Kröll, and M. Strohmaier. (2013). “Towards Linking Buyers and Sellers: Detecting Commercial Intent on Twitter”. In: *Proceedings of Web Conference*. 629–632.
- Hosmer Jr, D. W., S. Lemeshow, and R. X. Sturdivant. (2013). *Applied Logistic Regression*. Vol. 398. John Wiley & Sons.
- Hosseini-Asl, E., B. McCann, C.-S. Wu, S. Yavuz, and R. Socher. (2020). “A Simple Language Model for Task-Oriented Dialogue”. In: *Proceedings of NIPS*. 20179–20191.
- Hou, X., Z. Wang, Q. Liu, T. Qu, J. Cheng, and J. Lei. (2023). “Deep Context Interest Network for Click-Through Rate Prediction”. In: *Proceedings of CIKM*. 3948–3952.
- Hou, Y., G. Zhao, C. Liu, Z. Zu, and X. Zhu. (2021). “Conversion Prediction with Delayed Feedback: A Multi-task Learning Approach”. In: *Proceedings ICDM*. 191–199.
- Hsu, C.-C., S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku. (2018). “EmotionLines: An Emotion Corpus of Multi-Party Conversations”. In: *Proceedings of LREC*.
- Hu, B., Z. Lu, H. Li, and Q. Chen. (2014). “Convolutional Neural Network Architectures for Matching Natural Language Sentences”. In: *Proceedings of NIPS*. 2042–2050.
- Hu, B., C. Shi, and J. Liu. (2017). “Playlist Recommendation Based on Reinforcement Learning”. In: *Proceedings of ICIS*.
- Hu, B., C. Shi, W. X. Zhao, and P. S. Yu. (2018a). “Leveraging Meta-Path Based Context for Top-N Recommendation with a Neural Co-Attention Model”. In: *Proceedings of KDD*. 1531–1540.
- Hu, H. and X. He. (2019). “Sets2Sets: Learning from Sequential Sets with Neural Networks”. In: *Proceedings of KDD*. ACM. 1491–1499.

- Hu, H., X. He, J. Gao, and Z.-L. Zhang. (2020a). “Modeling Personalized Item Frequency Information for Next-basket Recommendation”. In: *Proceedings of SIGIR*. 1071–1080.
- Hu, Q., H.-F. Yu, V. Narayanan, I. Davchev, R. Bhagat, and I. S. Dhillon. (2020b). “Query Transformation for Multi-lingual Product Search”. In: *Proceedings of SIGIR*.
- Hu, Y., Q. Da, A. Zeng, Y. Yu, and Y. Xu. (2018b). “Reinforcement Learning to Rank in E-Commerce Search Engine: Formalization, Analysis, and Application”. In: *Proceedings of KDD*. 368–377.
- Hua, K., Z. Feng, C. Tao, R. Yan, and L. Zhang. (2020). “Learning to Detect Relevant Contexts and Knowledge for Response Selection in Retrieval-based Dialogue Systems”. In: *Proceedings of CIKM*. 525–534.
- Huang, C., X. Wu, X. Zhang, C. Zhang, J. Zhao, D. Yin, and N. V. Chawla. (2019). “Online Purchase Prediction via Multi-Scale Modeling of Behavior Dynamics”. In: *Proceedings of KDD*. 2613–2622.
- Huang, C., O. R. Zaiane, A. Trabelsi, and N. Dziri. (2018a). “Automatic Dialogue Generation with Expressed Emotions”. In: *Proceedings of ACL*. 49–54.
- Huang, H., B. Zhao, H. Zhao, Z. Zhuang, Z. Wang, X. Yao, X. Wang, H. Jin, and X. Fu. (2018b). “A Cross-Platform Consumer Behavior Analysis of Large-Scale Mobile Shopping Data”. In: *Proceedings of Web Conference*. 1785–1794.
- Huang, J., H. Oosterhuis, and M. de Rijke. (2022). “It Is Different When Items Are Older: Debiasing Recommendations When Selection Bias and User Preferences Are Dynamic”. In: *Proceedings of WSDM*. 381–389.
- Huang, P.-S., X. He, J. Gao, L. Deng, A. Acero, and L. Heck. (2013). “Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data”. In: *Proceedings of CIKM*. 2333–2338.
- Huang, Y., B. Cui, W. Zhang, J. Jiang, and Y. Xu. (2015). “Tencentrec: Real-time stream recommendation in practice”. In: *Proceedings of SIGMOD*. 227–238.
- Huebner, J., R. M. Frey, C. Ammendola, E. Fleisch, and A. Ilic. (2018). “What People Like in Mobile Finance Apps: An Analysis of User Reviews”. In: *Proceedings of MUM*. 293–304.

- Humeau, S., K. Shuster, M.-A. Lachaux, and J. Weston. (2019). “Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring”. In: *Proceedings of ICLR*.
- Ie, E., V. Jain, J. Wang, S. Narvekar, R. Agarwal, R. Wu, H.-T. Cheng, M. Lustman, V. Gatto, P. Covington, J. McFadden, T. Chandra, and C. Boutilier. (2019). “Reinforcement Learning for Slate-based Recommender Systems: A Tractable Decomposition and Practical Methodology”. *ArXiv*.
- Irene, R. T., C. Borrelli, M. Zanoni, M. Buccoli, and A. Sarti. (2019). “Automatic playlist generation using Convolutional Neural Networks and Recurrent Neural Networks”. In: *Proceedings of EUSIPCO*.
- Izacard, G. and É. Grave. (2021). “Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering”. In: *Proceedings of EACL*. 874–880.
- Jagerman, R., I. Markov, and M. de Rijke. (2019). “When People Change their Mind: Off-policy Evaluation in Non-stationary Recommendation Environments”. In: *Proceedings of WSDM*. 447–455.
- Jain, A., J. Rana, and C. Aggarwal. (2023). “Too Much of Product Information: Don’t Worry, Let’s Look for Evidence!” In: *Proceedings of EMNLP*. 732–738.
- Jannach, D. and M. Ludewig. (2017a). “Investigating Personalized Search in E-Commerce”. In: *Proceedings of FLAIRS*.
- Jannach, D. and M. Ludewig. (2017b). “When Recurrent Neural Networks Meet the Neighborhood for Session-Based Recommendation”. In: *Proceedings of RecSys*. 306–310.
- Jansen, B. J. and P. R. Molina. (2006). “The Effectiveness of Web Search Engines for Retrieving Relevant Ecommerce Links”. *Information Processing & Management*. 42(4): 1075–1098.
- Jin, X., M. Sloan, and J. Wang. (2013). “Interactive Exploratory Search for Multi Page Search Results”. In: *Proceedings of Web Conference*. 655–666.
- Jin, X., W. Lei, Z. Ren, H. Chen, S. Liang, Y. Zhao, and D. Yin. (2018). “Explicit State Tracking with Semi-Supervision for Neural Dialogue Generation”. In: *Proceedings of CIKM*. 1403–1412.

- Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. (2007). “Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search”. *ACM Transactions on Information Systems*. 25(2): 7–es.
- Joshi, M., E. Choi, D. S. Weld, and L. Zettlemoyer. (2017). “TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension”. In: *Proceedings of ACL*. 1601–1611.
- Ju, M., W. Yu, T. Zhao, C. Zhang, and Y. Ye. (2022). “Grape: Knowledge Graph Enhanced Passage Reader for Open domain Question Answering”. In: *Findings of EMNLP*. 169–181.
- Juan, Y., Y. Zhuang, W.-S. Chin, and C.-J. Lin. (2016). “Field-Aware Factorization Machines for CTR Prediction”. In: *Proceedings of RecSys*. 43–50.
- Jung, J., B. Son, and S. Lyu. (2020). “AttnIO: Knowledge Graph Exploration with In-and-Out Attention Flow for Knowledge-Grounded Dialogue”. In: *Proceedings of EMNLP*. 3484–3497.
- Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.
- Jurcicek, F., S. Keizer, M. Gasic, F. Mairesse, B. Thomson, K. Yu, and S. J. Young. (2011). “Real User Evaluation of Spoken Dialogue Systems Using Amazon Mechanical Turk”. In: *Proceedings of INTERSPEECH*.
- Kabbur, S., X. Ning, and G. Karypis. (2013). “FISM: Factored Item Similarity Models for Top-N Recommender Systems”. In: *Proceedings of KDD*. 659–667.
- Kaghazgaran, P., J. Caverlee, and A. Squicciarini. (2018). “Combating Crowdsourced Review Manipulators: A Neighborhood-Based Approach”. In: *Proceedings of WSDM*. 306–314.
- Kamal, A., M. Abulaish, and T. Anwar. (2012). “Mining Feature-Opinion Pairs and Their Reliability Scores from Web Opinion Sources”. In: *Proceedings of WIMS*. 15.
- Kang, W.-C. and J. McAuley. (2018). “Self-Attentive Sequential Recommendation”. In: *Proceedings of ICDM*. IEEE. 197–206.

- Kang, W.-C., J. Ni, N. Mehta, M. Sathiamoorthy, L. Hong, E. Chi, and D. Z. Cheng. (2023). “Do LLMs Understand User Preferences? Evaluating LLMs on User Rating Prediction”. *arXiv preprint arXiv:2305.06474*.
- Kann, K., A. Ebrahimi, J. Koh, S. Dudy, and A. Roncone. (2022). “Open-domain Dialogue Generation: What We Can Do, Cannot Do, And Should Do Next”. In: *Proceedings of ConvAI*. 148–165.
- Kanoje, S., S. Girase, and D. Mukhopadhyay. (2015). “User Profiling Trends, Techniques and Applications”. *arXiv preprint arXiv:1503.07474*.
- Karmaker Santu, S. K., P. Sondhi, and C. Zhai. (2017). “On Application of Learning to Rank for E-Commerce Search”. In: *Proceedings of SIGIR*. 475–484.
- Karpukhin, V., B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. (2020). “Dense Passage Retrieval for Open-Domain Question Answering”. In: *Proceedings of EMNLP*. 6769–6781.
- Karypis, G. (2001). “Evaluation of Item-Based Top-N Recommendation Algorithms”. In: *Proceedings of CIKM*. 247–254.
- Kedia, A., M. A. Zaidi, and H. Lee. (2022). “FiE: Building a Global Probability Space by Leveraging Early Fusion in Encoder for Open-Domain Question Answering”. In: *Proceedings of EMNLP*. 4246–4260.
- Kenter, T., A. Borisov, C. Van Gysel, M. Dehghani, M. de Rijke, and B. Mitra. (2017). “Neural Networks for Information Retrieval (NN4IR)”. In: *Proceedings of SIGIR*.
- Kenton, J. D. M.-W. C. and L. K. Toutanova. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of NAACL-HLT*. 4171–4186.
- Kersbergen, B. and S. Schelter. (2021). “Learnings from a Retail Recommendation System on Billions of Interactions at bol.com”. In: *Proceedings of ICDE*. 2447–2452.
- Kim, B., J. Ahn, and G. Kim. (2020a). “Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue”. In: *Proceedings of ICLR*.

- Kim, E., W. Kim, and Y. Lee. (2003). “Combination of Multiple Classifiers for the Customer’s Purchase Behavior Prediction”. *Decision Support Systems*. 34(2): 167–175.
- Kim, J., M. Won, C. C. S. Liem, and A. Hanjalic. (2018). “Towards Seed-Free Music Playlist Generation: Enhancing Collaborative Filtering with Playlist Title Information”. *Proceedings of ACM Recommender Systems Challenge*.
- Kim, K., E. Kwon, and J. Park. (2021). “Deep User Segment Interest Network Modeling for Click-through Rate Prediction of Online Advertising”. *IEEE Access*. 9: 9812–9821.
- Kim, S.-M., P. Pantel, T. Chklovski, and M. Pennacchiotti. (2006). “Automatically Assessing Review Helpfulness”. In: *Proceedings of EMNLP*. 423–430.
- Kim, S. Y., H. Park, K. Shin, and K.-M. Kim. (2022a). “Ask Me What You Need: Product Retrieval Using Knowledge from GPT-3”. *arXiv preprint arXiv:2207.02516*.
- Kim, S., S. Yang, G. Kim, and S.-W. Lee. (2020b). “Efficient Dialogue State Tracking by Selectively Overwriting Memory”. In: *Proceedings of ACL*. 567–582.
- Kim, T., H. Yoon, Y. Lee, P. Kang, and M. Kim. (2022b). “Mismatch between Multi-turn Dialogue and its Evaluation Metric in Dialogue State Tracking”. In: *Proceedings of ACL*. 297–309.
- Kingma, D. P. and M. Welling. (2014). “Auto-Encoding Variational Bayes”. In: *Proceedings of ICLR*.
- Kingma, D. P. and J. Ba. (2015). “Adam: A Method for Stochastic Optimization”. In: *Proceedings of ICLR*.
- Kipf, T. N. and M. Welling. (2017). “Semi-Supervised Classification with Graph Convolutional Networks”. In: *Proceedings of ICLR*.
- Kiseleva, J., A. Tuzhilin, J. Kamps, M. J. Mueller, L. Bernardi, C. Davis, I. Kovacek, M. S. Einarsen, and D. Hiemstra. (2016). “Beyond Movie Recommendations: Solving the Continuous Cold Start Problem in E-commerce Recommendations”. *arXiv preprint arXiv:1607.07904*.
- Kleinbaum, D. G., K. Dietz, M. Gail, M. Klein, and M. Klein. (2002). *Logistic Regression*. Springer.

- Konstan, J. A., B. N. Miller, D. A. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. (1997). “GroupLens: Applying Collaborative Filtering to Usenet News”. *Communication of the ACM*. 40(3): 77–87.
- Kooti, F., K. Lerman, L. M. Aiello, M. Grbovic, N. Djuric, and V. Radosavljevic. (2016). “Portrait of an Online Shopper: Understanding and Predicting Consumer Behavior”. In: *Proceedings of WSDM*. 205–214.
- Koren, Y. (2008). “Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model”. In: *Proceedings of KDD*. 426–434.
- Koren, Y. (2009). “Collaborative Filtering with Temporal Dynamics”. In: *Proceedings of KDD*. 447–456.
- Koren, Y., R. Bell, and C. Volinsky. (2009). “Matrix Factorization Techniques for Recommender Systems”. *Computer*. 42(8): 30–37.
- Krishnan, S., J. Patel, M. J. Franklin, and K. Goldberg. (2014). “A Methodology for Learning, Analyzing, and Mitigating Social Influence Bias in Recommender Systems”. In: *Proceedings of RecSys*. 137–144.
- Kuflik, T. and P. Shoval. (2000). “Generation of User Profiles for Information Filtering — Research Agenda”. In: *Proceedings of SIGIR*. 313–315.
- Kumar, A., P. Ku, A. Goyal, A. Metallinou, and D. Hakkani-Tur. (2020). “MA-DST: Multi-Attention-Based Scalable Dialog State Tracking”. In: *Proceedings of AAAI*. Vol. 34. No. 05. 8107–8114.
- Laenen, K., S. Zoghbi, and M.-F. Moens. (2018). “Web Search of Fashion Items with Multimodal Querying”. In: *Proceedings of WSDM*. 342–350.
- Lalmas, M. and L. Hong. (2018). “Tutorial on Metrics of User Engagement: Applications to News, Search and E-Commerce”. In: *Proceedings of WSDM*. 781–782.
- Lalmas, M., J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. (2015). “Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users”. In: *Proceedings of KDD*. 1929–1938.
- Lamel, L., S. Rosset, J.-L. Gauvain, S. Bannacef, M. Garnier-Rizet, and B. Prouts. (2000). “The LIMSI ARISE system”. *Speech Communication*. 31: 339–353.

- Le, D., H. W. Lauw, and Y. Fang. (2019). “Correlation-Sensitive Next-Basket Recommendation”. In: *Proceedings of IJCAI*. ijcai.org. 2808–2814.
- Le, Y., X. Li, S. Wang, P. Wang, H. Lin, and G. Jiang. (2018). “CVTE SLU: A Hybrid System for Command Understanding Task Oriented to the Music Field”. In: *Proceedings of CCKS*. 13–18.
- Ledford, J. L. (2015). *Search Engine Optimization Bible*. Vol. 584. John Wiley & Sons.
- Lee, C.-H., H. Cheng, and M. Ostendorf. (2021). “Dialogue State Tracking with a Language Model using Schema-Driven Prompting”. In: *Proceedings of EMNLP*. 4937–4949.
- Lee, J., M. Podlaseck, E. Schonberg, and R. Hoch. (2001). “Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising”. *Data Mining and Knowledge Discovery*. 5(1-2): 59–84.
- Lee, K., M.-W. Chang, and K. Toutanova. (2019). “Latent Retrieval for Weakly Supervised Open Domain Question Answering”. In: *Proceedings of ACL*. 6086–6096.
- Lee, K.-c., B. Orten, A. Dasdan, and W. Li. (2012). “Estimating Conversion Rate in Display Advertising from Past Performance Data”. In: *Proceedings of KDD*. 768–776.
- Lee, S. (2013). “Structured Discriminative Model For Dialog State Tracking”. In: *Proceedings of SIGDIAL*. 442–451.
- Lee, S. and M. Eskenazi. (2013). “Recipe For Building Robust Spoken Dialog State Trackers: Dialog State Tracking Challenge System Description”. In: *Proceedings of SIGDIAL*. 414–422.
- Leem, B. and H. Chun. (2014). “An Impact of Online Recommendation Network on Demand”. *Expert Systems with Applications*. 41(4): 1723–1729.
- Lehmann, J., M. Lalmas, E. Yom-Tov, and G. Dupret. (2012). “Models of User Engagement”. In: *Proceedings of UMAP*. 164–175.
- Lei, W., X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, and T.-S. Chua. (2020a). “Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems”. In: *Proceedings of WSDM*. 304–312.

