## Information Discovery in E-commerce

*Half-day SIGIR 2018 tutorial*

Ren, Z.; He, X.; Yin, D.; de Rijke, M.

[Link to publication](#)

# Information Discovery in E-commerce

## Half-day SIGIR 2018 Tutorial

### Zhaochun Ren
Data Science Lab, JD.com
Beijing, China
renzhaochun@jd.com

### Xiangnan He
National University of Singapore
Singapore
xiangnanhe@gmail.com

### Dawei Yin
Data Science Lab, JD.com
Beijing, China
yindawei@acm.org

### Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

## EXTENDED ABSTRACT

E-commerce (electronic commerce or EC) 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, eBay, JD.com and platforms targeting specific markets such as Bol.com and Booking.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 (exploratory search vs. lookup tasks), recommender systems, and natural language processing in e-commerce portals. Recently, the explosive popularity of e-commerce sites has made research on information discovery in e-commerce more important and more popular. There is increased attention for e-commerce information discovery methods in the community as witnessed by an increase in publications and dedicated workshops in this space. Methods for information discovery in e-commerce largely focus on improving the performance of e-commerce search and recommender systems, on enriching and using 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.

Below we describe why we believe that the time is right for an introductory tutorial on information discovery in e-commerce, the objectives of the proposed tutorial, its relevance, as well as more practical details, such as the format, schedule and support materials.

## 1 MOTIVATION

In recent years, the explosive popularity of e-commerce sites has re-shaped users' shopping habits. An increasing number of customers now prefer to spend more time shopping online. E-commerce corporations, e.g., Amazon, Alibaba, and JD.com, are amassing billions of user requests per day. As part of this process, 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 appropriate information from the sheer volume of e-commerce data to enhance performances of e-commerce products presents interesting challenges for both academic and industrial researchers.

Information discovery in e-commerce can be divided into five main directions:

- e-commerce user behavior modeling and profiling
- content analysis of e-commerce text,
- e-commerce search and ranking,
- e-commerce recommender systems, and
- e-commerce conversational interaction systems.

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; there is a click stream, but there is also an order stream. E-commerce information discovery problems are wide in scope and range from user interaction modalities. There is a growing body of established methods for information discovery in e-commerce (see the schedule below for a broad range of examples). Most of them are aimed at developing algorithms about product search in e-commerce [1, 10], candidate retrieval in e-commerce [15, 50], user behavior analysis [52, 59], recommender systems [14, 21, 35, 37, 46], content analysis, and conversational interactions [4, 11, 18, 22, 29, 29, 39]. These areas, and the methods developed, form the core around which most ongoing research efforts concerning information discovery for e-commerce is organized.

The time is right to organize and present this material to a broad audience of interested information retrieval researchers, whether junior or senior, whether academic or industrial. One of the key aims of the proposed tutorial is to bring together, and offer a unified perspective on, the large number of methods for e-commerce information discovery that are available today. To achieve this, we describe the basic architecture about information discovery in e-commerce, algorithms for e-commerce information discovery, and evaluation principles. We supplement this with an account of available datasets and packages based on these. We also present e-commerce applications accompanied by examples.

We expect the tutorial to be useful for 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 tutorial to improve product performance in e-commerce services.

## 2 OBJECTIVES

Information discovery 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 the traditional web-page ranking features are either not present or are present in a different form [9]; instead, discovery processes need to be supported based on unstructured information, structured information, semi-structured, or information that might have facets such as price, ratings, title, description, seller location, and so on.

Our treatment information discovery in e-commerce is organized in four groups. The first group concerns two main areas: (1) modeling user behavior and user profiling, and (2) content analysis in an e-commerce setting: from text, OCR files, and item images material; multiple content analysis tasks are addressed to enhance e-commerce services, e.g., review filtering. This group of topics and techniques prepares us for the next three groups.

Then, the second group of topics that make up this tutorial concerns e-commerce search. 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 achieved with new types of features, and the concept of relevance can be highly personal in product search. And, as pointed out above, the target corpora can structured, semi-structured, or unstructured; semantic search against diverse sources raises interesting research challenges.

