Explainable Recommendation: A Survey and New Perspectives
Other titles in Foundations and Trends® in Information Retrieval

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

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

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

*An Introduction to Neural Information Retrieval*
Bhaskar Mitra and Nick Craswell
ISBN: 978-1-68083-532-8
Explainable Recommendation: A Survey and New Perspectives

Yongfeng Zhang
Rutgers University, USA
yongfeng.zhang@rutgers.edu

Xu Chen
Tsinghua University, China
xu-ch14@mails.tsinghua.edu.cn

Full text available at: http://dx.doi.org/10.1561/1500000066
Editorial Scope

Topics

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

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

Information for Librarians

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

1 Introduction ................................................. 3
   1.1 Explainable Recommendation ......................... 3
   1.2 A Historical Overview ................................ 4
   1.3 Classification of the Methods ......................... 7
   1.4 Explainability and Effectiveness ....................... 10
   1.5 Explainability and Interpretability .................... 10
   1.6 How to Read the Survey ............................... 11

2 Information Source for Explanations ....................... 13
   2.1 Relevant User or Item Explanation ..................... 14
   2.2 Feature-based Explanation .............................. 17
   2.3 Opinion-based Explanation .............................. 19
   2.4 Sentence Explanation .................................. 21
   2.5 Visual Explanation .................................... 24
   2.6 Social Explanation .................................... 26
   2.7 Summary .............................................. 28

3 Explainable Recommendation Models ....................... 29
   3.1 Overview of Machine Learning for Recommendation .... 30
   3.2 Factorization Models for Explainable Recommendation .... 31
   3.3 Topic Modeling for Explainable Recommendation ........ 35
   3.4 Graph-based Models for Explainable Recommendation .... 38
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5 Deep Learning for Explainable Recommendation</td>
<td>40</td>
</tr>
<tr>
<td>3.6 Knowledge Graph-based Explainable Recommendation</td>
<td>46</td>
</tr>
<tr>
<td>3.7 Rule Mining for Explainable Recommendation</td>
<td>49</td>
</tr>
<tr>
<td>3.8 Model Agnostic and Post Hoc Explainable Recommendation</td>
<td>51</td>
</tr>
<tr>
<td>3.9 Summary</td>
<td>54</td>
</tr>
<tr>
<td>4 Evaluation of Explainable Recommendation</td>
<td>56</td>
</tr>
<tr>
<td>4.1 User Study</td>
<td>57</td>
</tr>
<tr>
<td>4.2 Online Evaluation</td>
<td>59</td>
</tr>
<tr>
<td>4.3 Offline Evaluation</td>
<td>61</td>
</tr>
<tr>
<td>4.4 Qualitative Evaluation by Case Study</td>
<td>62</td>
</tr>
<tr>
<td>4.5 Summary</td>
<td>64</td>
</tr>
<tr>
<td>5 Explainable Recommendation in Different Applications</td>
<td>65</td>
</tr>
<tr>
<td>5.1 Explainable E-commerce Recommendation</td>
<td>65</td>
</tr>
<tr>
<td>5.2 Explainable Point-of-Interest Recommendation</td>
<td>66</td>
</tr>
<tr>
<td>5.3 Explainable Social Recommendation</td>
<td>68</td>
</tr>
<tr>
<td>5.4 Explainable Multimedia Recommendation</td>
<td>69</td>
</tr>
<tr>
<td>5.5 Other Explainable Recommendation Applications</td>
<td>70</td>
</tr>
<tr>
<td>5.6 Summary</td>
<td>71</td>
</tr>
<tr>
<td>6 Open Directions and New Perspectives</td>
<td>72</td>
</tr>
<tr>
<td>6.1 Methods and New Applications</td>
<td>72</td>
</tr>
<tr>
<td>6.2 Evaluation and User Behavior Analysis</td>
<td>77</td>
</tr>
<tr>
<td>6.3 Explanation for Broader Impacts</td>
<td>78</td>
</tr>
<tr>
<td>6.4 Cognitive Science Foundations</td>
<td>79</td>
</tr>
<tr>
<td>7 Conclusions</td>
<td>80</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>82</td>
</tr>
<tr>
<td>References</td>
<td>83</td>
</tr>
</tbody>
</table>
ABSTRACT
Explainable recommendation attempts to develop models that generate not only high-quality recommendations but also intuitive explanations. The explanations may either be post-hoc or directly come from an explainable model (also called interpretable or transparent model in some contexts). Explainable recommendation tries to address the problem of why: by providing explanations to users or system designers, it helps humans to understand why certain items are recommended by the algorithm, where the human can either be users or system designers. Explainable recommendation helps to improve the transparency, persuasiveness, effectiveness, trustworthiness, and satisfaction of recommendation systems. It also facilitates system designers for better system debugging. In recent years, a large number of explainable recommendation approaches – especially model-based methods – have been proposed and applied in real-world systems.