- Lei, W., X. Jin, M.-Y. Kan, Z. Ren, X. He, and D. Yin. (2018). “Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures”. In: *Proceedings of ACL*. 1437–1447.
- Lei, W., G. Zhang, X. He, Y. Miao, X. Wang, L. Chen, and T.-S. Chua. (2020b). “Interactive Path Reasoning on Graph for Conversational Recommendation”. In: *Proceedings of KDD*. 2073–2083.
- Leng, Y., L. Yu, J. Xiong, and G. Xu. (2020). “Recurrent Convolution Basket Map for Diversity Next-Basket Recommendation”. In: *Proceedings of DASFAA*. Vol. 12114. *Lecture Notes in Computer Science*. Springer. 638–653.
- Lewis, P., P. Stenetorp, and S. Riedel. (2021). “Question and Answer Test-train Overlap in Open-domain Question Answering Datasets”. In: *Proceedings of ACL*. 1000–1008.
- Li, B., A. Ghose, and P. G. Ipeirotis. (2011). “Towards a Theory Model for Product Search”. In: *Proceedings of Web Conference*. 327–336.
- Li, C., Z. Liu, M. Wu, Y. Xu, H. Zhao, P. Huang, G. Kang, Q. Chen, W. Li, and D. L. Lee. (2019a). “Multi-Interest Network with Dynamic Routing for Recommendation at Tmall”. In: *Proceedings of CIKM*. ACM. 2615–2623.
- Li, F., Y. Tang, M. Huang, and X. Zhu. (2009). “Answering Opinion Questions with Random Walks on Graphs”. In: *Proceedings of ACL*. 737–745.
- Li, H., F. Pan, X. Ao, Z. Yang, M. Lu, J. Pan, D. Liu, L. Xiao, and Q. He. (2021a). “Follow the Prophet: Accurate Online Conversion Rate Prediction in the Face of Delayed Feedback”. In: *Proceedings of SIGIR*. 1915–1919.
- Li, J., C. Tao, H. Hu, C. Xu, Y. Chen, and D. Jiang. (2022a). “Unsupervised Cross-Domain Adaptation for Response Selection Using Self-Supervised and Adversarial Training”. In: *Proceedings of WSDM*. 562–570.
- Li, J., H. Liu, C. Gui, J. Chen, Z. Ni, N. Wang, and Y. Chen. (2018a). “The Design and Implementation of a Real Time Visual Search System on JD E-commerce Platform”. In: *Proceedings of IMCI*. 9–16.

- Li, J., P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma. (2017a). “Neural Attentive Session-based Recommendation”. In: *Proceedings of CIKM*. 1419–1428.
- Li, J. and X. Sun. (2018). “A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation”. In: *Proceedings of EMNLP*. 678–683.
- Li, J., M. Galley, C. Brockett, J. Gao, and B. Dolan. (2016a). “A Diversity-Promoting Objective Function for Neural Conversation Models”. In: *Proceedings of NAACL*. 110–119.
- Li, J., M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. (2016b). “A Persona-Based Neural Conversation Model”. In: *Proceedings of ACL*. 994–1003.
- Li, J., Z. Dou, Y. Zhu, X. Zuo, and J.-R. Wen. (2020a). “Deep Cross-platform Product Matching in E-commerce”. *Information Retrieval Journal*. 23(2): 136–158.
- Li, J., H. Yang, and C. Zong. (2018b). “Document-level Multi-aspect Sentiment Classification by Jointly Modeling Users, Aspects, and Overall Ratings”. In: *Proceedings of COLING*. 925–936.
- Li, J., T. Tang, W. X. Zhao, and J.-R. Wen. (2021b). “Pretrained Language Models for Text Generation: A Survey”. In: *Proceedings of IJCAI*. 4492–4499.
- Li, L., J. Y. Kim, and I. Zitouni. (2015). “Toward Predicting the Outcome of an A/B Experiment for Search Relevance”. In: *Proceedings of WSDM*. 37–46.
- Li, L., C. Xu, W. Wu, Y. Zhao, X. Zhao, and C. Tao. (2020b). “Zero-Resource Knowledge-Grounded Dialogue Generation”. In: *Proceedings of NIPS*.
- Li, M., M. Ariannezhad, A. Yates, and M. de Rijke. (2023a). “Who Will Purchase this Item Next? Reverse Next Period Recommendation in Grocery Shopping”. *ACM Transactions on Recommender Systems*.
- Li, M., S. Jullien, M. Ariannezhad, and M. de Rijke. (2023b). “A Next Basket Recommendation Reality Check”. *ACM Transactions on Information Systems*. 41(4): Article 116.
- Li, P., Z. Wang, L. Bing, and W. Lam. (2019b). “Persona-Aware Tips Generation?” In: *Proceedings of Web Conference*. 1006–1016.

- Li, P., Z. Wang, Z. Ren, L. Bing, and W. Lam. (2017b). “Neural Rating Regression with Abstractive Tips Generation for Recommendation”. In: *Proceedings of SIGIR*. 345–354.
- Li, Q., H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen. (2020c). “EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation”. In: *Proceedings of COLING*. 4454–4466.
- Li, Q., P. Li, Z. Chen, and Z. Ren. (2022b). “Empathetic Dialogue Generation via Knowledge Enhancing and Emotion Dependency Modeling”. In: *Proceedings of AAAI*.
- Li, Q., P. Li, X. Li, Z. Ren, Z. Chen, and M. de Rijke. (2021c). “Abstractive Opinion Tagging”. In: *Proceedings of WSDM*. 337–345.
- Li, Q., P. Li, Z. Ren, P. Ren, and Z. Chen. (2022c). “Knowledge Bridging for Empathetic Dialogue Generation”. In: *Proceedings of AAAI*. 10993–11001.
- Li, R., S. E. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal. (2018c). “Towards Deep Conversational Recommendations”. In: *Proceedings of NIPS*. 9748–9758.
- Li, R., Y. Jiang, W. Yang, G. Tang, S. Wang, C. Ma, W. He, X. Xiong, Y. Xiao, and E. Y. Zhao. (2019c). “From Semantic Retrieval to Pairwise Ranking: Applying Deep Learning in E-Commerce Search”. In: *Proceedings of SIGIR*. 1383–1384.
- Li, S., W. Lei, Q. Wu, X. He, P. Jiang, and T.-S. Chua. (2021d). “Seamlessly Unifying Attributes and Items: Conversational Recommendation for Cold-Start Users”. *ACM Transactions on Information Systems*. 39(4): 1–29.
- Li, X., C. Wu, and F. Mai. (2018d). “The Effect of Online Reviews on Product Sales: A Joint Sentiment-topic Analysis”. *Information & Management*. 56(2): 172–184.
- Li, X. and D. Roth. (2002). “Learning Question Classifiers”. In: *Proceedings of COLING*. 1–7.
- Li, X., Q. Li, W. Wu, and Q. Yin. (2021e). “Generation and Extraction Combined Dialogue State Tracking with Hierarchical Ontology Integration”. In: *Proceedings of EMNLP*. 2241–2249.
- Li, Y., H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. (2017c). “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”. In: *Proceedings of IJCNLP*. 986–995.

- Li, Y., N. Yang, L. Wang, F. Wei, and W. Li. (2023c). “Multiview Identifiers Enhanced Generative Retrieval”. *arXiv preprint arXiv:2305.16675*.
- Li, Y., N. Yang, L. Wang, F. Wei, and W. Li. (2024). “Learning to Rank in Generative Retrieval”. In: *Proceedings of AAAI*. Vol. 38. No. 8. 8716–8723.
- Li, Z., C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou. (2019d). “Incremental Transformer with Deliberation Decoder for Document Grounded Conversations”. In: *Proceedings of ACL*. 12–21.
- Lian, J., I. Batal, Z. Liu, A. Soni, E. Y. Kang, Y. Wang, and X. Xie. (2021). “Multi-Interest-Aware User Modeling for Large-Scale Sequential Recommendations”. *arXiv preprint arXiv:2102.09211*.
- Lian, J., X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun. (2018). “xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems”. In: *Proceedings of KDD*. 1754–1763.
- Lian, R., M. Xie, F. Wang, J. Peng, and H. Wu. (2019). “Learning to Select Knowledge for Response Generation in Dialog Systems”. In: *Proceedings of IJCAI*.
- Liang, D., L. Charlin, J. McInerney, and D. M. Blei. (2016). “Modeling User Exposure in Recommendation”. In: *Proceedings of Web Conference*. 951–961.
- Liang, D., R. G. Krishnan, M. D. Hoffman, and T. Jebara. (2018a). “Variational Autoencoders for Collaborative Filtering”. In: *Proceedings of Web Conference*. 689–698.
- Liang, S., Z. Ren, Y. Zhao, J. Ma, E. Yilmaz, and M. de Rijke. (2017). “Inferring Dynamic User Interests in Streams of Short Texts for User Clustering”. *ACM Transactions on Information Systems*. 36(1): 1–37.
- Liang, S., X. Zhang, Z. Ren, and E. Kanoulas. (2018b). “Dynamic Embeddings for User Profiling in Twitter”. In: *Proceedings of KDD*. 1764–1773.
- Liang, T.-P., X. Li, C.-T. Yang, and M. Wang. (2015). “What in Consumer Reviews Affects the Sales of Mobile Apps: A Multifacet Sentiment Analysis Approach”. *International Journal of Electronic Commerce*. 20(2): 236–260.

- Liang, W., Y. Tian, C. Chen, and Z. Yu. (2020). “MOSS: End-to-End Dialog System Framework with Modular Supervision”. In: *Proceedings of AAAI*. 8327–8335.
- Liao, L., X. He, B. Zhao, C.-W. Ngo, and T.-S. Chua. (2018). “Interpretable Multimodal Retrieval for Fashion Products”. In: *Proceedings of MM*. 1571–1579.
- Liao, L., T. Zhu, L. H. Long, and T. S. Chua. (2021). “Multi-Domain Dialogue State Tracking with Recursive Inference”. In: *Proceedings of Web Conference*. 2568–2577.
- Liebman, E., M. Saar-Tsechansky, and P. Stone. (2015). “DJ-MC: A Reinforcement-Learning Agent for Music Playlist Recommendation”. In: *Proceedings of AAMAS*. ACM. 591–599.
- Lin, J. (2006). “The Role of Information Retrieval in Answering Complex Questions”. In: *Proceedings of COLING*. 523–530.
- Lin, J., R. Nogueira, and A. Yates. (2021a). “Pretrained Transformers for Text Ranking: BERT and Beyond”. *Synthesis Lectures on Human Language Technologies*. 14(4): 1–325.
- Lin, P., Y. Zou, L. Wu, M. Ma, Z. Ding, and B. Long. (2022). “Automatic Scene-based Topic Channel Construction System for E-Commerce”. In: *Proceedings of EMNLP*.
- Lin, T.-H., T.-C. Chi, and A. Rumshisky. (2021b). “Domain-Adaptive Pretraining Methods for Dialogue Understanding”. In: *Proceedings of ACL-IJCNLP*. 665–669.
- Lin, W., B.-H. Tseng, and B. Byrne. (2021c). “Knowledge-Aware Graph-Enhanced GPT-2 for Dialogue State Tracking”. In: *Proceedings of EMNLP*. 7871–7881.
- Lin, X., W. Jian, J. He, T. Wang, and W. Chu. (2020a). “Generating Informative Conversational Response using Recurrent Knowledge-Interaction and Knowledge-Copy”. In: *Proceedings of ACL*. 41–52.
- Lin, Y., H. Ji, Z. Liu, and M. Sun. (2018). “Denoising Distantly Supervised Open-Domain Question Answering”. In: *Proceedings of ACL*. 1736–1745.
- Lin, Z., B. Liu, A. Madotto, S. Moon, P. Crook, Z. Zhou, Z. Wang, Z. Yu, E. Cho, R. Subba, *et al.* (2021d). “Zero-shot Dialogue State Tracking via Cross-task Transfer”. In: *Proceedings of EMNLP*. 7890–7900.

- Lin, Z., B. Liu, S. Moon, P. A. Crook, Z. Zhou, Z. Wang, Z. Yu, A. Madotto, E. Cho, and R. Subba. (2021e). “Leveraging Slot Descriptions for Zero-Shot Cross-Domain Dialogue State Tracking”. In: *Proceedings of NAACL-HLT*. 5640–5648.
- Lin, Z., A. Madotto, J. Shin, P. Xu, and P. Fung. (2019). “MoEL: Mixture of Empathetic Listeners”. In: *Proceedings of EMNLP-IJCNLP*. 121–132.
- Lin, Z., P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung. (2020b). “CAiRE: An End-to-End Empathetic Chatbot”. In: *Proceedings of AACL*. 13622–13623.
- Liu, B., N. Xue, H. Guo, R. Tang, S. Zafeiriou, X. He, and Z. Li. (2020a). “AutoGroup: Automatic Feature Grouping for Modelling Explicit High-Order Feature Interactions in CTR Prediction”. In: *Proceedings of SIGIR*. 199–208.
- Liu, B., C. Zhu, G. Li, W. Zhang, J. Lai, R. Tang, X. He, Z. Li, and Y. Yu. (2020b). “AutoFIS: Automatic Feature Interaction Selection in Factorization Models for Click-Through Rate Prediction”. In: *Proceedings of KDD*. 2636–2645.
- Liu, C.-L., W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou. (2012). “Movie Rating and Review Summarization in Mobile Environment”. *IEEE Transactions on Systems, Man, and Cybernetics*. 42(3): 397–407.
- Liu, G., T. T. Nguyen, G. Zhao, W. Zha, J. Yang, J. Cao, M. Wu, P. Zhao, and W. Chen. (2016a). “Repeat Buyer Prediction for E-Commerce”. In: *Proceedings of KDD*. 155–164.
- Liu, H., Y. Gao, P. Lv, M. Li, S. Geng, M. Li, and H. Wang. (2017). “Using Argument-based Features to Predict and Analyse Review Helpfulness”. In: *Proceedings of EMNLP*. 1358–1363.
- Liu, H., M. Chen, Y. Wu, X. He, and B. Zhou. (2021a). “Conversational Query Rewriting with Self-Supervised Learning”. In: *Proceedings of ICASSP*. IEEE. 7628–7632.
- Liu, H., J. Jia, W. Qu, and N. Z. Gong. (2021b). “EncoderMI: Membership inference against pre-trained encoders in contrastive learning”. In: *Proceedings of CCS*. 2081–2095.
- Liu, J., P. Dolan, and E. R. Pedersen. (2010). “Personalized News Recommendation Based on Click Behavior”. In: *Proceedings of IUI*. 31–40.

- Liu, J., Z. Dou, Q. Zhu, and J.-R. Wen. (2022a). “A Category-aware Multi-interest Model for Personalized Product Search”. In: *Proceedings of Web Conference*. 360–368.
- Liu, J., Z. Hai, M. Yang, and L. Bing. (2021c). “Multi-perspective Coherent Reasoning for Helpfulness Prediction of Multimodal Reviews”. In: *Proceedings of ACL*. 5927–5936.
- Liu, M., X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua. (2018a). “Attentive Moment Retrieval in Videos”. In: *Proceedings of SIGIR*. 15–24.
- Liu, M., Y. Fang, D. H. Park, X. Hu, and Z. Yu. (2016b). “Retrieving Non-Redundant Questions to Summarize a Product Review”. In: *Proceedings of SIGIR*. 385–394.
- Liu, P., L. Zhang, and J. A. Gulla. (2023a). “Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems”. *Transactions of the Association for Computational Linguistics*. 11: 1553–1571.
- Liu, Q., P. Cao, C. Liu, J. Chen, X. Cai, F. Yang, S. He, K. Liu, and J. Zhao. (2021d). “Domain-Lifelong Learning for Dialogue State Tracking via Knowledge Preservation Networks”. In: *Proceedings of EMNLP*. 2301–2311.
- Liu, S., W. Gu, G. Cong, and F. Zhang. (2020c). “Structural Relationship Representation Learning with Graph Embedding for Personalized Product Search”. In: *Proceedings of CIKM*. 915–924.
- Liu, S., L. Li, J. Song, Y. Yang, and X. Zeng. (2023b). “Multimodal Pre-training with Self-distillation for Product Understanding in E-commerce”. In: *Proceedings of WSDM*. 1039–1047.
- Liu, S., H. Chen, Z. Ren, Y. Feng, Q. Liu, and D. Yin. (2018b). “Knowledge Diffusion for Neural Dialogue Generation”. In: *Proceedings of ACL*. 1489–1498.
- Liu, X., Z. Li, Y. Gao, J. Yang, T. Cao, Z. Wang, B. Yin, and Y. Song. (2023c). “Enhancing User Intent Capture in Session-Based Recommendation with Attribute Patterns”. *arXiv preprint arXiv:2312.16199*.