The third group of topics covered in this tutorial concerns e-commerce recommendations. In contrast to traditional research on recommender systems that focuses on rating prediction, e-commerce recommender systems usually aim to optimize the top-$N$ online recommendation results based on implicit feedback. Due to the existence of a very large number of candidate items in e-commerce portals, of which only a small fraction attracts users' attention, personalized e-commerce recommendation procedure usually consist of two procedures. The first is to generate candidate items. The second is to rank candidate items. 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 [6, 13–15].

The fourth group of topics direction of this tutorial concerns conversational interaction systems in e-commerce. Conversational interaction helps various e-commerce applications, e.g., customer service, to interact with humans in an intelligent natural way. The target of conversational interaction systems is to generate proper utterances given previous utterances, which raises research challenges for both task-oriented and non-task-oriented scenarios.

### 2.1 Specific goals

This tutorial targets practitioners and researchers from academia and industry and aims to present them with the challenges, state-of-the-art approaches, and most urgent open questions in information discovery for e-commerce. Specifically, in terms of content, the objectives of the proposed tutorial 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.

- To describe existing 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 the importance of balancing exploration and exploitation in information discovery in an e-commerce setting.
- 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.
- To describe how to deploy e-commerce information discovery algorithms in an industrial setting.
- To discuss future directions of research in e-commerce information discovery.

### 2.2 Topics not covered as part of the tutorial

E-commerce impacts large parts of our economy and society, including markets and retailers, supply chain management, employment. While all of these are important, scientifically challenging, and deserving of attention from the information retrieval community, in the proposed tutorial we will restrict ourselves to information discovery in the context of e-commerce. In particular, we will not cover computational advertising, marketing strategies or information mangement in e-commerce.

## 3 FORMAT AND DETAILED SCHEDULE

The tutorial is organized into four parts, each mixing theoretical principles and experimental outcomes, with formal analyses of e-commerce information discovery methods interleaved with discussions of experimental outcomes. In Part I we aim at providing preliminaries of e-commerce information discovery, including user behavior modeling, profiling, and content analysis in e-commerce. Part II is aimed at e-commerce search and ranking and deals with the key concepts and algorithms. In Part III we focus on the recommender systems for e-commerce, and select a small number of topics for which we provide a more in-depth technical treatment. In Part IV we discuss conversational interaction systems in e-commerce portals.

### 3.1 Part I: Preliminaries

**[5min]** Introduction, aims and historical notes about e-commerce

- Here we first discuss the context in which e-commerce information discovery is applied and the most important historical milestones in its development. We describe the four parts of the tutorial, and introduce each part individually.

**[20min]** User behavior modeling and profiling

- The unique characteristics of e-commerce search make personalization essential. To discover users behaviors in e-commerce portals, we describe recent research on user behavior modeling, including post-click behaviors tracking [21, 51, 52], purchasing behavior modeling [19, 25, 40], and micro user behavior modeling [59].

**[25min]** Content analysis in e-commerce

- Here we discuss the context in which content analysis with e-commerce text, OCRs, and images is applied and the most important historical milestones in its development.

- We discuss recent work on knowledge graph construction given massive unstructured data generated in e-commerce sites.
- We introduce review analysis in e-commerce, e.g., review quality analysis [28, 29, 55], review summarization and filtering [18, 22], and explainable recommendation based on review analysis [33, 56].
- We introduce knowledge harvesting and other related research from unstructured e-commerce data [16, 17, 58], e.g., generating titles and keywords of products [47], etc.

## 3.2  Part II: Search and Ranking

**[5min]** Introduction

- We first introduce the basic architecture of e-commerce retrieval systems, with several real-world examples. Specifically, we focus on the product search models in e-commerce retrieval systems.

**[25min]** Product search models in e-commerce

- We describe existing approaches for product search, starting with a basic probabilistic mixture model [57] and arriving at more sophisticated neural embedding models [1, 43].
- We discuss the difference between product search in e-commerce portals and other traditional ad hoc web-search approaches. We analyze structured aspects in product search, the gap between the language of product descriptions and free-form user queries, modelÂÃÂŹ assumptions, observed and hidden random variables, and model parameters.
- The discussion is concluded with a summary of differences and similarities between the presented product search models.
- We outline directions for future research on e-commerce search.