In this survey, we provide a comprehensive review for the explainable recommendation research. We first highlight the position of explainable recommendation in recommender system research by categorizing recommendation problems into the 5W, i.e., what, when, who, where, and why. We then

conduct a comprehensive survey of explainable recommendation on three perspectives: 1) We provide a chronological research timeline of explainable recommendation, including user study approaches in the early years and more recent model-based approaches. 2) We provide a two-dimensional taxonomy to classify existing explainable recommendation research: one dimension is the information source (or display style) of the explanations, and the other dimension is the algorithmic mechanism to generate explainable recommendations. 3) We summarize how explainable recommendation applies to different recommendation tasks, such as product recommendation, social recommendation, and POI recommendation.

We also devote a section to discuss the explanation perspectives in broader IR and AI/ML research. We end the survey by discussing potential future directions to promote the explainable recommendation research area and beyond.
1

Introduction

1.1 Explainable Recommendation

Explainable recommendation refers to personalized recommendation algorithms that address the problem of why—they not only provide users or system designers with recommendation results, but also explanations to clarify why such items are recommended. In this way, it helps to improve the transparency, persuasiveness, effectiveness, trustworthiness, and user satisfaction of the recommendation systems. It also facilitates system designers to diagnose, debug, and refine the recommendation algorithm.

To highlight the position of explainable recommendation in the recommender system research area, we classify personalized recommendation with a broad conceptual taxonomy. Specifically, personalized recommendation research can be classified into the 5W problems – when, where, who, what, and why, corresponding to time-aware recommendation (when), location-based recommendation (where), social recommendation (who), application-aware recommendation (what), and explainable recommendation (why), where explainable recommendation aims to answer why-type questions in recommender systems.
Explainable recommendation models can either be model-intrinsic or model-agnostic (Lipton, 2018; Molnar, 2019). The model-intrinsic approach develops interpretable models, whose decision mechanism is transparent, and thus, we can naturally provide explanations for the model decisions (Zhang et al., 2014a). The model-agnostic approach (Wang et al., 2018d), or sometimes called the post-hoc explanation approach (Peake and Wang, 2018), allows the decision mechanism to be a blackbox. Instead, it develops an explanation model to generate explanations after a decision has been made. The philosophy of these two approaches is deeply rooted in our understanding of human cognitive psychology – sometimes we make decisions by careful, rational reasoning and we can explain why we make certain decisions; other times we make decisions first and then find explanations for the decisions to support or justify ourselves (Lipton, 2018; Miller, 2019).

The scope of explainable recommendation not only includes developing transparent machine learning, information retrieval, or data mining models. It also includes developing effective methods to deliver the recommendations or explanations to users or system designers, because explainable recommendations naturally involve humans in the loop. Significant research efforts in user behavior analysis and human-computer interaction community aim to understand how users interact with explanations.

With this section, we will introduce not only the explainable recommendation problem, but also a big picture of the recommender system research area. It will help readers to understand what is unique about the explainable recommendation problem, what is the position of explainable recommendation in the research area, and why explainable recommendation is important to the area.

1.2 A Historical Overview

In this section, we will provide a historical overview of the explainable recommendation research. Though the term explainable recommendation was formally introduced in recent years (Zhang et al., 2014a), the basic concept, however, dates back to some of the earliest works in personalized
recommendation research. For example, Schafer et al. (1999) noted that recommendations could be explained by other items that the user is familiar with, such as this product you are looking at is similar to these other products you liked before, which leads to the fundamental idea of item-based collaborative filtering (CF); Herlocker et al. (2000) studied how to explain CF algorithms in MovieLens based on user surveys; and Sinha and Swearingen (2002) highlighted the role of transparency in recommender systems. Besides, even before explainable recommendation has attracted serious research attention, the industry has been using manual or semi-automatic explanations in practical systems, such as the people also viewed explanation in e-commerce systems (Tintarev and Masthoff, 2007a).

To help the readers understand the “pre-history” research of recommendation explanation and how explainable recommendation emerged as an essential research task in the recent years, we provide a historical overview of the research line in this section.