- Liu, X., W. Guan, L. Li, H. Li, C. Lin, X. Li, S. Chen, J. Xu, H. Deng, and B. Zheng. (2022b). “Pretraining Representations of Multi-modal Multi-query E-commerce Search”. In: *Proceedings of KDD*. 3429–3437.
- Liu, Y. and M. Lapata. (2019). “Hierarchical Transformers for Multi-Document Summarization”. In: *Proceedings of ACL*. 5070–5081.
- Liu, Y., Y. Gu, Z. Ding, J. Gao, Z. Guo, Y. Bao, and W. Yan. (2020d). “Decoupled Graph Convolution Network for Inferring Substitutable and Complementary Items”. In: *Proceedings of CIKM*. 2621–2628.
- Liu, Y., Y. Xiao, Q. Wu, C. Miao, J. Zhang, B. Zhao, and H. Tang. (2020e). “Diversified Interactive Recommendation with Implicit Feedback”. In: *Proceedings of AAAI*. 4932–4939.
- Liu, Y., M. Li, X. Li, F. Giunchiglia, X. Feng, and R. Guan. (2022c). “Few-Shot Node Classification on Attributed Networks with Graph Meta-Learning”. In: *Proceedings of SIGIR*. 471–481.
- Liu, Y., Z. Ren, W.-N. Zhang, W. Che, T. Liu, and D. Yin. (2020f). “Keywords Generation Improves E-Commerce Session-Based Recommendation”. In: *Proceedings of Web Conference*. 1604–1614.
- Liu, Y., H. Lee, P. Achananuparp, E.-P. Lim, T.-L. Cheng, and S.-D. Lin. (2019a). “Characterizing and Predicting Repeat Food Consumption Behavior for Just-in-Time Interventions”. In: *Proceedings of DPH*. ACM. 11–20.
- Liu, Z., H. Wang, Z.-Y. Niu, H. Wu, W. Che, and T. Liu. (2020g). “Towards Conversational Recommendation over Multi-Type Dialogs”. *arXiv preprint arXiv:2005.03954*.
- Liu, Z., W. Zhang, Y. Chen, W. Sun, T. Du, and B. Schroeder. (2022d). “Towards Generalizeable Semantic Product Search by Text Similarity Pre-training on Search Click Logs”. In: *Proceedings of Fifth Workshop on e-Commerce and NLP*. 224–233.
- Liu, Z., Z.-Y. Niu, H. Wu, and H. Wang. (2019b). “Knowledge Aware Conversation Generation with Reasoning on Augmented Graph”. In: *Proceedings of EMNLP*. 1782–1792.
- Liu, Z., Y. Dou, P. S. Yu, Y. Deng, and H. Peng. (2020h). “Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection”. In: *Proceedings of SIGIR*. 1569–1572.

- Liu, Z., P. Ren, Z. Chen, Z. Ren, M. de Rijke, and M. Zhou. (2021e). “Learning to Ask Conversational Questions by Optimizing Levenshtein Distance”. In: *Proceedings of ACL*. 5638–5650.
- Liu, Z., M. A. Patwary, R. J. Prenger, S. Prabhunoye, W. Ping, M. Shoeybi, and B. Catanzaro. (2022e). “Multi-Stage Prompting for Knowledgeable Dialogue Generation”. In: *Proceedings of ACL*. 1317–1337.
- Lo, C., D. Frankowski, and J. Leskovec. (2016). “Understanding Behaviors That Lead to Purchasing: A Case Study of Pinterest”. In: *Proceedings of KDD*. 531–540.
- Lu, H., Y. Hu, T. Zhao, T. Wu, Y. Song, and B. Yin. (2021). “Graph-based Multilingual Product Retrieval in E-Commerce Search”. In: *Proceedings of NAACL-HLT*. 146–153.
- Lu, H., M. Zhang, and S. Ma. (2018). “Between Clicks and Satisfaction: Study on Multi-phase User Preferences and Satisfaction for Online News Reading”. In: *Proceedings of SIGIR*. 435–444.
- Lu, J., G. Pergola, L. Gui, B. Li, and Y. He. (2020). “CHIME: Cross-passage Hierarchical Memory Network for Generative Review Question Answering”. *arXiv preprint arXiv:2011.00519*.
- Lu, Q., S. Pan, L. Wang, J. Pan, F. Wan, and H. Yang. (2017). “A Practical Framework of Conversion Rate Prediction for Online Display Advertising”. In: *Proceedings of ADKDD*. 1–9.
- Lu, Y., H. He, Q. Peng, W. Meng, and C. Yu. (2006). “Clustering E-commerce Search Engines based on their Search Interface Pages using WISE-Cluster”. *Data & Knowledge Engineering*. 59(2): 231–246.
- Lucic, A., M. Srikumar, U. Bhatt, A. Xiang, A. Taly, Q. V. Liao, and M. de Rijke. (2021). “A Multistakeholder Approach Towards Evaluating AI Transparency Mechanisms”. In: *ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI*. ACM.
- Luo, K., H. Yang, G. Wu, and S. Sanner. (2020). “Deep Critiquing for VAE-based Recommender Systems”. In: *Proceedings of SIGIR*. 1269–1278.

- Luo, X., Y. Wu, and X.-S. Xu. (2018). “Scalable Supervised Discrete Hashing for Large-Scale Search”. In: *Proceedings of Web Conference*. 1603–1612.
- Ly, D. K., K. Sugiyama, Z. Lin, and M.-Y. Kan. (2011). “Product Review Summarization from a Deeper Perspective”. In: *Proceedings of JCDL*. 311–314.
- Ma, L., W.-N. Zhang, M. Li, and T. Liu. (2020a). “A Survey of Document Grounded Dialogue Systems (DGDS)”. *arXiv preprint arXiv:2004.13818*.
- Ma, L., W. Zhang, R. Sun, and T. Liu. (2020b). “A Compare Aggregate Transformer for Understanding Document-grounded Dialogue”. In: *Proceedings of EMNLP*. 1358–1367.
- Ma, W., R. Takanobu, and M. Huang. (2021). “CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation”. In: *Proceedings of EMNLP*. 1839–1851.
- Ma, X., L. Zhao, G. Huang, Z. Wang, Z. Hu, X. Zhu, and K. Gai. (2018). “Entire Space Multi-task Model: An Effective Approach for Estimating Post-click Conversion Rate”. In: *Proceedings of SIGIR*. 1137–1140.
- Ma, Y., Z. Guo, Z. Ren, J. Tang, and D. Yin. (2020c). “Streaming Graph Neural Networks”. In: *Proceedings of SIGIR*. 719–728.
- Madotto, A., S. Cahyawijaya, G. I. Winata, Y. Xu, Z. Liu, Z. Lin, and P. Fung. (2020). “Learning Knowledge Bases with Parameters for Task-Oriented Dialogue Systems”. In: *Proceedings of EMNLP*. 2372–2394.
- Madotto, A., C.-S. Wu, and P. Fung. (2018). “Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems”. In: *Proceedings of ACL*. 1468–1478.
- Magnani, A., F. Liu, S. Chaidaroon, S. Yadav, P. Reddy Suram, A. Puthenpuhussery, S. Chen, M. Xie, A. Kashi, T. Lee, *et al.* (2022). “Semantic Retrieval at Walmart”. In: *Proceedings of KDD*. 3495–3503.
- Magnani, A., F. Liu, M. Xie, and S. Banerjee. (2019). “Neural Product Retrieval at Walmart.Com”. In: *Proceedings of Web Conference*. 367–372.

- Majumder, N., P. Hong, S. Peng, J. Lu, D. Ghosal, A. F. Gelbukh, R. Mihalcea, and S. Poria. (2020). “MIME: MIMicking Emotions for Empathetic Response Generation”. In: *Proceedings of EMNLP*. 8968–8979.
- Malone, T. W., K. R. Grant, F. A. Turbak, S. A. Brobst, and M. D. Cohen. (1987). “Intelligent Information Sharing Systems”. *Communications of the ACM*. 30(5): 390–402.
- Mangili, F., D. Brogini, A. Antonucci, M. Alberti, and L. Cimasoni. (2020). “A Bayesian Approach to Conversational Recommendation Systems”. *AAAI 2020 Workshop on Interactive and Conversational Recommendation Systems*.
- Manning, C. D., P. Raghavan, and H. Schütze. (2008). *Introduction to Information Retrieval*. Vol. 39. Cambridge University Press Cambridge.
- Mao, J., Y. Liu, M. Zhang, and S. Ma. (2014). “Estimating Credibility of User Clicks with Mouse Movement and Eye-Tracking Information”. In: *Proceedings of SIGIR*. 263–274.
- Mao, J., W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. (2015). “Deep Captioning with Multimodal Recurrent Neural Networks”. In: *Proceedings of ICLR*.
- Mao, Y., P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen. (2021). “Generation-augmented Retrieval for Open-domain Question Answering”. In: *Proceedings of ACL*. 4089–4100.
- Marlin, B. M., R. S. Zemel, S. Roweis, and M. Slaney. (2007). “Collaborative Filtering and the Missing at Random Assumption”. In: *Proceedings of UAI*. 267–275.
- Martin, L. and P. Pu. (2014). “Prediction of Helpful Reviews Using Emotions Extraction”. In: *Proceedings of AAAI*. 1551–1557.
- Mathur, A., N. D. Lane, and F. Kawsar. (2016). “Engagement-Aware Computing: Modelling User Engagement from Mobile Contexts”. In: *Proceedings of UbiComp*. 622–633.
- McAuley, J., C. Targett, Q. Shi, and A. Van Den Hengel. (2015). “Image-Based Recommendations on Styles and Substitutes”. In: *Proceedings of SIGIR*. 43–52.

- McAuley, J. and A. Yang. (2016). “Addressing Complex and Subjective Product-Related Queries with Customer Reviews”. In: *Proceedings of Web Conference*. 625–635.
- McDuff, D., R. El Kaliouby, J. F. Cohn, and R. W. Picard. (2015). “Predicting Ad Liking and Purchase Intent: Large-Scale Analysis of Facial Responses to Ads”. *IEEE Transactions on Affective Computing*. 6(3): 223–235.
- McLaughlin, M. R. and J. L. Herlocker. (2004). “A Collaborative Filtering Algorithm and Evaluation Metric That Accurately Model the User Experience”. In: *Proceedings of SIGIR*. 329–336.
- Mehrotra, R., M. Lalmas, D. Kenney, T. Lim-Meng, and G. Hashemian. (2019). “Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations”. In: *Proceedings of Web Conference*. 1256–1267.
- Meng, C., P. Ren, Z. Chen, C. Monz, J. Ma, and M. de Rijke. (2020a). “RefNet: A Reference-aware Network for Background Based Conversation”. In: *Proceedings of AAAI*. 8496–8503.
- Meng, C., P. Ren, Z. Chen, Z. Ren, T. Xi, and M. de Rijke. (2021a). “Initiative-Aware Self-Supervised Learning for Knowledge-Grounded Conversations”. In: *Proceedings of SIGIR*. 522–532.
- Meng, C., P. Ren, Z. Chen, W. Sun, Z. Ren, Z. Tu, and M. de Rijke. (2020b). “DukeNet: A Dual Knowledge Interaction Network for Knowledge-Grounded Conversation”. In: *Proceedings of SIGIR*. 1151–1160.
- Meng, W., D. Yang, and Y. Xiao. (2020c). “Incorporating User Micro-Behaviors and Item Knowledge into Multi-Task Learning for Session-Based Recommendation”. In: *Proceedings of SIGIR*. 1091–1100.
- Meng, Z., J. Zhang, Y. Li, J. Li, T. Zhu, and L. Sun. (2021b). “A General Method For Automatic Discovery of Powerful Interactions In Click-Through Rate Prediction”. In: *Proceedings of SIGIR*. 1298–1307.
- Mesnil, G., X. He, L. Deng, and Y. Bengio. (2013). “Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding”. In: *Proceedings of Interspeech*. 3771–3775.

- Miao, L., D. Cao, J. Li, and W. Guan. (2020). “Multi-modal Product Title Compression”. *Information Processing & Management*. 57(1): 102123.
- Mihaylova, T., G. Karadzhov, P. Atanasova, R. Baly, M. Mohtarami, and P. Nakov. (2019). “SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums”. In: *Proceedings of 13th International Workshop on Semantic Evaluation*. 860–869.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean. (2013). “Efficient Estimation of Word Representations in Vector Space”. In: *Proceedings of ICLR*.
- Mishne, G. and M. de Rijke. (2006a). “Deriving Wishlists from Blogs: Show us your Blog, and We’ll Tell you What Books to Buy”. In: *Proceedings of Web Conference*. 925–926.
- Mishne, G. and M. de Rijke. (2006b). “Deriving Wishlists from Blogs: Show us your Blog, and We’ll Tell you What Books to Buy”. In: *Proceedings of Web Conference*. ACM.
- Mislove, A., B. Viswanath, K. P. Gummadi, and P. Druschel. (2010). “You Are Who You Know: Inferring User Profiles in Online Social Networks”. In: *Proceedings of WSDM*. 251–260.
- Mitra, B. and N. Craswell. (2017). “An Introduction to Neural Information Retrieval”. *Foundations and Trends in Information Retrieval*. 13: 1–126.
- Mitra, B., F. Diaz, and N. Craswell. (2017). “Learning to Match Using Local and Distributed Representations of Text for Web Search”. In: *Proceedings of Web Conference*. 1291–1299.
- Mittal, H., A. Chakrabarti, B. Bayar, A. A. Sharma, and N. Rasiwasia. (2021). “Distantly Supervised Transformers For E-Commerce Product QA”. *arXiv preprint arXiv:2104.02947*.
- Mnih, V., N. Heess, A. Graves, *et al.* (2014). “Recurrent Models of Visual Attention”. In: *Proceedings of NIPS*. 2204–2212.
- Moghaddam, S. and M. Ester. (2011). “AQA: Aspect-based Opinion Question Answering”. In: *Proceedings of ICDMW*. 89–96.
- Moghe, N., S. Arora, S. Banerjee, and M. M. Khapra. (2018). “Towards Exploiting Background Knowledge for Building Conversation Systems”. In: *Proceedings of EMNLP*. 2322–2332.

- Monti, D., E. Palumbo, G. Rizzo, P. Lisena, R. Troncy, M. Fell, E. Cabrio, and M. Morisio. (2018). “An Ensemble Approach of Recurrent Neural Networks using Pre-Trained Embeddings for Playlist Completion”. *Proceedings of ACM Recommender Systems Challenge 2018*.
- Moon, S., P. Shah, A. Kumar, and R. Subba. (2019). “OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs”. In: *Proceedings of ACL*. 845–854.
- Morik, M., A. Singh, J. Hong, and T. Joachims. (2020). “Controlling Fairness and Bias in Dynamic Learning-to-Rank”. In: *Proceedings of SIGIR*. 429–438.
- Mrkšić, N., D. Ó. Séaghdha, B. Thomson, M. Gasic, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young. (2015). “Multi-domain Dialog State Tracking using Recurrent Neural Networks”. In: *Proceedings of ACL*. 794–799.
- Mrkšić, N., D. Ó. Séaghdha, T.-H. Wen, B. Thomson, and S. Young. (2017). “Neural Belief Tracker: Data-Driven Dialogue State Tracking”. In: *Proceedings of ACL*. 1777–1788.
- Mukherjee, S. (2021). “Unsupervised Meta Learning for One Shot Title Compression in Voice Commerce”. *arXiv preprint arXiv:2102.10760*.
- Mukherjee, S., P. Sayapaneni, and S. Subramanya. (2020). “Discriminative Pre-training for Low Resource Title Compression in Conversational Grocery”. *arXiv preprint arXiv:2012.06943*.
- Nemat, R. (2011). “Taking a Look at Different Types of E-commerce”. *World Applied Programming*. 1(2): 100–104.
- Nguyen, T. V., N. Rao, and K. Subbian. (2020). “Learning Robust Models for E-Commerce Product Search”. In: *Proceedings of ACL*. 6861–6869.
- Nguyen, T., X. Wu, A.-T. Luu, C.-D. Nguyen, Z. Hai, and L. Bing. (2022). “Adaptive Contrastive Learning on Multimodal Transformer for Review Helpfulness Predictions”. In: *Proceedings of EMNLP*.
- Ni, J., Z. C. Lipton, S. Vikram, and J. McAuley. (2017). “Estimating Reactions and Recommending Products with Generative Models of Reviews”. In: *Proceedings of IJCNLP*. 783–791.

- Ni, J. and J. McAuley. (2018). “Personalized Review Generation by Expanding Phrases and Attending on Aspect-Aware Representations”. In: *Proceedings of ACL*. 706–711.
- Niu, X., B. Li, C. Li, R. Xiao, H. Sun, H. Deng, and Z. Chen. (2020). “A Dual Heterogeneous Graph Attention Network to Improve Long-Tail Performance for Shop Search in E-Commerce”. In: *Proceedings of KDD*. 3405–3415.
- Nogueira, R., W. Yang, K. Cho, and J. Lin. (2019). “Multi-stage Document Ranking with BERT”. *arXiv preprint arXiv:1910.14424*.
- Nurmi, P., E. Lagerspetz, W. Buntine, P. Floréen, and J. Kukkonen. (2008). “Product Retrieval for Grocery Stores”. In: *Proceedings of SIGIR*. 781–782.
- O’Brien, H. L. and E. G. Toms. (2010). “The Development and Evaluation of a Survey to Measure User Engagement”. *Journal of the American Society for Information Science and Technology*. 61(1): 50–69.
- O’Hare, N., P. De Juan, R. Schifanella, Y. He, D. Yin, and Y. Chang. (2016). “Leveraging User Interaction Signals for Web Image Search”. In: *Proceedings of SIGIR*. 559–568.
- Oentaryo, R. J., E.-P. Lim, J.-W. Low, D. Lo, and M. Finegold. (2014). “Predicting Response in Mobile Advertising with Hierarchical Importance-Aware Factorization Machine”. In: *Proceedings of WSDM*. 123–132.
- Onal, K. D., Y. Zhang, I. S. Altıngövdü, M. M. Rahman, P. Karagoz, A. Braylan, B. Dang, H.-L. Chang, H. Kim, Q. McNamara, A. Angert, E. Banner, V. Khetan, T. McDonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. de Rijke, and M. Lease. (2018). “Neural Information Retrieval: At the End of the Early Years”. *Information Retrieval Journal*. 21(2–3): 111–182.
- Oord, A. van den, O. Vinyals, and K. Kavukcuoglu. (2017). “Neural Discrete Representation Learning”. In: *Proceedings of NIPS*. 6309–6318.
- Oosterhuis, H. and M. de Rijke. (2017). “Balancing Speed and Quality in Online Learning to Rank for Information Retrieval”. In: *Proceedings of CIKM*. 277–286.