## 3.3  Part III: Recommendation

**[10min]** Introduction, aims and historical notes about e-commerce recommendation

- Here we discuss the context in which e-commerce recommendation research is applied and the most important historical milestones in its development.
- We then describe the gap between traditional recommendation algorithms and models for e-commerce recommendation.

**[25min]** Candidate product retrieval in e-commerce

- Candidate product retrieval refers to a process that retrieves personalized candidates out of billions of items, which provides relevant products for e-commerce search and e-commerce recommendation. We describe several strategies for retrieving personalized product candidates, especially for distinguishing substitutable and complementary products [50, 57].
- We describe how to employ network embeddings to generate relevant product pairs or relevant communities to enhance the performance of e-commerce candidate retrieval. We start with traditional network embeddings, such as Deepwalk [30] and LINES [41], and arrive at state-of-the-art work on multi-dimension network embeddings [27] and heterogeneous embeddings [2, 32, 45].

**[25min]** Recommendation models for e-commerce

- We first introduce an example of the basic architecture of e-commerce recommender systems. We then describe basic recommendation models, especially for several widely used models in an e-commerce setting, such as item-to-item collaborative filtering and latent factor models [34, 36], tree-based recommendation

models (e.g., XGBoost, etc.) [6, 48], and neural network recommendation models (e.g., wide and deep neural networks) [7]. We also demonstrate online performance of real-world e-commerce recommender systems based on these models.
- We describe state-of-the-art neural network based recommendation models, e.g., neural collaborative filtering [15] and neural factorization machines [12], etc. We also introduce approaches that address effectiveness and efficiency (e.g., eALS) [14], and multi-modal recommendation models [5, 49].
- We introduce recent work based on bandits and reinforcement learning models for online recommendation [38, 44].
- The discussion is concluded with a summary of existing recommendation models applied in e-commerce systems, and an outlook to future directions of e-commerce recommendation.

## 3.4  Part IV: Conversational Interaction

**[5min]** Introduction and aims

- Here we introduce conversational interaction systems for e-commerce services and the most important milestones in its development.

**[15min]** Question-answering in e-commerce

- We discuss research on question-answering in e-commerce [23, 39, 42, 53, 54], especially for those research on knowledge-based question-answering and non-factoid question-answering systems. Thereafter, we discuss how to employ those models to e-commerce services.

**[20min]** Dialog systems in e-commerce

- We discuss research on dialogue systems in e-commerce sites [20, 24].
- Task-oriented dialog systems has gained an increase of attention. We introduce recent work on task-oriented dialog systems [3, 8, 26, 31], and discuss how to employ those models to e-commerce custom service chatbots or dialog-based online recommender systems.
- We conclude with a summary of existing chatbots models applied in e-commerce, and outlook to future directions of e-commerce chatbots.

## 4  TYPE OF SUPPORT MATERIALS TO BE SUPPLIED TO ATTENDEES

**Slides**  All slides will be made publicly available.

**Survey**  The authors are writing a survey on information discovery in e-commerce; a complete draft will be shared with attendees.

**Bibliography**  An annotated compilation of references will list all works discussed in the tutorial and should provide a good basis for further study.

**Code**  Code and datasets used for demonstration purposes during the tutorial will be shared. In addition, a list of pointers to open source code and datasets about e-commerce recommendation and content analysis will be shared with attendees.

## 5  ACKNOWLEDGEMENTS

## REFERENCES

[1] Q. Ai, Y. Zhang, K. Bi, X. Chen, and W. B. Croft. Learning a hierarchical embedding model for personalized product search. In *SIGIR*, pages 645–654. ACM, 2017.

[2] S. Chang, W. Han, J. Tang, G.-J. Qi, C. C. Aggarwal, and T. S. Huang. Heterogeneous network embedding via deep architectures. In *KDD*, pages 119–128, ACM, 2015.

[3] H. Chen, X. Liu, D. Yin, and J. Tang. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2), 2017.

[4] H. Chen, Z. Ren, J. Tang, Y. E. Zhao, and D. Yin. Hierarchical variational memory network for dialogue generation. In *WWW*. ACM, 2018.

[5] J. Chen, H. Zhang, X. He, W. Liu, W. Liu, and T.-S. Chua. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In *SIGIR*, pages 335–344. ACM, 2017.

[6] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, pages 785–794. ACM, 2016.