- Ouyang, Y., M. Chen, X. Dai, Y. Zhao, S. Huang, and J. Chen. (2020). “Dialogue State Tracking with Explicit Slot Connection Modeling”. In: *Proceedings of ACL*. 34–40.
- Ovaysi, Z., R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva. (2020). “Correcting for Selection Bias in Learning-to-Rank Systems”. In: *Proceedings of Web Conference*. 1863–1873.
- Oved, N. and R. Levy. (2021). “PASS: Perturb-and-Select Summarizer for Product Reviews”. In: *Proceedings of ACL*. 351–365.
- Pan, S. J., X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. (2010). “Cross-Domain Sentiment Classification via Spectral Feature Alignment”. In: *Proceedings of Web Conference*. 751–760.
- Pang, B., L. Lee, and S. Vaithyanathan. (2002). “Thumbs up? Sentiment Classification using Machine Learning Techniques”. In: *Proceedings of EMNLP*. 79–86.
- Pang, L., Y. Lan, J. Guo, J. Xu, L. Su, and X. Cheng. (2019). “HAS-QA: Hierarchical Answer Spans Model for Open-Domain Question Answering”. In: *Proceedings of AAAI*. 6875–6882.
- Pang, L., Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng. (2016). “Text Matching as Image Recognition”. In: *Proceedings of AAAI*. Vol. 30. No. 1.
- Papernot, N., M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar. (2016). “Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data”. In: *Proceedings of ICLR*.
- Park, D. H., H. D. Kim, C. Zhai, and L. Guo. (2015). “Retrieval of Relevant Opinion Sentences for New Products”. In: *Proceedings of SIGIR*. 393–402.
- Parthasarathi, P. and J. Pineau. (2018). “Extending Neural Generative Conversational Model using External Knowledge Sources”. In: *Proceedings of EMNLP*. 690–695.
- Pavlou, P. A. and M. Fygenon. (2006). “Understanding and Predicting Electronic Commerce Adoption: An Extension of the Theory of Planned Behavior”. *MIS quarterly*. 30(1): 115–143.
- Peeters, R., C. Bizer, and G. Glavaš. (2020). “Intermediate Training of BERT for Product Matching”. In: *Proceedings of CEUR Workshop*. 1–2.

- Pei, C., Y. Zhang, Y. Zhang, F. Sun, X. Lin, H. Sun, J. Wu, P. Jiang, J. Ge, W. Ou, and D. Pei. (2019). “Personalized Re-Ranking for Recommendation”. In: *Proceedings of RecSys*. 3–11.
- Peng, B., C. Zhu, C. Li, X. Li, J. Li, M. Zeng, and J. Gao. (2020). “Few-shot Natural Language Generation for Task-oriented Dialog”. In: *Proceedings of EMNLP*. 172–182.
- Peng, C.-Y. J., K. L. Lee, and G. M. Ingersoll. (2002). “An Introduction to Logistic Regression Analysis and Reporting”. *The Journal of Educational Research*. 96(1): 3–14.
- Perez, J. and F. Liu. (2017). “Dialog State Tracking, A Machine Reading Approach Using Memory Network”. In: *Proceedings EACL*.
- Perozzi, B., R. Al-Rfou, and S. Skiena. (2014). “DeepWalk: Online Learning of Social Representations”. In: *Proceedings of KDD*. 701–710.
- Petroni, F., T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel. (2019). “Language Models as Knowledge Bases?” In: *Proceedings of EMNLP*. 2463–2473.
- Pi, Q., W. Bian, G. Zhou, X. Zhu, and K. Gai. (2019). “Practice on Long Sequential User Behavior Modeling for Click-through Rate Prediction”. In: *Proceedings of KDD*. 2671–2679.
- Pontiki, M., D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos. (2015). “SemEval-2015 Task 12: Aspect Based Sentiment Analysis”. In: *Proceedings of SemEval*. 486–495.
- Qin, L., M. Galley, C. Brockett, and X. Liu. (2019). “Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading”. In: *Proceedings of ACL*. 5427–5436.
- Qin, T. (2020). “Dual Learning for Machine Translation and Beyond”. In: *Dual Learning*. Springer. 49–72.
- Qiu, J., Z. Lin, and Y. Li. (2015). “Predicting Customer Purchase Behavior in the E-commerce Context”. *Electronic Commerce Research*. 15(4): 427–452.
- Qiu, X. and X. Huang. (2015). “Convolutional Neural Tensor Network Architecture for Community-Based Question Answering”. In: *Proceedings of IJCAI*. 1305–1311.

- Qiu, Y., C. Zhao, H. Zhang, J. Zhuo, T. Li, X. Zhang, S. Wang, S. Xu, B. Long, and W.-Y. Yang. (2022). “Pre-training Tasks for User Intent Detection and Embedding Retrieval in E-commerce Search”. In: *Proceedings of CIKM*. 4424–4428.
- Qiu, Z., X. Wu, J. Gao, and W. Fan. (2021). “U-BERT: Pre-training user representations for improved recommendation”. In: *Proceedings of AAAI*. Vol. 35. No. 5. 4320–4327.
- Qu, L., M. Liu, J. Wu, Z. Gao, and L. Nie. (2021). “Dynamic Modality Interaction Modeling for Image-text Retrieval”. In: *Proceedings of SIGIR*. 1104–1113.
- Qu, Y., H. Cai, K. Ren, W. Zhang, Y. Yu, Y. Wen, and J. Wang. (2016). “Product-Based Neural Networks for User Response Prediction”. In: *Proceedings of ICDM*. IEEE. 1149–1154.
- Qu, Y., B. Fang, W. Zhang, R. Tang, M. Niu, H. Guo, Y. Yu, and X. He. (2018). “Product-Based Neural Networks for User Response Prediction over Multi-Field Categorical Data”. *ACM Transactions on Information Systems*. 37(1): 1–35.
- Quadrana, M., A. Karatzoglou, B. Hidasi, and P. Cremonesi. (2017). “Personalizing Session-Based Recommendations with Hierarchical Recurrent Neural Networks”. In: *Proceedings of RecSys*. 130–137.
- Quijano-Sanchez, L., C. Sauer, J. A. Recio-Garcia, and B. Diaz-Agudo. (2017). “Make It Personal: A Social Explanation System Applied to Group Recommendations”. *Expert Systems with Applications*. 76: 36–48.
- Radev, D. R., H. Qi, H. Wu, and W. Fan. (2002). “Evaluating Web-based Question Answering Systems.” In: *Proceedings of LREC*. Citeseer.
- Radford, A., R. Jozefowicz, and I. Sutskever. (2018). “Learning to Generate Reviews and Discovering Sentiment”. In: *Proceedings of ICLR*.
- Radlinski, F. and N. Craswell. (2017). “A Theoretical Framework for Conversational Search”. In: *Proceedings of CHIIR*. 117–126.
- Radlinski, F., R. Kleinberg, and T. Joachims. (2008). “Learning Diverse Rankings with Multi-Armed Bandits”. In: *Proceedings of ICML*. 784–791.

- Rahdari, B., T. Arabghalizi, and M. Brambilla. (2017). “Analysis of Online User Behaviour for Art and Culture Events”. In: *Proceedings of International Cross-Domain Conference*. 219–236.
- Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. (2016). “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. In: *Proceedings of EMNLP*. 2383–2392.
- Ramadan, O., P. Budzianowski, and M. Gasic. (2018). “Large-Scale Multi-Domain Belief Tracking with Knowledge Sharing”. In: *Proceedings of ACL*. 432–437.
- Rao, S. X., S. Zhang, Z. Han, Z. Zhang, W. Min, Z. Chen, Y. Shan, Y. Zhao, and C. Zhang. (2020). “xFraud: Explainable Fraud Transaction Detection on Heterogeneous Graphs”. *arXiv preprint arXiv:2011.12193*.
- Rashkin, H., E. M. Smith, M. Li, and Y. Boureau. (2018). “I Know the Feeling: Learning to Converse with Empathy”. *CoRR*. abs/1811.00207.
- Rashkin, H., E. M. Smith, M. Li, and Y.-L. Boureau. (2019). “Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset”. In: *Proceedings of ACL*. 5370–5381.
- Rastogi, A., X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. (2020). “Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset”. In: *Proceedings of AAAI*. 8689–8696.
- Regelson, M. and D. Fain. (2006). “Predicting Click-Through Rate Using Keyword Clusters”. In: *Proceedings of Second Workshop on Sponsored Search Auctions*. Vol. 9623. 1–6.
- Ren, H., W. Xu, Y. Zhang, and Y. Yan. (2013). “Dialog State Tracking using Conditional Random Fields”. In: *Proceedings of SIGDIAL*. 457–461.
- Ren, K., J. Qin, Y. Fang, W. Zhang, L. Zheng, W. Bian, G. Zhou, J. Xu, Y. Yu, X. Zhu, *et al.* (2019a). “Lifelong Sequential Modeling with Personalized Memorization for User Response Prediction”. In: *Proceedings of SIGIR*. 565–574.
- Ren, L., J. Ni, and J. McAuley. (2019b). “Scalable and Accurate Dialogue State Tracking via Hierarchical Sequence Generation”. In: *Proceedings of EMNLP-IJCNLP*. 1876–1885.

- Ren, P., Z. Chen, J. Li, Z. Ren, J. Ma, and M. de Rijke. (2019c). “RepeatNet: A Repeat Aware Neural Recommendation Machine for Session-based Recommendation”. In: *Proceedings of AAAI*. 4806–4813.
- Ren, P., Z. Chen, C. Monz, J. Ma, and M. de Rijke. (2020). “Thinking Globally, Acting Locally: Distantly Supervised Global-to-Local Knowledge Selection for Background Based Conversation”. In: *Proceedings of AAAI*.
- Ren, P., Z. Chen, Z. Ren, F. Wei, J. Ma, and M. de Rijke. (2017). “Leveraging Contextual Sentence Relations for Extractive Summarization using a Neural Attention Model”. In: *Proceedings of SIGIR*. 95–104.
- Ren, P., Z. Liu, X. Song, H. Tian, Z. Chen, Z. Ren, and M. de Rijke. (2021). “Wizard of Search Engine: Access to Information Through Conversations with Search Engines”. *arXiv preprint arXiv:2105.08301*.
- Ren, Z., X. He, D. Yin, and M. de Rijke. (2018). “Information Discovery in E-commerce: Half-day SIGIR 2018 Tutorial”. In: *Proceedings of SIGIR*. 1379–1382.
- Ren, Z., Z. Tian, D. Li, P. Ren, L. Yang, X. Xin, H. Liang, M. de Rijke, and Z. Chen. (2022). “Variational Reasoning about User Preferences for Conversational Recommendation”. In: *Proceedings of SIGIR*. 165–175.
- Rendle, S. (2010). “Factorization Machines”. In: *Proceedings of ICDM*. 995–1000.
- Rendle, S., C. Freudenthaler, and L. Schmidt-Thieme. (2010). “Factorizing Personalized Markov Chains for Next-basket Recommendation”. In: *Proceedings of Web Conference*. 811–820.
- Rendle, S. and L. Schmidt-Thieme. (2010). “Pairwise Interaction Tensor Factorization for Personalized Tag Recommendation”. In: *Proceedings of WSDM*. 81–90.
- Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. (1994). “GroupLens: An Open Architecture for Collaborative Filtering of Netnews”. In: *Proceedings of CSCW*. ACM. 175–186.

- Richardson, M., E. Dominowska, and R. Ragno. (2007). “Predicting Clicks: Estimating the Click-through Rate for New Ads”. In: *Proceedings of Web Conference*. 521–530.
- Ritter, A., C. Cherry, and W. B. Dolan. (2011). “Data-Driven Response Generation in Social Media”. In: *Proceedings of EMNLP*. 583–593.
- Roberts, A., C. Raffel, and N. M. Shazeer. (2020). “How Much Knowledge Can You Pack into the Parameters of a Language Model?” In: *Proceedings of EMNLP*. 5418–5426.
- Roegiest, A., A. Lipani, A. Beutel, A. Olteanu, A. Lucic, A.-A. Stoica, A. Das, A. Biega, B. Voorn, C. Hauff, D. Spina, D. Lewis, D. W. Oard, E. Yilmaz, F. Hasibi, G. Kazai, G. McDonald, H. Haned, I. Ounis, I. van der Linden, J. Garcia-Gathright, J. Baan, K. N. Lau, K. Balog, M. de Rijke, M. Sayed, M. Panteli, M. Sanderson, M. Lease, M. D. Ekstrand, P. Lahoti, and T. Kamishima. (2019). “FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval”. *SIGIR Forum*. 53(2): 20–43.
- Roller, S., E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, Y.-L. Boureau, and J. Weston. (2021). “Recipes for Building an Open-Domain Chatbot”. In: *Proceedings of EACL*. 300–325.
- Rosales, R., H. Cheng, and E. Manavoglu. (2012). “Post-Click Conversion Modeling and Analysis for Non-Guaranteed Delivery Display Advertising”. In: *Proceedings of WSDM*. 293–302.
- Rowley, J. (2000). “Product Search in E-shopping: A Review and Research Propositions”. *Journal of Consumer Marketing*. 17(1): 20–35.
- Rozen, O., D. Carmel, A. Mejer, V. Mirkis, and Y. Ziser. (2021). “Answering Product-Questions by Utilizing Questions from Other Contextually Similar Products”. *arXiv preprint arXiv:2105.08956*.
- Ryu, P.-M., M.-G. Jang, and H.-K. Kim. (2014). “Open Domain Question Answering Using Wikipedia-based Knowledge Model”. *Information Processing & Management*. 50(5): 683–692.
- Saad, D. (1998). “Online Algorithms and Stochastic Approximations”. *Online Learning*. 5: 6–3.

- Sabour, S., C. Zheng, and M. Huang. (2022). “CEM: Commonsense-Aware Empathetic Response Generation”. In: *Proceedings of AAAI*. 11229–11237.
- Sachan, D. S., M. Patwary, M. Shoeybi, N. Kant, W. Ping, W. L. Hamilton, and B. Catanzaro. (2021). “End-to-end Training of Neural Retrievers for Open-domain Question Answering”. In: *Proceedings of ACL*. 6648–6662.
- Sakurai, K., R. Togo, T. Ogawa, and M. Haseyama. (2020). “Music Playlist Generation Based on Reinforcement Learning Using Acoustic Feature Map”. In: *Proceedings of GCCE*. IEEE. 942–943.
- Sakurai, K., R. Togo, T. Ogawa, and M. Haseyama. (2021). “Music Playlist Generation Based on Graph Exploration Using Reinforcement Learning”. In: *Proceedings of LifeTech*. IEEE. 53–54.
- Sakurai, K., R. Togo, T. Ogawa, and M. Haseyama. (2022). “Controlable Music Playlist Generation Based on Knowledge Graph and Reinforcement Learning”. *Sensors*. 22(10): 3722.
- Santhanam, S. and S. Shaikh. (2019). “Emotional Neural Language Generation Grounded in Situational Contexts”. In: *Proceedings of 4th Workshop on Computational Creativity in Language Generation*. 22–27.
- Sarikaya, R., G. E. Hinton, and B. Ramabhadran. (2011). “Deep Belief Nets for Natural Language Call-routing”. In: *Proceedings of ICASSP*. 5680–5683.
- Sarvi, F., M. Heuss, M. Aliannejadi, S. Schelter, and M. de Rijke. (2022). “Understanding and Mitigating the Effect of Outliers in Fair Ranking”. In: *Proceedings of WSDM*. ACM. 861–869.
- Sarvi, F., N. Voskarides, L. Mooiman, S. Schelter, and M. de Rijke. (2020). “A Comparison of Supervised Learning to Match Methods for Product Search”. In: *Proceedings of SIGIR Workshop on eCommerce*. ACM.
- Sarwar, B., G. Karypis, J. Konstan, and J. Riedl. (2001). “Item-Based Collaborative Filtering Recommendation Algorithms”. In: *Proceedings of Web Conference*. 285–295.
- Scarselli, F., M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. (2009). “The Graph Neural Network Model”. *IEEE Transactions on Neural Networks*. 20(1): 61–80.

- Schuth, A., H. Oosterhuis, S. Whiteson, and M. de Rijke. (2016). “Multileave Gradient Descent for Fast Online Learning to Rank”. In: *Proceedings of WSDM*. 457–466.
- Sculley, D., R. G. Malkin, S. Basu, and R. J. Bayardo. (2009). “Predicting Bounce Rates in Sponsored Search Advertisements”. In: *Proceedings of KDD*. ACM. 1325–1334.
- Seo, M., A. Kembhavi, A. Farhadi, and H. Hajishirzi. (2017). “Bidirectional Attention Flow for Machine Comprehension”. In: *Proceedings of ICLR*.
- Serban, I., A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. (2017a). “A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues”. In: *Proceedings of AAAI*.
- Serban, I. V., T. Klinger, G. Tesauro, K. Talamadupula, B. Zhou, Y. Bengio, and A. C. Courville. (2017b). “Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation.” In: *Proceedings of AAAI*.
- Serban, I. V., A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. (2016). “Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models”. In: *Proceedings of AAAI*. 3776–3784.
- Shan, Y., T. R. Hoens, J. Jiao, H. Wang, D. Yu, and J. Mao. (2016). “Deep Crossing: Web-Scale Modeling without Manually Crafted Combinatorial Features”. In: *Proceedings of KDD*. 255–262.
- Shang, L., Z. Lu, and H. Li. (2015). “Neural Responding Machine for Short-Text Conversation”. In: *Proceedings of ACL*. 1577–1586.
- Shardanand, U. and P. Maes. (1995). “Social Information Filtering: Algorithms for Automating “Word of Mouth””. In: *Proceedings of CHI*. 210–217.
- Sharma, A. and D. Cosley. (2013). “Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems”. In: *Proceedings of Web Conference*. ACM. 1133–1144.
- Sharma, V., H. Sharma, A. Bishnu, and L. Patel. (2018). “Cyclegen: Cyclic Consistency based Product Review Generator from Attributes”. In: *Proceedings of INLG*. 426–430.

- Shawar, B. A. and E. Atwell. (2007). “Chatbots: Are they Really Useful?” *Ldv Forum*. 22(1): 29–49.
- Shen, L. and Y. Feng. (2020). “CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation”. In: *Proceedings of ACL*. 556–566.
- Shen, T., J. Zuo, F. Shi, J. Zhang, L. Jiang, M. Chen, Z. Zhang, W. Zhang, X. He, and T. Mei. (2021). “ViDA-MAN: Visual Dialog with Digital Humans”. In: *Proceedings of MM*. 2789–2791.
- Shen, Y., X. He, J. Gao, L. Deng, and G. Mesnil. (2014). “Learning Semantic Representations Using Convolutional Neural Networks for Web Search”. In: *Proceedings of Web Conference*. 373–374.
- Sheng, Z., T. Zhang, and Y. Zhang. (2021). “HTDA: Hierarchical Time-based Directional Attention Network for Sequential User Behavior Modeling”. *Neurocomputing*. 441: 323–334.
- Shih, S. and H. Chi. (2018). “Automatic, Personalized, and Flexible Playlist Generation using Reinforcement Learning”. In: *Proceedings of ISMIR*. 168–174.
- Shin, J., P. Xu, A. Madotto, and P. Fung. (2019). “HappyBot: Generating Empathetic Dialogue Responses by Improving User Experience Look-ahead”. *CoRR*. abs/1906.08487.
- Si, W. M., M. Backes, J. Blackburn, E. De Cristofaro, G. Stringhini, S. Zannettou, and Y. Zhang. (2022). “Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots”. In: *Proceedings of CCS*. 2659–2673.
- Si, Z., Z. Sun, X. Zhang, J. Xu, X. Zang, Y. Song, K. Gai, and J.-R. Wen. (2023). “When Search Meets Recommendation: Learning Disentangled Search Representation for Recommendation”. In: *Proceedings of SIGIR*. 1313–1323.
- Simmons, R. F. (1965). “Answering English Questions by Computer: A Survey”. *Communications of the ACM*. 8(1): 53–70.
- Singh, A. and T. Joachims. (2018). “Fairness of Exposure in Rankings”. In: *Proceedings of KDD*. 2219–2228.
- Singh, D., S. Reddy, W. Hamilton, C. Dyer, and D. Yogatama. (2021). “End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering”. In: *Proceedings of NIPS*. 25968–25981.

- Singh, G., N. Parikh, and N. Sundaresan. (2012). “Rewriting Null E-Commerce Queries to Recommend Products”. In: *Proceedings of Web Conference*. 73–82.
- Siro, C., M. Aliannejadi, and M. de Rijke. (2022). “Understanding User Satisfaction with Task-Oriented Dialogue Systems”. In: *Proceedings of SIGIR*. ACM. 2018–2023.
- Sismeyro, C. and R. E. Bucklin. (2004). “Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach”. *Journal of marketing research*. 41(3): 306–323.
- Skowron, M., M. Theunis, S. Rank, and A. Kappas. (2013). “Affect and Social Processes in Online Communication—Experiments with an Affective Dialog System”. *IEEE Transactions on Affective Computing*. 4(3): 267–279.
- Smola, A. and V. Vapnik. (1997). “Support Vector Regression Machines”. In: *Proceedings of NIPS*. 155–161.
- Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. (2013). “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”. In: *Proceedings of EMNLP*. 1631–1642.
- Solomon, M. R. and C. Behavior. (1994). “Consumer Buying, Having and Being”. *London: Prentice Hall*.
- Song, H., Z. Ren, S. Liang, P. Li, J. Ma, and M. de Rijke. (2017). “Summarizing Answers in Non-factoid Community Question-answering”. In: *Proceedings of WSDM*. 405–414.
- Song, Q., D. Cheng, H. Zhou, J. Yang, Y. Tian, and X. Hu. (2020a). “Towards Automated Neural Interaction Discovery for Click-through Rate Prediction”. In: *Proceedings of KDD*. 945–955.
- Song, S. (2021). “An Online Question Answering System based on Sub-graph Searching”. *arXiv preprint arXiv:2107.13684*.
- Song, S., C. Wang, H. Chen, and H. Chen. (2020b). “TCNN: Triple Convolutional Neural Network Models for Retrieval-based Question Answering System in E-commerce”. In: *Proceedings of Web Conference*. 844–845.
- Song, Y., A. M. Elkahky, and X. He. (2016). “Multi-Rate Deep Learning for Temporal Recommendation”. In: *Proceedings of SIGIR*. 909–912.

- Sordoni, A., M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. (2015). “A Neural Network Approach to Context-Sensitive Generation of Conversational Responses”. In: *Proceedings of NAACL*. 196–205.
- Sprangers, O., S. Schelter, and M. de Rijke. (2023). “Parameter Efficient Deep Probabilistic Forecasting”. *International Journal of Forecasting*. 39(1): 332–345.
- Srihari, R. and W. Li. (1999). “Information Extraction Supported Question Answering”. *Tech. rep.* Williamsville NY: Cymfony Net Inc.
- Steck, H. (2013). “Evaluation of Recommendations: Rating-Prediction and Ranking”. In: *Proceedings of RecSys*. 213–220.
- Su, N., J. He, Y. Liu, M. Zhang, and S. Ma. (2018a). “User Intent, Behaviour, and Perceived Satisfaction in Product Search”. In: *Proceedings of WSDM*. 547–555.
- Su, N., Y. Liu, Z. Li, Y. Liu, M. Zhang, and S. Ma. (2018b). “Detecting Crowdturfing “Add to Favorites” Activities in Online Shopping”. In: *Proceedings of Web Conference*. 1673–1682.
- Su, P.-H., M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young. (2016). “Continuously Learning Neural Dialogue Management”. *arXiv preprint arXiv:1606.02689*.
- Su, X. and T. M. Khoshgoftaar. (2009). “A Survey of Collaborative Filtering Techniques”. *Advances in Artificial Intelligence*. 2009.
- Su, Y., R. Zhang, S. Erfani, and Z. Xu. (2021). “Detecting Beneficial Feature Interactions for Recommender Systems”. In: *Proceedings of AAAI*.
- Su, Y., D. Cai, Q. Zhou, Z. Lin, S. Baker, Y. Cao, S. Shi, N. Collier, and Y. Wang. (2020a). “Dialogue Response Selection with Hierarchical Curriculum Learning”. *arXiv preprint arXiv:2012.14756*.
- Su, Y., L. Zhang, Q. Dai, B. Zhang, J. Yan, D. Wang, Y. Bao, S. Xu, Y. He, and W. Yan. (2020b). “An Attention-based Model for Conversion Rate Prediction with Delayed Feedback via Post-click Calibration”. In: *Proceedings of IJCAI*. 3522–3528.

- Suh, E., S. Lim, H. Hwang, and S. Kim. (2004). “A Prediction Model for the Purchase Probability of Anonymous Customers to Support Real Time Web Marketing: A Case Study”. *Expert Systems with Applications*. 27(2): 245–255.
- Sun, F., P. Jiang, H. Sun, C. Pei, W. Ou, and X. Wang. (2018a). “Multi-Source Pointer Network for Product Title Summarization”. In: *Proceedings of CIKM*. 7–16.
- Sun, F., J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. (2019a). “BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer”. In: *Proceedings of CIKM*. 1441–1450.
- Sun, H., B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. W. Cohen. (2018b). “Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text”. In: *Proceedings of EMNLP*. 4231–4242.
- Sun, L., Y. Bai, B. Du, C. Liu, H. Xiong, and W. Lv. (2020a). “Dual Sequential Network for Temporal Sets Prediction”. In: *Proceedings of SIGIR*. ACM. 1439–1448.
- Sun, S., C. Luo, and J. Chen. (2017). “A Review of Natural Language Processing Techniques for Opinion Mining Systems”. *Information Fusion*. 36: 10–25.
- Sun, W., C. Meng, Q. Meng, Z. Ren, P. Ren, Z. Chen, and M. de Rijke. (2021a). “Conversations Powered by Cross-Lingual Knowledge”. In: *Proceedings of SIGIR*. 1442–1451.
- Sun, W., Z. Shi, S. Gao, P. Ren, M. de Rijke, and Z. Ren. (2023a). “Contrastive Learning Reduces Hallucination in Conversations”. In: *Proceedings of AAAI*. 13618–13626.
- Sun, W., L. Yan, Z. Chen, S. Wang, H. Zhu, P. Ren, Z. Chen, D. Yin, M. de Rijke, and Z. Ren. (2024). “Learning to Tokenize for Generative Retrieval”. *Proceedings of NIPS*. 36.
- Sun, W., L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, and Z. Ren. (2023b). “Is ChatGPT Good at Search? Investigating Large Language Models as Re-ranking Agents”. *arXiv preprint arXiv:2304.09542*.

- Sun, W., S. Zhang, K. Balog, Z. Ren, P. Ren, Z. Chen, and M. de Rijke. (2021b). “Simulating User Satisfaction for the Evaluation of Task-Oriented Dialogue Systems”. In: *Proceedings of SIGIR*. 2499–2506.
- Sun, W., S. Khenissi, O. Nasraoui, and P. Shafto. (2019b). “Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering”. In: *Proceedings of Web Conference*. 645–651.
- Sun, Y., Y. Hu, L. Xing, J. Yu, and Y. Xie. (2020b). “History-Adaption Knowledge Incorporation Mechanism for Multi-Turn Dialogue System.” In: *Proceedings of AAAI*. 8944–8951.
- Sun, Y. and Y. Zhang. (2018). “Conversational Recommender System”. In: *Proceedings of SIGIR*. 235–244.
- Sun, Y., Y. Zhang, Y. Chen, and R. Jin. (2016). “Conversational Recommendation System with Unsupervised Learning”. In: *Proceedings of RecSys*. 397–398.
- Sunehag, P., R. Evans, G. Dulac-Arnold, Y. Zwols, D. Visentin, and B. Coppin. (2015). “Deep Reinforcement Learning with Attention for Slate Markov Decision Processes with High-Dimensional States and Actions”. *ArXiv*.
- Susan, M. M. and S. David. (2010). “What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com”. *MIS Quarterly*. 34(1): 185–200.
- Sutskever, I., O. Vinyals, and Q. V. Le. (2014). “Sequence to Sequence Learning with Neural Networks”. In: *Proceedings of NIPS*. 3104–3112.
- Swaminathan, A., A. Krishnamurthy, A. Agarwal, M. Dudík, J. Langford, D. Jose, and I. Zitouni. (2017). “Off-policy Evaluation for Slate Recommendation”. In: *Proceedings of NIPS*. 3632–3642.
- Swinyard, W. R. and S. M. Smith. (2004). “Activities, Interests, and Opinions of Online Shoppers and Non-Shoppers”. *International Business and Economics Research Journal*. 3: 37–48.
- Takanobu, R., T. Zhuang, M. Huang, J. Feng, H. Tang, and B. Zheng. (2019). “Aggregating E-Commerce Search Results from Heterogeneous Sources via Hierarchical Reinforcement Learning”. In: *Proceedings of Web Conference*. 1771–1781.

- Tan, J., A. Kotov, R. Pir Mohammadiani, and Y. Huo. (2017). “Sentence Retrieval with Sentiment-Specific Topical Anchoring for Review Summarization”. In: *Proceedings of CIKM*. 2323–2326.
- Tan, W., L. Zhu, W. Guan, J. Li, and Z. Cheng. (2022). “Bit-aware Semantic Transformer Hashing for Multi-modal Retrieval”. In: *Proceedings of SIGIR*. 982–991.
- Tang, D., B. Qin, and T. Liu. (2015). “Learning Semantic Representations of Users and Products for Document Level Sentiment Classification”. In: *Proceedings of ACL-IJCNLP*. Vol. 1. 1014–1023.
- Tang, D., F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. (2014). “Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification”. In: *Proceedings of ACL*. 1555–1565.
- Tang, J. and K. Wang. (2018). “Personalized Top-N Sequential Recommendation via Convolutional Sequence Embedding”. In: *Proceedings of WSDM*. 565–573.
- Tang, J., L. Yao, D. Zhang, and J. Zhang. (2010). “A Combination Approach to Web User Profiling”. *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 5(1): 2.
- Tao, C., C. Chen, J. Feng, J.-R. Wen, and R. Yan. (2021a). “A Pre-training Strategy for Zero-Resource Response Selection in Knowledge-Grounded Conversations”. In: *Proceedings of ACL-IJCNLP*. 4446–4457.
- Tao, C., J. Feng, R. Yan, W. Wu, and D. Jiang. (2021b). “A Survey on Response Selection for Retrieval-based Dialogues”. In: *Proceedings of IJCAI*.
- Tapeh, A. G. and M. Rahgozar. (2008). “A Knowledge-based Question Answering System for B2C eCommerce”. *Knowledge-Based Systems*. 21(8): 946–950.
- Tay, Y., V. Tran, M. Dehghani, J. Ni, D. Bahri, H. Mehta, Z. Qin, K. Hui, Z. Zhao, J. Gupta, *et al.* (2022). “Transformer Memory as a Differentiable Search Index”. *Proceedings of NIPS*. 35: 21831–21843.
- Testa, F., M. V. Russo, T. B. Cornwell, A. McDonald, and B. Reich. (2018). “Social Sustainability as Buying Local: Effects of Soft Policy, Meso-Level Actors, and Social Influences on Purchase Intentions”. *Journal of Public Policy & Marketing*. 37(1): 152–166.

- Tian, Z., W. Bi, D. Lee, L. Xue, Y. Song, X. Liu, and N. L. Zhang. (2020). “Response-Anticipated Memory for On-Demand Knowledge Integration in Response Generation”. In: *Proceedings of ACL*. 650–659.
- Tomasi, F., J. Cauteruccio, S. Kanoria, K. Ciosek, M. Rinaldi, and Z. Dai. (2023). “Automatic Music Playlist Generation via Simulation-based Reinforcement Learning”. In: *Proceedings of KDD*. ACM. 4948–4957.
- Tracz, J., P. I. Wójcik, K. Jasinska-Kobus, R. Belluzzo, R. Mroczkowski, and I. Gawlik. (2020). “BERT-based Similarity Learning for Product Matching”. In: *Proceedings of Workshop on Natural Language Processing in E-Commerce*. 66–75.
- Tran, V.-K. and L.-M. Nguyen. (2017). “Semantic Refinement GRU-Based Neural Language Generation for Spoken Dialogue Systems”. In: *Proceedings of PACLING*. 63–75.
- Tripathi, R., B. Dhamodharaswamy, S. Jagannathan, and A. Nandi. (2019). “Detecting Sensitive Content in Spoken Language”. In: *Proceedings of DSAA*. 374–381.
- Tripathy, A., A. Agrawal, and S. K. Rath. (2016). “Classification of Sentiment Reviews using N-gram Machine Learning Approach”. *Expert Systems with Applications*. 57: 117–126.
- Trippas, J. R., D. Spina, L. Cavedon, H. Joho, and M. Sanderson. (2018). “Informing the Design of Spoken Conversational Search: Perspective Paper”. In: *Proceedings of CHIIR*. 32–41.
- Trippas, J. R., D. Spina, P. Thomas, M. Sanderson, H. Joho, and L. Cavedon. (2020). “Towards a Model for Spoken Conversational Search”. *Information Processing and Management*. 57: 102162.
- Trotman, A., J. Degenhardt, and S. Kallumadi. (2017). “The Architecture of eBay Search”. In: *Proceedings of SIGIR*.
- Tsagkias, M., T. H. King, S. Kallumadi, V. Murdock, and M. de Rijke. (2020). “Challenges and Research Opportunities in eCommerce Search and Recommendations”. *SIGIR Forum*. 54(1).
- Tsytsarau, M. and T. Palpanas. (2016). “Managing Diverse Sentiments at Large Scale”. *IEEE Transactions on Knowledge and Data Engineering*. 28(11): 3028–3040.