[7] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 7–10. ACM, 2016.

[8] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou. Superagent: A customer service chatbot for e-commerce websites. In *ACL*, pages 97–102. ACL, 2017.

[9] J. Degenhardt, S. Kallumadi, M. de Rijke, L. Si, A. Trotman, and X. Yinghui. eCom: The SIGIR 2017 workshop on ecommerce. In *SIGIR*. ACM, 2017.

[10] H. Duan, C. Zhai, J. Cheng, and A. Gattani. Supporting keyword search in product database: a probabilistic approach. *Proceedings of the VLDB Endowment*, 6(14): 1786–1797, 2013.

[11] C. Geigle and C. Zhai. Scaling up online question answering via similar question retrieval. In *L@S*, pages 257–260. ACM, 2016.

[12] X. He and T.-S. Chua. Neural factorization machines for sparse predictive analytics. In *SIGIR*, pages 355–364. ACM, 2017.

[13] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *ADKDD*, pages 1–9. ACM, 2014.

[14] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua. Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR*, pages 549–558. ACM, 2016.

[15] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *WWW*, pages 173–182. ACM, 2017.

[16] Z. He, W. Chen, Z. Li, M. Zhang, W. Zhang, and M. Zhang. SEE: Syntax-aware entity embedding for neural relation extraction. In *AAAI*. AAAI Press, 2018.

[17] G. Ji, K. Liu, S. He, J. Zhao, et al. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *AAAI*, pages 3060–3066. AAAI Press, 2017.

[18] S. K. Karmaker Santu, P. Sondhi, and C. Zhai. Generative feature language models for mining implicit features from customer reviews. In *CIKM*, pages 929–938. ACM, 2016.

[19] A. Kumar, A. Biswas, and S. Sanyal. eCommerceGAN: A generative adversarial network for e-commerce. In *arXiv preprint arXiv:1801.03244*, 2018.

[20] W. Lei, X. Jin, Z. Ren, X. He, M.-Y. Kan, and D. Yin. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *ACL*, 2018.

[21] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, and J. Ma. Neural attentive session-based recommendation. In *CIKM*, pages 1419–1428. ACM, 2017.

[22] P. Li, Z. Wang, Z. Ren, L. Bing, and W. Lam. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*, pages 345–354, 2017.

[23] M. Liu, Y. Fang, D. H. Park, X. Hu, and Z. Yu. Retrieving non-redundant questions to summarize a product review. In *SIGIR*, pages 385–394. ACM, 2016.

[24] S. Liu, H. Chen, Z. Ren, Y. Feng, and D. Yin. Knowledge diffusion for neural dialogue generation. In *ACL*, 2018.

[25] C. Lo, D. Frankowski, and J. Leskovec. Understanding behaviors that lead to purchasing: A case study of Pinterest. In *KDD*, pages 531–540. ACM, 2016.

[26] X. Luo, Z. Lin, Y. Wang, and Z. Nie. CoChat: Enabling bot and human collaboration for task completion. In *AAAI*, 2018.

[27] Y. Ma, Z. Ren, Z. Jiang, J. Tang, and D. Yin. Multi-dimensional network embedding with hierarchical structures. In *WSDM*, pages 387–395. ACM, 2018.

[28] S. Mukherjee, K. Popat, and G. Weikum. Exploring latent semantic factors to find useful product reviews. In *SDM*, pages 480–488. SIAM, 2017.

[29] D. H. Park, C. Zhai, and L. Guo. Speclda: Modeling product reviews and specifications to generate augmented specifications. In *SDM*, pages 837–845. SIAM, 2015.

[30] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *KDD*, pages 701–710. ACM, 2014.

[31] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu. Alime chat: A sequence to sequence and rerank based chatbot engine. In *ACL*, pages 498–503. ACL, 2017.

[32] M. Qu, J. Tang, and J. Han. Curriculum learning for heterogeneous star network embedding via deep reinforcement learning. In *WSDM*, pages 468–476. ACM, 2018.

[33] Z. Ren, S. Liang, P. Li, S. Wang, and M. de Rijke. Social collaborative viewpoint regression with explainable recommendations. In *WSDM*, pages 485–494. ACM, 2017.

[34] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2011.