- Tuan, Y.-L., Y.-N. Chen, and H.-y. Lee. (2019). “DyKgChat: Benchmarking Dialogue Generation Grounding on Dynamic Knowledge Graphs”. In: *Proceedings of EMNLP*. 1855–1865.
- Tur, G., L. Deng, D. Hakkani-Tür, and X. He. (2012). “Towards Deeper Understanding: Deep Convex Networks for Semantic Utterance Classification”. In: *Proceedings of ICASSP*. 5045–5048.
- Vakulenko, S., E. Kanoulas, and M. de Rijke. (2021). “A Large-scale Analysis of Mixed Initiative in Information-seeking Dialogues for Conversational Search”. *ACM Transactions on Information Systems*. 39(4): 1–32.
- Vall, A., M. Quadrana, M. Schedl, and G. Widmer. (2018). “The Importance of Song Context and Song Order in Automated Music Playlist Generation”. *ArXiv*. abs/1807.04690.
- Van den Poel, D. and W. Buckinx. (2005). “Predicting Online-purchasing Behaviour”. *European Journal of Operational Research*. 166(2): 557–575.
- Van Gysel, C., M. de Rijke, and E. Kanoulas. (2016a). “Learning Latent Vector Spaces for Product Search”. In: *Proceedings of CIKM*. 165–174.
- Van Gysel, C., M. de Rijke, and E. Kanoulas. (2018). “Mix’n Match: Integrating Text Matching and Product Substitutability within Product Search”. In: *Proceedings of CIKM*. 1373–1382.
- Van Gysel, C., M. de Rijke, and M. Worring. (2016b). “Unsupervised, Efficient and Semantic Expertise Retrieval”. In: *Proceedings of Web Conference*. 1069–1079.
- Vanderveld, A., A. Pandey, A. Han, and R. Parekh. (2016). “An Engagement-Based Customer Lifetime Value System for E-Commerce”. In: *Proceedings of KDD*. 293–302.
- Vardasbi, A., F. Sarvi, and M. de Rijke. (2022). “Probabilistic Permutation Graph Search: Black-Box Optimization for Fairness in Ranking”. In: *Proceedings of SIGIR*. ACM. 715–725.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017). “Attention is All you Need”. In: *Proceedings of NIPS*. 5998–6008.

- Vendrov, I., T. Lu, Q. Huang, and C. Boutilier. (2020). “Gradient-Based Optimization for Bayesian Preference Elicitation”. In: *Proceedings of AAAI*. 10292–10301.
- Vergo, J., S. Noronha, J. Kramer, J. Lechner, and T. Cofino. (2002). “E-Commerce Interface Design”. In: *The Human-computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. 757–771.
- Vinyals, O., M. Fortunato, and N. Jaitly. (2015). “Pointer Networks”. In: *Proceedings of NIPS*. 2692–2700.
- Vinyals, O. and Q. Le. (2015). “A Neural Conversational Model”. In: *ICML Deep Learning Workshop*.
- Voorhees, E. M., D. M. Tice, *et al.* (1999). “The TREC-8 Question Answering Track Evaluation”. In: *Proceedings of TREC*. 82.
- Vtyurina, A., D. Savenkov, E. Agichtein, and C. L. A. Clarke. (2017). “Exploring Conversational Search With Humans, Assistants, and Wizards”. In: *Proceedings of CHI*. 2187–2193.
- Wan, M. and J. McAuley. (2016). “Modeling Ambiguity, Subjectivity, and Diverging Viewpoints in Opinion Question Answering Systems”. In: *Proceedings of ICDM*. 489–498.
- Wan, M. and J. McAuley. (2018). “Item Recommendation on Monotonic Behavior Chains”. In: *Proceedings of RecSys*. 86–94.
- Wan, M., D. Wang, M. Goldman, M. Taddy, J. Rao, J. Liu, D. Lymberopoulos, and J. McAuley. (2017). “Modeling Consumer Preferences and Price Sensitivities from Large-Scale Grocery Shopping Transaction Logs”. In: *Proceedings of Web Conference*. 1103–1112.
- Wan, S., Y. Lan, J. Guo, J. Xu, L. Pang, and X. Cheng. (2016). “A Deep Architecture for Semantic Matching with Multiple Positional Sentence Representations”. In: *Proceedings of AAAI*. 2835–2841.
- Wang, B., M. Li, Z. Zeng, J. Zhuo, S. Wang, S. Xu, B. Long, and W. Yan. (2023a). “Learning Multi-Stage Multi-Grained Semantic Embeddings for E-Commerce Search”. In: *Proceedings of Web Conference*. 411–415.
- Wang, C., H. Yu, and Y. Zhang. (2023b). “RFiD: Towards Rational Fusion-in-Decoder for Open-Domain Question Answering”. In: *Findings of ACL*. Ed. by A. Rogers, J. L. Boyd-Graber, and N. Okazaki. 2473–2481.

- Wang, H., Q. Wu, and H. Wang. (2017a). “Factorization Bandits for Interactive Recommendation.” In: *Proceedings of AAAI*. 2695–2702.
- Wang, J., J. Liu, W. Bi, X. Liu, K. He, R. Xu, and M. Yang. (2020a). “Improving Knowledge-Aware Dialogue Generation via Knowledge Base Question Answering”. In: *Proceedings of AAAI*. 9169–9176.
- Wang, J., J. Tian, L. Qiu, S. Li, J. Lang, L. Si, and M. Lan. (2018a). “A Multi-task Learning Approach for Improving Product Title Compression with User Search Log Data”. In: *Proceedings of AAAI*. 451–458.
- Wang, J., Y. Hou, J. Liu, Y. Cao, and C.-Y. Lin. (2017b). “A Statistical Framework for Product Description Generation”. In: *Proceedings of IJCNLP*. 187–192.
- Wang, J., P. Huang, H. Zhao, Z. Zhang, B. Zhao, and D. L. Lee. (2018b). “Billion-Scale Commodity Embedding for E-Commerce Recommendation in Alibaba”. In: *Proceedings of KDD*. 839–848.
- Wang, P., J. Guo, Y. Lan, J. Xu, S. Wan, and X. Cheng. (2015). “Learning Hierarchical Representation Model for NextBasket Recommendation”. In: *Proceedings of SIGIR*. ACM. 403–412.
- Wang, P., Y. Zhang, S. Niu, and J. Guo. (2019a). “Modeling Temporal Dynamics of Users’ Purchase Behaviors for Next Basket Prediction”. *J. Comput. Sci. Technol.* 34(6): 1230–1240.
- Wang, Q., X. Liu, W. Liu, A.-A. Liu, W. Liu, and T. Mei. (2020b). “Metasearch: Incremental Product Search via Deep Meta-learning”. *IEEE Transactions on Image Processing*. 29(1): 7549–7564.
- Wang, Q., C. Zeng, W. Zhou, T. Li, S. S. Iyengar, L. Shwartz, and G. Y. Grabarnik. (2018c). “Online Interactive Collaborative Filtering Using Multi-armed Bandit with Dependent Arms”. *IEEE Transactions on Knowledge and Data Engineering*. 31(8): 1569–1580.
- Wang, R., B. Fu, G. Fu, and M. Wang. (2017c). “Deep & Cross Network for Ad Click Predictions”. In: *Proceedings of ADKDD*. 12:1–12:7.
- Wang, S., L. Hu, Y. Wang, Q. Z. Sheng, M. A. Orgun, and L. Cao. (2020c). “Intention Nets: Psychology-Inspired User Choice Behavior Modeling for Next-Basket Prediction”. In: *Proceedings of AAAI*. AAAI Press. 6259–6266.

- Wang, S., M. Yu, X. Guo, Z. Wang, T. Klinger, W. Zhang, S. Chang, G. Tesauro, B. Zhou, and J. Jiang. (2018d). “R³: Reinforced Ranker-Reader for Open-domain Question Answering”. In: *Proceedings of AAAI*. 5981–5988.
- Wang, W., X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang. (2016a). “Effective Deep Learning-based Multi-modal Retrieval”. *The VLDB Journal*. 25(1): 79–101.
- Wang, W., N. Yang, F. Wei, B. Chang, and M. Zhou. (2017d). “Gated Self-Matching Networks for Reading Comprehension and Question Answering”. In: *Proceedings of ACL*. 189–198.
- Wang, W., F. Feng, X. He, H. Zhang, and T.-S. Chua. (2021). “Clicks Can Be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue”. In: *Proceedings of SIGIR*. 1288–1297.
- Wang, X., X. He, Y. Cao, M. Liu, and T.-S. Chua. (2019b). “KGAT: Knowledge Graph Attention Network for Recommendation”. In: *Proceedings of KDD*. 950–958.
- Wang, X., X. He, M. Wang, F. Feng, and T.-S. Chua. (2019c). “Neural Graph Collaborative Filtering”. In: *Proceedings of SIGIR*. 165–174.
- Wang, X., D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua. (2019d). “Explainable Reasoning over Knowledge Graphs for Recommendation”. In: *Proceedings of AAAI*. 5329–5336.
- Wang, X., M. Bendersky, D. Metzler, and M. Najork. (2016b). “Learning to Rank with Selection Bias in Personal Search”. In: *Proceedings of SIGIR*. 115–124.
- Wang, Y., J. Zhao, J. Bao, C. Duan, Y. Wu, and X. He. (2022a). “LUNA: Learning Slot-Turn Alignment for Dialogue State Tracking”. In: *Proceedings of NAACL*. 3319–3328.
- Wang, Y., Y. Hou, H. Wang, Z. Miao, S. Wu, Q. Chen, Y. Xia, C. Chi, G. Zhao, Z. Liu, *et al.* (2022b). “A Neural Corpus Indexer for Document Retrieval”. *Proceedings of NIPS*. 35: 25600–25614.
- Wang, Y., M. Liu, Y. Wei, Z. Cheng, Y. Wang, and L. Nie. (2022c). “Siamese Alignment Network for Weakly Supervised Video Moment Retrieval”. *IEEE Transactions on Multimedia*. 14(8): 1–13.
- Wang, Z., G. Lin, H. Tan, Q. Chen, and X. Liu. (2020d). “CKAN: Collaborative Knowledge-aware Attentive Network for Recommender Systems”. In: *Proceedings of SIGIR*. 219–228.

- Wang, Z. and O. Lemon. (2013). “A Simple and Generic Belief Tracking Mechanism for the Dialog State Tracking Challenge: On the believability of observed information”. In: *Proceedings of SIGDIAL*. 423–432.
- Wang, Z., Z. Jiang, Z. Ren, J. Tang, and D. Yin. (2018e). “A Path-constrained Framework for Discriminating Substitutable and Complementary Products in E-commerce”. In: *Proceedings of WSDM*. 619–627.
- Wei, J., Y. Yang, X. Xu, X. Zhu, and H. T. Shen. (2022). “Universal weighting metric learning for cross-modal retrieval”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 44(10): 6534–6545.
- Wei, W., J. Liu, X. Mao, G. Guo, F. Zhu, P. Zhou, and Y. Hu. (2019). “Emotion-aware Chat Machine: Automatic Emotional Response Generation for Human-like Emotional Interaction”. In: *Proceedings of CIKM*. 1401–1410.
- Wen, H., J. Zhang, Q. Lin, K. Yang, and P. Huang. (2019a). “Multi-Level Deep Cascade Trees for Conversion Rate Prediction in Recommendation System”. In: *Proceedings of AAAI*. 338–345.
- Wen, H., J. Zhang, F. Lv, W. Bao, T. Wang, and Z. Chen. (2021). “Hierarchically Modeling Micro and Macro Behaviors via Multi-Task Learning for Conversion Rate Prediction”. In: *Proceedings of SIGIR*. 2187–2191.
- Wen, H., J. Zhang, Y. Wang, F. Lv, W. Bao, Q. Lin, and K. Yang. (2020). “Entire Space Multi-Task Modeling via Post-Click Behavior Decomposition for Conversion Rate Prediction”. In: *Proceedings of SIGIR*. 2377–2386.
- Wen, H., L. Yang, and D. Estrin. (2019b). “Leveraging Post-click Feedback for Content Recommendations”. In: *Proceedings of RecSys*. 278–286.
- Wen, T.-H., M. Gasic, D. Kim, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young. (2015a). “Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking”. In: *Proceedings of SIGDIAL*. 275–284.

- Wen, T.-H., M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young. (2015b). “Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems”. In: *Proceedings of EMNLP*. 1711–1721.
- Wen, T.-H., Y. Miao, P. Blunsom, and S. Young. (2017a). “Latent Intention Dialogue Models”. In: *Proceedings of ICML*. 3732–3741.
- Wen, T.-H., D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young. (2017b). “A Network-Based End-to-End Trainable Task-Oriented Dialogue System”. In: *Proceedings of EACL*. 438–449.
- Whang, T., D. Lee, C. Lee, K. Yang, D. Oh, and H. Lim. (2020). “An Effective Domain Adaptive Post-Training Method for BERT in Response Selection.” In: *Proceedings of INTERSPEECH*. 1585–1589.
- Wilhelm, M., A. Ramanathan, A. Bonomo, S. Jain, E. H. Chi, and J. Gillenwater. (2018). “Practical Diversified Recommendations on YouTube with Determinantal Point Processes”. In: *Proceedings of CIKM*. 2165–2173.
- Williams, J. (2012). “A Belief Tracking Challenge Task for Spoken Dialog Systems”. In: *Proceedings of NAACL*. 23–24.
- Williams, J. (2013). “Multi-domain Learning and Generalization in Dialog State Tracking”. In: *Proceedings of SIGDIAL*. 433–441.
- Williams, J., A. Raux, and M. Henderson. (2016). “The Dialog State Tracking Challenge Series: A Review”. *Dialogue & Discourse*. 7(3): 4–33.
- Williams, J. D. (2014). “Web-style Ranking and SLU Combination for Dialog State Tracking”. In: *Proceedings of SIGDIAL*.
- Williams, J. D., K. Asadi, and G. Zweig. (2017). “Hybrid Code Networks: Practical and Efficient End-to-end Dialog Control with Supervised and Reinforcement Learning”. In: *Proceedings of ACL*. 665–677.
- Williams, J. D., M. Henderson, A. Raux, B. Thomson, A. Black, and D. Ramachandran. (2014). “The Dialog State Tracking Challenge Series”. *AI Magazine*. 35(4): 121–124.
- Wu, C.-S., S. C. Hoi, R. Socher, and C. Xiong. (2020a). “TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue”. In: *Proceedings of EMNLP*. 917–929.

- Wu, C.-S., A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung. (2019a). “Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems”. In: *Proceedings of ACL*. 808–819.
- Wu, C., F. Wu, Y. Yu, T. Qi, Y. Huang, and X. Xie. (2021a). “UserBERT: Contrastive User Model Pre-training”. *arXiv preprint arXiv:2109.01274*.
- Wu, H., Y. Gu, S. Sun, and X. Gu. (2016a). “Aspect-based Opinion Summarization with Convolutional Neural Networks”. In: *Proceedings of IJCNN*. IEEE. 3157–3163.
- Wu, J., X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie. (2021b). “Self-Supervised Graph Learning for Recommendation”. In: *Proceedings of SIGIR*. 726–735.
- Wu, L., D. Hu, L. Hong, and H. Liu. (2018a). “Turning Clicks into Purchases: Revenue Optimization for Product Search in E-Commerce”. In: *Proceedings of SIGIR*. 365–374.
- Wu, Q., H. Wang, L. Hong, and Y. Shi. (2017a). “Returning is Believing: Optimizing Long-Term User Engagement in Recommender Systems”. In: *Proceedings of CIKM*. 1927–1936.
- Wu, S., Y. Li, D. Zhang, and Z. Wu. (2020b). “Improving Knowledge-Aware Dialogue Response Generation by Using Human-Written Prototype Dialogues”. In: *Proceedings of EMNLP*. 1402–1411.
- Wu, S., Y. Li, D. Zhang, Y. Zhou, and Z. Wu. (2020c). “Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness”. In: *Proceedings of ACL*. 5811–5820.
- Wu, W. and R. Yan. (2019). “Deep Chit-Chat: Deep Learning for Chatbots”. In: *Proceedings of SIGIR*. 1413–1414.
- Wu, W., Z. Guo, X. Zhou, H. Wu, X. Zhang, R. Lian, and H. Wang. (2019b). “Proactive Human-Machine Conversation with Explicit Conversation Goal”. In: *Proceedings of ACL*. 3794–3804.
- Wu, X., A. Magnani, S. Chaidaroon, A. Puthenputhussery, C. Liao, and Y. Fang. (2022a). “A Multi-task Learning Framework for Product Ranking with BERT”. In: *Proceedings of Web Conference*. 493–501.
- Wu, Y., C. DuBois, A. X. Zheng, and M. Ester. (2016b). “Collaborative Denoising Auto-Encoders for Top-N Recommender Systems”. In: *Proceedings of WSDM*. 153–162.

- Wu, Y., Z. Li, W. Wu, and M. Zhou. (2018b). “Response Selection with Topic Clues for Retrieval-based Chatbots”. *Neurocomputing*. 316: 251–261.
- Wu, Y., W. Wu, C. Xing, M. Zhou, and Z. Li. (2017b). “Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots”. In: *Proceedings of ACL*. 496–505.
- Wu, Z., X.-Y. Dai, C. Yin, S. Huang, and J. Chen. (2018c). “Improving Review Representations With User Attention and Product Attention for Sentiment Classification”. In: *Proceedings of AAAI*. 5989–5996.
- Wu, Z., W. Bi, X. Li, L. Kong, and B. C. Kao. (2022b). “Lexical Knowledge Internalization for Neural Dialog Generation”. In: *Proceedings of ACL*. 7945–7958.
- Xia, R., F. Xu, C. Zong, Q. Li, Y. Qi, T. Li, *et al.* (2015). “Dual Sentiment Analysis: Considering Two Sides of One Review”. *IEEE Transactions on Knowledge and Data Engineering*. 27(8): 2120–2133.
- Xian, Y., Z. Fu, S. Muthukrishnan, G. De Melo, and Y. Zhang. (2019). “Reinforcement Knowledge Graph Reasoning for Explainable Recommendation”. In: *Proceedings of SIGIR*. 285–294.
- Xiao, J., H. Ye, X. He, H. Zhang, F. Wu, and T. Chua. (2017). “Attentional Factorization Machines: Learning the Weight of Feature Interactions via Attention Networks”. In: *Proceedings of IJCAI*. 3119–3125.
- Xiao, L., J. Ma, X. L. Dong, P. Martinez-Gomez, N. Zalmout, W. Chen, T. Zhao, H. He, and Y. Jin. (2021). “End-to-End Conversational Search for Online Shopping with Utterance Transfer”. In: *Proceedings of EMNLP*. 3477–3486.
- Xiao, T., J. Ren, Z. Meng, H. Sun, and S. Liang. (2019). “Dynamic Bayesian Metric Learning for Personalized Product Search”. In: *Proceedings of CIKM*. 1693–1702.
- Xiao, Z., L. Yang, W. Jiang, Y. Wei, Y. Hu, and H. Wang. (2020). “Deep Multi-interest Network for Click-through Rate Prediction”. In: *Proceedings of CIKM*. 2265–2268.
- Xiong, W. and D. Litman. (2011). “Automatically Predicting Peer-Review Helpfulness”. In: *Proceedings of ACL*. 502–507.

- Xiong, W. and D. Litman. (2014). “Empirical Analysis of Exploiting Review Helpfulness for Extractive Summarization of Online Reviews”. In: *Proceedings of COLING*. 1985–1995.
- Xu, D., J. Liang, W. Cheng, H. Wei, H. Chen, and X. Zhang. (2021a). “Transformer-style Relational Reasoning with Dynamic Memory Updating for Temporal Network Modeling”. In: *Proceedings of AAAI*. 4546–4554.
- Xu, H., B. Liu, L. Shu, and P. S. Yu. (2019). “Review Conversational Reading Comprehension”. *arXiv preprint arXiv:1902.00821*.
- Xu, J., X. He, and H. Li. (2018). “Deep Learning for Matching in Search and Recommendation”. In: *Proceedings of SIGIR*. 1365–1368.
- Xu, J., Z. Lei, H. Wang, Z.-Y. Niu, H. Wu, and W. Che. (2021b). “Discovering Dialog Structure Graph for Coherent Dialog Generation”. In: *Proceedings of ACL-IJCNLP*. 1726–1739.
- Xu, J. and H. Li. (2007). “AdaRank: A Boosting Algorithm for Information Retrieval”. In: *Proceedings of SIGIR*. 391–398.
- Xu, J., H. Wang, Z.-Y. Niu, H. Wu, W. Che, and T. Liu. (2020a). “Conversational Graph Grounded Policy Learning for Open-Domain Conversation Generation”. In: *Proceedings of ACL*. 1835–1845.
- Xu, J., H. Wang, Z. Niu, H. Wu, and W. Che. (2020b). “Knowledge Graph Grounded Goal Planning for Open-Domain Conversation Generation.” In: *Proceedings of AAAI*. 9338–9345.
- Xu, P. and Q. Hu. (2018). “An End-to-end Approach for Handling Unknown Slot Values in Dialogue State Tracking”. In: *Proceedings of ACL*. 1448–1457.
- Xu, R., X. Zhang, B. Li, Y. Zhang, X. Chen, and P. Cui. (2022a). “Product Ranking for Revenue Maximization with Multiple Purchases”. *Advances in Neural Information Processing Systems*. 35: 25132–25145.
- Xu, R., C. Tao, J. Feng, W. Wu, R. Yan, and D. Zhao. (2021c). “Response Ranking with Multi-types of Deep Interactive Representations in Retrieval-based Dialogues”. *ACM Transactions on Information Systems*. 39(4): 1–28.

- Xu, R., C. Tao, D. Jiang, X. Zhao, D. Zhao, and R. Yan. (2021d). “Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues”. In: *Proceedings of AAAI*. 14158–14166.
- Xu, Y., E. Ishii, Z. Liu, G. I. Winata, D. Su, A. Madotto, and P. Fung. (2022b). “Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters”. In: *Proceedings of ACL*. 93–107.
- Xu, Y., H. Zhao, and Z. Zhang. (2021e). “Topic-Aware Multi-turn Dialogue Modeling”. In: *Proceedings of AAAI*. 14176–14184.
- Xu, Z., C. Chen, T. Lukasiewicz, Y. Miao, and X. Meng. (2016). “Tag-Aware Personalized Recommendation Using a Deep-Semantic Similarity Model with Negative Sampling”. In: *Proceedings of CIKM*. 1921–1924.
- Xue, X., S. Huston, and W. B. Croft. (2010). “Improving Verbose Queries Using Subset Distribution”. In: *Proceedings of CIKM*. 1059–1068.
- Yan, R., Y. Song, and H. Wu. (2016). “Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System”. In: *Proceedings of SIGIR*. 55–64.
- Yan, Z., N. Duan, J. Bao, P. Chen, M. Zhou, and Z. Li. (2018). “Response Selection from Unstructured Documents for Human-computer Conversation Systems”. *Knowledge-Based Systems*. 142: 149–159.
- Yan, Z., N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li. (2017). “Building Task-oriented Dialogue Systems for Online Shopping”. In: *Proceedings of AAAI*. 4618–4625.
- Yang, F., A. Kale, Y. Bubnov, L. Stein, Q. Wang, H. Kiapour, and R. Piramuthu. (2017a). “Visual Search at EBay”. In: *Proceedings of KDD*. 2101–2110.
- Yang, H., Q. Lu, A. X. Qiu, and C. Han. (2016). “Large Scale CVR Prediction through Dynamic Transfer Learning of Global and Local Features”. In: *Proceedings of Workshop on Big Data, Streams and Heterogeneous Source Mining*. 103–119.
- Yang, H., Y. Zhao, J. Xia, B. Yao, M. Zhang, and K. Zheng. (2019). “Music Playlist Recommendation with Long Short-Term Memory”. In: *Proceedings of DASFAA*. Vol. 11447. Springer. 416–432.

- Yang, J.-Q., X. Li, S. Han, T. Zhuang, D.-C. Zhan, X. Zeng, and B. Tong. (2021a). “Capturing Delayed Feedback in Conversion Rate Prediction via Elapsed-time Sampling”. In: *Proceedings of AAAI*. 4582–4589.
- Yang, J.-Y., H.-j. Kim, and S.-g. Lee. (2010). “Feature-based Product Review Summarization Utilizing User Score”. *Journal of Information Science and Engineering*. 26(6): 1973–1990.
- Yang, L., M. Qiu, C. Qu, J. Guo, Y. Zhang, W. B. Croft, J. Huang, and H. Chen. (2018a). “Response Ranking with Deep Matching Networks and External Knowledge in Information-Seeking Conversation Systems”. In: *Proceedings of SIGIR*. 245–254.
- Yang, P., H.-Y. Huang, and X.-L. Mao. (2021b). “Comprehensive Study: How the Context Information of Different Granularity Affects Dialogue State Tracking?” In: *Proceedings of ACL-IJCNLP*. 2481–2491.
- Yang, Y., W.-t. Yih, and C. Meek. (2015a). “WIKIQA: A Challenge Dataset for Open-Domain Question Answering”. In: *Proceedings of EMNLP*. 2013–2018.
- Yang, Y., C. Chen, M. Qiu, and F. Bao. (2017b). “Aspect Extraction from Product Reviews Using Category Hierarchy Information”. In: *Proceedings of EACL*. 675–680.
- Yang, Y., Y. Yan, M. Qiu, and F. Bao. (2015b). “Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews”. In: *Proceedings of ACL*. 38–44.
- Yang, Z., P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. (2018b). “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering”. In: *Proceedings of EMNLP*. 2369–2380.
- Yao, J., Z. Dou, R. Xie, Y. Lu, Z. Wang, and J.-R. Wen. (2021). “USER: A Unified Information Search and Recommendation Model based on Integrated Behavior Sequence”. In: *Proceedings of CIKM*. 2373–2382.
- Yao, K., B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi. (2014). “Spoken Language Understanding using Long Short-term Memory Neural Networks”. In: *Proceedings of SLT*. 189–194.

- Yao, K., G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu. (2013). “Recurrent Neural Networks for Language Understanding”. In: *Proceedings of Interspeech*. 2524–2528.
- Yao, Q., X. Chen, J. T. Kwok, Y. Li, and C. Hsieh. (2020). “Efficient Neural Interaction Function Search for Collaborative Filtering”. In: *Proceedings of Web Conference*. 1660–1670.
- Yao, S., J. Tan, X. Chen, J. Zhang, X. Zeng, and K. Yang. (2022). “ReprBERT: Distilling BERT to an Efficient Representation-Based Relevance Model for E-Commerce”. In: *Proceedings of KDD*. 4363–4371.
- Yasui, S., G. Morishita, F. Komei, and M. Shibata. (2020). “A Feedback Shift Correction in Predicting Conversion Rates under Delayed Feedback”. In: *Proceedings of Web Conference*. 2740–2746.
- Ye, C., L. Liao, F. Feng, W. Ji, and T.-S. Chua. (2022). “Structured and Natural Responses Co-generation for Conversational Search”. In: *Proceedings of SIGIR*. 155–164.
- Yee, K.-P., K. Swearingen, K. Li, and M. Hearst. (2003). “Faceted Metadata for Image Search and Browsing”. In: *Proceedings of CHI*. 401–408.
- Yi, J., F. Wu, C. Wu, Q. Li, G. Sun, and X. Xie. (2021). “DebiasedRec: Bias-aware User Modeling and Click Prediction for Personalized News Recommendation”. *arXiv preprint arXiv:2104.07360*.
- Yi, X., L. Hong, E. Zhong, N. N. Liu, and S. Rajan. (2014). “Beyond Clicks: Dwell Time for Personalization”. In: *Proceedings of RecSys*. 113–120.
- Yi, X., J. Yang, L. Hong, D. Z. Cheng, L. Heldt, A. Kumthekar, Z. Zhao, L. Wei, and E. Chi. (2019). “Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations”. In: *Proceedings of RecSys*. 269–277.
- Yilmaz, E., M. Shokouhi, N. Craswell, and S. Robertson. (2010). “Expected Browsing Utility for Web Search Evaluation”. In: *Proceedings of CIKM*. 1561–1564.
- Yin, H., B. Cui, L. Chen, Z. Hu, and Z. Huang. (2014). “A Temporal Context-Aware Model for User Behavior Modeling in Social Media Systems”. In: *Proceedings of SIGMOD*. 1543–1554.

- Yin, H., B. Cui, L. Chen, Z. Hu, and X. Zhou. (2015). “Dynamic User Modeling in Social Media Systems”. *ACM Transactions on Information Systems (TOIS)*. 33(3): 10.
- Yin, J., X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li. (2016). “Neural Generative Question Answering”. In: *Proceedings of IJCAI*. 2972–2978.
- Yin, P., P. Luo, W.-C. Lee, and M. Wang. (2013). “Silence is Also Evidence: Interpreting Dwell Time for Recommendation from Psychological Perspective”. In: *Proceedings of KDD*. 989–997.
- Yin, Y., L. Shang, X. Jiang, X. Chen, and Q. Liu. (2020). “Dialog State Tracking with Reinforced Data Augmentation”. In: *Proceedings of AAAI*. 9474–9481.
- Ying, R., R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec. (2018). “Graph Convolutional Neural Networks for Web-Scale Recommender Systems”. In: *Proceedings of KDD*. 974–983.
- Yoshikawa, Y. and Y. Imai. (2018). “A Nonparametric Delayed Feedback Model for Conversion Rate Prediction”. *arXiv preprint arXiv:1802.00255*.
- Young, S., M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. (2010). “The Hidden Information State model: A Practical Framework for POMDP-based Spoken Dialogue Management”. *Computer Speech & Language*. 24(2): 150–174.
- Young, S., M. Gašić, B. Thomson, and J. D. Williams. (2013). “POMDP-Based Statistical Spoken Dialog Systems: A Review”. *Proceedings of IEEE*. 101(5): 1160–1179.
- Young, T., E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang. (2018). “Augmenting End-to-End Dialogue Systems With Commonsense Knowledge”. In: *Proceedings of AAAI*. 4970–4977.
- Young Kim, E. and Y.-K. Kim. (2004). “Predicting Online Purchase Intentions for Clothing Products”. *European Journal of Marketing*. 38(7): 883–897.
- Yu, A. W., D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. (2018a). “QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension”. In: *Proceedings of ICLR*.

- Yu, D., C. Zhu, Y. Fang, W. Yu, S. Wang, Y. Xu, X. Ren, Y. Yang, and M. Zeng. (2022a). “KG-FiD: Infusing Knowledge Graph in Fusion-in-decoder for Open-domain Question Answering”. In: *Proceedings of ACL*. 4961–4974.
- Yu, F., Q. Liu, S. Wu, L. Wang, and T. Tan. (2016). “A Dynamic Recurrent Model for Next Basket Recommendation”. In: *Proceedings of SIGIR*. ACM. 729–732.
- Yu, H.-F., K. Zhong, J. Zhang, W.-C. Chang, and I. S. Dhillon. (2022b). “Pecos: Prediction for Enormous and Correlated Output Spaces”. *Journal of Machine Learning Research*. 23(98): 1–32.
- Yu, J., M. Qiu, J. Jiang, J. Huang, S. Song, W. Chu, and H. Chen. (2018b). “Modelling Domain Relationships for Transfer Learning on Retrieval-based Question Answering Systems in E-commerce”. In: *Proceedings of WSDM*. 682–690.
- Yu, J., Z.-J. Zha, and T.-S. Chua. (2012). “Answering Opinion Questions on Products by Exploiting Hierarchical Organization of Consumer Reviews”. In: *Proceedings of EMNLP-CoNLL*. 391–401.
- Yu, J., Z.-J. Zha, M. Wang, K. Wang, and T.-S. Chua. (2011). “Domain-Assisted Product Aspect Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews”. In: *Proceedings of EMNLP*. 140–150.
- Yu, J., X. Zhang, Y. Xu, X. Lei, X. Guan, J. Zhang, L. Hou, J. Li, and J. Tang. (2022c). “XDAI: A Tuning-free Framework for Exploiting Pre-trained Language Models in Knowledge Grounded Dialogue Generation”. In: *Proceedings of KDD*. 4422–4432.
- Yu, L., L. Sun, B. Du, C. Liu, H. Xiong, and W. Lv. (2020). “Predicting Temporal Sets with Deep Neural Networks”. In: *Proceedings of KDD*. 1083–1091.
- Yu, Q. and W. Lam. (2018). “Review-Aware Answer Prediction for Product-Related Questions Incorporating Aspects”. In: *Proceedings of WSDM*. 691–699.
- Yu, T., Y. Shen, and H. Jin. (2019). “A Visual Dialog Augmented Interactive Recommender System”. In: *Proceedings of KDD*. 157–165.

- Yuan, C., Y. Qiu, M. Li, H. Hu, S. Wang, and S. Xu. (2023). “A Multi-Granularity Matching Attention Network for Query Intent Classification in E-commerce Retrieval”. In: *Proceedings of Web Conference*. 416–420.
- Yuan, C., W. Zhou, M. Li, S. Lv, F. Zhu, J. Han, and S. Hu. (2019). “Multi-hop Selector Network for Multi-turn Response Selection in Retrieval-based Chatbots”. In: *Proceedings of EMNLP-IJCNLP*. 111–120.
- Yuan, J., W. Ji, D. Zhang, J. Pan, and X. Wang. (2022a). “Micro-Behavior Encoding for Session-based Recommendation”. In: *Proceedings ICDE*.
- Yuan, S., X. Shen, Y. Zhao, H. Liu, Z. Yan, R. Liu, and M. Chen. (2022b). “MCIC: Multimodal Conversational Intent Classification for E-commerce Customer Service”. In: *Proceedings of NLPCC*. 749–761.
- Yue, Y. and T. Joachims. (2009). “Interactively Optimizing Information Retrieval Systems as a Dueling Bandits Problem”. In: *Proceedings of ICML*. 1201–1208.
- Zamani, H., S. T. Dumais, N. Craswell, P. N. Bennett, and G. Lueck. (2020). “Generating Clarifying Questions for Information Retrieval”. In: *Proceedings of Web Conference*. 418–428.
- Zamani, H., J. R. Trippas, J. Dalton, and F. Radlinski. (2022). “Conversational Information Seeking”. *arXiv preprint arXiv:2201.08808*.
- Zang, H. and X. Wan. (2017). “Towards Automatic Generation of Product Reviews from Aspect-Sentiment Scores”. In: *Proceedings of INLG*. 168–177.
- Zehlike, M., F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. (2017). “FA*IR: A Fair Top-k Ranking Algorithm”. In: *Proceedings of CIKM*. 1569–1578.
- Zhang, H., S. Wang, K. Zhang, Z. Tang, Y. Jiang, Y. Xiao, W. Yan, and W.-Y. Yang. (2020a). “Towards Personalized and Semantic Retrieval: An End-to-End Solution for E-Commerce Search via Embedding Learning”. In: *Proceedings of SIGIR*. 2407–2416.