[35] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *EC*, pages 158–167. ACM, 2000.

[36] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295. ACM, 2001.

[37] J. B. Schafer, J. Konstan, and J. Riedl. Recommender systems in e-commerce. In *EC*, pages 158–166. ACM, 1999.

[38] T. Schnabel, P. N. Bennett, S. T. Dumais, and T. Joachims. Using shortlists to support decision making and improve recommender system performance. In *WWW*, pages 987–997. ACM, 2016.

[39] H. Song, Z. Ren, S. Liang, P. Li, J. Ma, and M. de Rijke. Summarizing answers in non-factoid community question-answering. In *WSDM*, pages 405–414. ACM, 2017.

[40] A. Tamhane, S. Arora, and D. Warrier. Modeling contextual changes in user behaviour in fashion e-commerce. In *PAKDD*, pages 539–550. ACM, 2017.

[41] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *WWW*, pages 1067–1077. ACM, 2015.

[42] A. G. Tapeh and M. Rahgozar. A knowledge-based question answering system for b2c ecommerce. *Knowledge-Based Systems*, 21(8):946–950, 2008.

[43] C. Van Gysel, M. de Rijke, and E. Kanoulas. Learning latent vector spaces for product search. In *CIKM*, pages 165–174. ACM, 2016.

[44] H. Wang, Q. Wu, and H. Wang. Factorization bandits for interactive recommendation. In *AAAI*, pages 2695–2702. AAAI Press, 2017.

[45] H. Wang, F. Zhang, M. Hou, X. Xie, M. Guo, and Q. Liu. Shine: Signed heterogeneous information network embedding for sentiment link prediction. In *WSDM*, pages 592–600. ACM, 2018.

[46] J. Wang, B. Sarwar, and N. Sundaresan. Utilizing related products for post-purchase recommendation in e-commerce. In *RecSys*, pages 329–332. ACM, 2011.

[47] J. Wang, J. Tian, L. Qiu, S. Li, J. Lang, L. Si, and M. Lan. A multi-task learning approach for improving product title compression with user search log data. In *AAAI*. AAAI Press, 2018.

[48] X. Wang, X. He, F. Feng, L. Nie, and T.-S. Chua. Tem: Tree-enhanced embedding model for explainable recommendation. In *WWW*, pages 1543–1552. ACM, 2018.

[49] Y. Wang, J. Xu, A. Wu, M. Li, Y. He, J. Hu, and W. P. Yan. Telepath: Understanding users from a human vision perspective in large-scale recommender systems. In *AAAI*. AAAI Press, 2018.

[50] Z. Wang, Z. Jiang, Z. Ren, J. Tang, and D. Yin. A path-constrained framework for discriminating substitutable and complementary products in e-commerce. In *WSDM*, pages 619–627. ACM, 2018.

[51] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan. Beyond clicks: dwell time for personalization. In *RecSys*, pages 113–120. ACM, 2014.

[52] D. Yin, L. Hong, and B. D. Davison. Exploiting session-like behaviors in tag prediction. In *WWW*, pages 167–168. ACM, 2011.

[53] J. Yin, X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li. Neural generative question answering. In *IJCAI*, pages 2972–2978. IJCAI, 2016.

[54] J. Yu, M. Qiu, J. Jiang, J. Huang, S. Song, W. Chu, and H. Chen. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. *arXiv preprint arXiv:1711.08726*, 2017.

[55] J. Zhang, M. Ren, X. Xiao, and J. Zhang. Providing consumers with a representative subset from online reviews. *Online Information Review*, 41(6):877–899, 2017.

[56] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR*, pages 83–92. ACM, 2014.

[57] H. Zhao, L. Si, X. Li, and Q. Zhang. Recommending complementary products in e-commerce push notifications with a mixture model approach. In *SIGIR*, pages 909–912. ACM, 2017.

[58] C. Zhou, Y. Liu, X. Liu, Z. Liu, and J. Gao. Scalable graph embedding for asymmetric proximity. In *AAAI*, pages 2942–2948. AAAI Press, 2017.

[59] M. Zhou, Z. Ding, Z. Jiang, and D. Yin. Micro behaviors: A new perspective in e-commerce recommender systems. In *WSDM*, pages 727–735. ACM, 2018.