- Zhang, H., Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. (2013). “Attribute-augmented Semantic Hierarchy: Towards Bridging Semantic Gap and Intention Gap in Image Retrieval”. In: *Proceedings of MM*. 33–42.
- Zhang, H., Y. Zhang, L.-M. Zhan, J. Chen, G. Shi, X.-M. Wu, and A. Lam. (2021a). “Effectiveness of Pre-training for Few-shot Intent Classification”. In: *Findings of EMNLP*. 1114–1120.
- Zhang, J., P. Zou, Z. Li, Y. Wan, X. Pan, Y. Gong, and S. Y. Philip. (2019a). “Multi-Modal Generative Adversarial Network for Short Product Title Generation in Mobile E-Commerce”. In: *Proceedings of NAACL*. 64–72.
- Zhang, S., L. Yao, A. Sun, and Y. Tay. (2019b). “Deep Learning Based Recommender System: A Survey and New Perspectives”. *ACM Computing Surveys*. 52(1): 1–38.
- Zhang, S. and K. Balog. (2020). “Evaluating Conversational Recommender Systems via User Simulation”. In: *Proceedings of KDD*. 1512–1520.
- Zhang, S., M.-C. Wang, and K. Balog. (2022). “Analyzing and Simulating User Utterance Reformulation in Conversational Recommender Systems”. In: *Proceedings of SIGIR*. 133–143.
- Zhang, T., J. Zhang, C. Huo, and W. Ren. (2019c). “Automatic Generation of Pattern-Controlled Product Description in E-Commerce”. In: *Proceedings of Web Conference*. 2355–2365.
- Zhang, W., T. Du, and J. Wang. (2016a). “Deep Learning over Multi-field Categorical Data”. In: *Proceedings of ECIR*. 45–57.
- Zhang, W., J. Qin, W. Guo, R. Tang, and X. He. (2021b). “Deep Learning for Click-through Rate Estimation”. *arXiv preprint arXiv:2104.10584*.
- Zhang, W., W. Bao, X.-Y. Liu, K. Yang, Q. Lin, H. Wen, and R. Ramezani. (2020b). “Large-scale Causal Approaches to Debiasing Post-click Conversion Rate Estimation with Multi-task Learning”. In: *Proceedings of Web Conference*. 2775–2781.
- Zhang, W., Y. Deng, and W. Lam. (2020c). “Answer Ranking for Product-Related Questions via Multiple Semantic Relations Modeling”. In: *Proceedings of SIGIR*. 569–578.

- Zhang, W., Y. Deng, J. Ma, and W. Lam. (2020d). “AnswerFact: Fact Checking in Product Question Answering”. In: *Proceedings of EMNLP*. 2407–2417.
- Zhang, W., Q. Yu, and W. Lam. (2020e). “Answering Product-related Questions with Heterogeneous Information”. In: *Proceedings of ACL-IJCNLP*. 696–705.
- Zhang, X., H. Xie, H. Li, and J. CS Lui. (2020f). “Conversational Contextual Bandit: Algorithm and Application”. In: *Proceedings of Web Conference*. 662–672.
- Zhang, X., X. Xin, D. Li, W. Liu, P. Ren, Z. Chen, J. Ma, and Z. Ren. (2023). “Variational Reasoning over Incomplete Knowledge Graphs for Conversational Recommendation”. In: *Proceedings of WSDM*. 231–239.
- Zhang, X., Y. Jiang, Y. Shang, Z. Cheng, C. Zhang, X. Fan, Y. Xiao, and B. Long. (2021c). “DSGPT: Domain-Specific Generative Pre-Training of Transformers for Text Generation in E-Commerce Title and Review Summarization”. In: *Proceedings of SIGIR*. 2146–2150.
- Zhang, Y., F. Feng, C. Wang, X. He, M. Wang, Y. Li, and Y. Zhang. (2020g). “How to Retrain Recommender System?: A Sequential Meta-Learning Method”. In: *Proceedings of SIGIR*. 1479–1488.
- Zhang, Y., P. Ren, and M. de Rijke. (2021d). “A Human-machine Collaborative Framework for Evaluating Malevolence in Dialogues”. In: *Proceedings of ACL-IJCNLP*. 5612–5623.
- Zhang, Y., P. Pan, Y. Zheng, K. Zhao, Y. Zhang, X. Ren, and R. Jin. (2018a). “Visual Search at Alibaba”. In: *Proceedings of KDD*. 993–1001.
- Zhang, Y., H. Lu, W. Niu, and J. Caverlee. (2018b). “Quality-Aware Neural Complementary Item Recommendation”. In: *Proceedings of RecSys*. 77–85.
- Zhang, Y., S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan. (2020h). “DialoGPT: Large-scale Generative Pre-training for Conversational Response Generation”. In: *Proceedings of ACL*. 270–278.
- Zhang, Y., X. Chen, Q. Ai, L. Yang, and W. B. Croft. (2018c). “Towards Conversational Search and Recommendation: System Ask, User Respond”. In: *Proceedings of CIKM*. 177–186.

- Zhang, Y., G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. (2014). “Explicit Factor Models for Explainable Recommendation Based on Phrase-level Sentiment Analysis”. In: *Proceedings of SIGIR*. 83–92.
- Zhang, Y., Q. Zhao, Y. Zhang, D. Friedman, M. Zhang, Y. Liu, and S. Ma. (2016b). “Economic Recommendation with Surplus Maximization”. In: *Proceedings of Web Conference*. 73–83.
- Zhang, Y. and M. Pennacchiotti. (2013). “Predicting Purchase Behaviors from Social Media”. In: *Proceedings of Web Conference*. 1521–1532.
- Zhang, Y., D. Wang, and Y. Zhang. (2019d). “Neural IR Meets Graph Embedding: A Ranking Model for Product Search”. In: *Proceedings of Web Conference*. 2390–2400.
- Zhang, Y., W. Chen, D. Wang, and Q. Yang. (2011). “User-click Modeling for Understanding and Predicting Search-behavior”. In: *Proceedings of KDD*. 1388–1396.
- Zhang, Z., M. Huang, Z. Zhao, F. Ji, H. Chen, and X. Zhu. (2019e). “Memory-augmented Dialogue Management for Task-oriented Dialogue Systems”. *ACM Transactions on Information Systems*. 37(3): 1–30.
- Zhang, Z., R. Takanobu, M. Huang, and X. Zhu. (2020i). “Recent Advances and Challenges in Task-oriented Dialog System”. *ArXiv*. abs/2003.07490.
- Zhang, Z., J. Li, P. Zhu, H. Zhao, and G. Liu. (2018d). “Modeling Multi-turn Conversation with Deep Utterance Aggregation”. In: *Proceedings of COLING*. 3740–3752.
- Zhao, J., Z. Guan, and H. Sun. (2019a). “Riker: Mining Rich Keyword Representations for Interpretable Product Question Answering”. In: *Proceedings of KDD*. 1389–1398.
- Zhao, K., Y. Zheng, T. Zhuang, X. Li, and X. Zeng. (2022a). “Joint Learning of E-commerce Search and Recommendation with a Unified Graph Neural Network”. In: *Proceedings of WSDM*. 1461–1469.
- Zhao, M., Y. Yang, M. Li, J. Wang, W. Wu, P. Ren, M. de Rijke, and Z. Ren. (2022b). “Personalized Abstractive Opinion Tagging”. In: *Proceedings of SIGIR*. 1066–1076.
- Zhao, M. (2020). “Data-driven Scene Marketing Based on Consumer Insight”. In: *Proceedings of BDEIM*. 61–65.

- Zhao, N., H. Li, Y. Wu, X. He, and B. Zhou. (2021a). “The JDDC 2.0 Corpus: A Large-Scale Multimodal Multi-Turn Chinese Dialogue Dataset for E-commerce Customer Service”. *arXiv preprint arXiv:2109.12913*.
- Zhao, Q., Y. Zhang, Y. Zhang, and D. Friedman. (2017). “Multi-Product Utility Maximization for Economic Recommendation”. In: *Proceedings of WSDM*. 435–443.
- Zhao, T. and M. Eskenazi. (2016). “Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning”. In: *Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1.
- Zhao, X., H. Liu, W. Fan, H. Liu, J. Tang, and C. Wang. (2021b). “AutoLoss: Automated Loss Function Search in Recommendations”. In: *Proceedings of KDD*. 3959–3967.
- Zhao, X., L. Wang, R. He, T. Yang, J. Chang, and R. Wang. (2020a). “Multiple Knowledge Syncretic Transformer for Natural Dialogue Generation”. In: *Proceedings of Web Conference*. 752–762.
- Zhao, X., L. Xia, Z. Ding, D. Yin, and J. Tang. (2019b). “Toward Simulating Environments in Reinforcement Learning Based Recommendations”. *arXiv preprint arXiv:1906.11462*.
- Zhao, X., W. Zhang, and J. Wang. (2013). “Interactive Collaborative Filtering”. In: *Proceedings of CIKM*. 1411–1420.
- Zhao, X., C. Tao, W. Wu, C. Xu, D. Zhao, and R. Yan. (2019c). “A Document-grounded Matching Network for Response Selection in Retrieval-based Chatbots”. In: *Proceedings of IJCAI*. 5443–5449.
- Zhao, X., W. Wu, C. Tao, C. Xu, D. Zhao, and R. Yan. (2019d). “Low-Resource Knowledge-Grounded Dialogue Generation”. In: *Proceedings of ICLR*.
- Zhao, X., W. Wu, C. Xu, C. Tao, D. Zhao, and R. Yan. (2020b). “Knowledge-Grounded Dialogue Generation with Pre-trained Language Models”. In: *Proceedings of EMNLP*. 3377–3390.
- Zhao, Y., W. Wu, and C. Xu. (2020c). “Are Pre-trained Language Models Knowledgeable to Ground Open Domain Dialogues?” *arXiv preprint arXiv:2011.09708*.

- Zhao, Y., J. Qi, Q. Liu, and R. Zhang. (2021c). “WGCN: Graph Convolutional Networks with Weighted Structural Features”. In: *Proceedings of SIGIR*. 624–633.
- Zheng, C., Y. Cao, D. Jiang, and M. Huang. (2020a). “Difference-aware Knowledge Selection for Knowledge-grounded Conversation Generation”. In: *Proceedings of EMNLP*. 115–125.
- Zheng, W., N. Milic-Frayling, and K. Zhou. (2020b). “Approximation of Response Knowledge Retrieval in Knowledge-grounded Dialogue Generation”. In: *Proceedings of EMNLP*. 3581–3591.
- Zheng, W. and K. Zhou. (2019). “Enhancing Conversational Dialogue Models with Grounded Knowledge”. In: *Proceedings of CIKM*. 709–718.
- Zhong, F., D. Wang, G. Wang, W. Chen, Y. Zhang, Z. Chen, and H. Wang. (2010). “Incorporating Post-click Behaviors into a Click Model”. In: *Proceedings of SIGIR*. 355–362.
- Zhong, M., Y. Liu, Y. Xu, C. Zhu, and M. Zeng. (2022). “DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization”. In: *Proceedings of AAAI*. 11765–11773.
- Zhong, P., D. Wang, and C. Miao. (2019). “An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss”. In: *Proceedings of AAAI*. 7492–7500.
- Zhong, P., C. Zhang, H. Wang, Y. Liu, and C. Miao. (2020). “Towards Persona-Based Empathetic Conversational Models”. In: *Proceedings of EMNLP*. 6556–6566.
- Zhou, G., N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai. (2019). “Deep Interest Evolution Network for Click-through Rate Prediction”. In: *Proceedings of AAAI*. Vol. 33. 5941–5948.
- Zhou, G., X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai. (2018a). “Deep Interest Network for Click-through Rate Prediction”. In: *Proceedings of KDD*. 1059–1068.
- Zhou, H., M. Huang, *et al.* (2016). “Context-aware Natural Language Generation for Spoken Dialogue Systems”. In: *Proceedings of COLING*. 2032–2041.
- Zhou, H., M. Huang, T. Zhang, X. Zhu, and B. Liu. (2018b). “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory”. In: *Proceedings of AAAI*. 730–739.

- Zhou, H., T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu. (2018c). “Commonsense Knowledge Aware Conversation Generation with Graph Attention”. In: *Proceedings of IJCAI*. 4623–4629.
- Zhou, H., C. Zheng, K. Huang, M. Huang, and X. Zhu. (2020a). “Kd-Conv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation”. In: *Proceedings of ACL*. 7098–7108.
- Zhou, K., S. Prabhunoye, and A. W. Black. (2018d). “A Dataset for Document Grounded Conversations”. In: *Proceedings of EMNLP*. 708–713.
- Zhou, K., W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, and J. Yu. (2020b). “Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion”. In: *Proceedings of KDD*. 1006–1014.
- Zhou, K., Y. Zhou, W. X. Zhao, X. Wang, and J.-R. Wen. (2020c). “Towards Topic-Guided Conversational Recommender System”. In: *Proceedings of COLING*. 4128–4139.
- Zhou, L., J. Gao, D. Li, and H.-Y. Shum. (2020d). “The Design and Implementation of XiaoIce, an Empathetic Social Chatbot”. *Computational Linguistics*. 46(1): 53–93.
- Zhou, L. and K. Small. (2019). “Multi-domain Dialogue State Tracking as Dynamic Knowledge Graph Enhanced Question Answering”. *arXiv preprint arXiv:1911.06192*.
- Zhou, M., Z. Ding, Z. Jiang, and D. Yin. (2018e). “Micro Behaviors: A New Perspective in E-commerce Recommender Systems”. In: *Proceedings of WSDM*. 727–735.
- Zhou, X. and W. Y. Wang. (2018). “Mojitalk: Generating Emotional Responses at Scale”. In: *Proceedings of ACL*. 1128–1137.
- Zhou, X., L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu. (2018f). “Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network”. In: *Proceedings of ACL*. 1118–1127.
- Zhou, X. and R. Zafarani. (2020). “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”. *ACM Computing Surveys*. 53(5): 1–40.

- Zhou, Y., K. Zhou, W. X. Zhao, C. Wang, P. Jiang, and H. Hu. (2022a). “C²-CRS: Coarse-to-Fine Contrastive Learning for Conversational Recommender System”. In: *Proceedings of WSDM*. 1488–1496.
- Zhou, Y., J. Yao, Z. Dou, L. Wu, P. Zhang, and J.-R. Wen. (2022b). “Ultron: An Ultimate Retriever on Corpus with a Model-based Indexer”. *arXiv preprint arXiv:2208.09257*.
- Zhu, C., B. Chen, W. Zhang, J. Lai, R. Tang, X. He, Z. Li, and Y. Yu. (2021a). “AIM: Automatic Interaction Machine for Click-Through Rate Prediction”. *IEEE Transactions on Knowledge and Data Engineering*.
- Zhu, H., J. Jin, C. Tan, F. Pan, Y. Zeng, H. Li, and K. Gai. (2017). “Optimized Cost per Click in Taobao Display Advertising”. In: *Proceedings of KDD*. 2191–2200.
- Zhu, H., X. Li, P. Zhang, G. Li, J. He, H. Li, and K. Gai. (2018). “Learning Tree-Based Deep Model for Recommender Systems”. In: *Proceedings of KDD*. 1079–1088.
- Zhu, J., Q. Dai, L. Su, R. Ma, J. Liu, G. Cai, X. Xiao, and R. Zhang. (2022a). “Bars: Towards Open Benchmarking for Recommender Systems”. In: *Proceedings of SIGIR*. 2912–2923.
- Zhu, J., J. Liu, S. Yang, Q. Zhang, and X. He. (2021b). “Open Benchmarking for Click-through Rate Prediction”. In: *Proceedings of CIKM*. 2759–2769.
- Zhu, L., X. Lu, Z. Cheng, J. Li, and H. Zhang. (2020a). “Deep Collaborative Multi-View Hashing for Large-Scale Image Search”. *IEEE Transactions on Image Processing*. 29: 4643–4655.
- Zhu, R., Y. Zhao, W. Qu, Z. Liu, and C. Li. (2022b). “Cross-Domain Product Search with Knowledge Graph”. In: *Proceedings of CIKM*. 3746–3755.
- Zhu, S., J. Li, L. Chen, and K. Yu. (2020b). “Efficient Context and Schema Fusion Networks for Multi-Domain Dialogue State Tracking”. In: *Proceedings of EMNLP*. 766–781.
- Zhu, Y., L. Pang, Y. Lan, H. Shen, and X. Cheng. (2021c). “Adaptive Information Seeking for Open-Domain Question Answering”. *arXiv preprint arXiv:2109.06747*.
- Zoghi, M., S. Whiteson, Z. Karnin, and M. de Rijke. (2015). “Copeland Dueling Bandits”. In: *Proceedings of NIPS*. 307–315.

- Zou, J., J. X. Huang, Z. Ren, and E. Kanoulas. (2022a). “Learning to Ask: Conversational Product Search via Representation Learning”. *ACM Transactions on Information Systems*.
- Zou, J., E. Kanoulas, P. Ren, Z. Ren, A. Sun, and C. Long. (2022b). “Improving Conversational Recommender Systems via Transformer-based Sequential Modelling”. In: *Proceedings of SIGIR*. 2319–2324.
- Zou, L., L. Xia, Z. Ding, W. Liu, Y. Zhao, and D. Yin. (2020a). “Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems”. In: *Proceedings of KDD*. 95–103.
- Zou, L., L. Xia, Y. Gu, X. Zhao, W. Liu, J. X. Huang, and D. Yin. (2020b). “Neural Interactive Collaborative Filtering”. In: *Proceedings of SIGIR*. 749–758.
- Zou, L., S. Zhang, H. Cai, D. Ma, S. Cheng, S. Wang, D. Shi, Z. Cheng, and D. Yin. (2021). “Pre-trained Language Model based Ranking in Baidu Search”. In: *Proceedings of KDD*. 4014–4022.