Early approaches to personalized recommender systems mostly focused on content-based or collaborative filtering (CF)-based recommendation (Ricci et al., 2011). Content-based recommender systems model user and item profiles with various available content information, such as the price, color, brand of the goods in e-commerce, or the genre, director, duration of the movies in review systems (Balabanović and Shoham, 1997; Pazzani and Billsus, 2007). Because the item contents are easily understandable to users, it was usually intuitive to explain to users why an item is recommended. For example, one straightforward way is to let users know the content features he/she might be interested in the recommended item. Ferwerda et al. (2012) provided a comprehensive study of possible protocols to provide explanations for content-based recommendations.

However, collecting content information in different application domains is time-consuming. Collaborative filtering (CF)-based approaches (Ekstrand et al., 2011), on the other hand, attempt to avoid this difficulty by leveraging the wisdom of crowds. One of the earliest CF algorithms is User-based CF for the GroupLens news recommendation system (Resnick et al., 1994). User-based CF represents each user as a vector of ratings, and predicts the user’s missing rating on a news
message based on the weighted average of other users’ ratings on the message. Symmetrically, Sarwar et al. (2001) introduced the Item-based CF method, and Linden et al. (2003) further described its application in Amazon product recommendation system. Item-based CF takes each item as a vector of ratings, and predicts the missing rating based on the weighted average of ratings from similar items.

Though the rating prediction mechanism would be relatively difficult to understand for average users, user- and item-based CF are somewhat explainable due to the philosophy of their algorithm design. For example, the items recommended by user-based CF can be explained as “users that are similar to you loved this item”, while item-based CF can be explained as “the item is similar to your previously loved items”. However, although the idea of CF has achieved significant improvement in recommendation accuracy, it is less intuitive to explain compared with content-based algorithms. Research pioneers in very early stages also noticed the importance of the problem (Herlocker and Konstan, 2000; Herlocker et al., 2000; Sinha and Swearingen, 2002).

The idea of CF achieved further success when integrated with Latent Factor Models (LFM) introduced by Koren (2008) in the late 2000s. Among the many LFMs, Matrix Factorization (MF) and its variants were especially successful in rating prediction tasks (Koren et al., 2009). Latent factor models have been leading the research and application of recommender systems for many years. However, though successful in recommendation performance, the “latent factors” in LFMs do not possess intuitive meanings, which makes it difficult to understand why an item got good predictions or why it got recommended out of other candidates. This lack of model explainability also makes it challenging to provide intuitive explanations to users, since it is hardly acceptable to tell users that we recommend an item only because it gets higher prediction scores by the model.

To make recommendation models better understandable, researchers have gradually turned to Explainable Recommendation Systems, where the recommendation algorithm not only outputs a recommendation list, but also explanations for the recommendations by working in an explainable way. For example, Zhang et al. (2014a) defined the explainable recommendation problem, and proposed an Explicit Factor
Model (EFM) by aligning the latent dimensions with explicit features for explainable recommendation. More approaches were also proposed to address the explainability problem, which we will introduce in detail in the survey. It is worthwhile noting that deep learning (DL) models for personalized recommendation have emerged in recent years. We acknowledge that whether DL models truly improve the recommendation performance is controversial (Dacrema et al., 2019), but this problem is out of the scope of this survey. In this survey, we will focus on the problem that the black-box nature of deep models brings difficulty in model explainability. We will review the research efforts on explainable recommendation over deep models.

In a broader sense, the explainability of AI systems was already a core discussion in the 1980s era of “old” or logical AI research, when knowledge-based systems predicted (or diagnosed) well but could not explain why. For example, the work of Clancy showed that being able to explain predictions requires far more knowledge than just making correct predictions (Clancey, 1982). The recent boom in big data and computational power have brought AI performance to a new level, but researchers in the broader AI community have again realized the importance of Explainable AI in recent years (Gunning, 2017), which aims to address a wide range of AI explainability problems in deep learning, computer vision, autonomous driving systems, and natural language processing tasks. As an essential branch of AI research, this also highlights the importance of the IR/RecSys community to address the explainability issues of various search and recommendation systems. Moreover, explainable recommendation has also become a very suitable problem setting to develop new Explainable Machine Learning theories and algorithms.

1.3 Classification of the Methods

In this survey, we provide a classification taxonomy of existing explainable recommendation methods, which can help readers to understand the state-of-the-art of explainable recommendation research.

Specifically, we classify existing explainable recommendation research with two orthogonal dimensions: 1) The information source or
Introduction
display style of the explanations (e.g., textual sentence explanation, or visual explanation), which represents the human-computer interaction (HCI) perspective of explainable recommendation research, and 2) the model to generate such explanations, which represents the machine learning (ML) perspective of explainable recommendation research. Potential explainable models include the nearest-neighbor, matrix factorization, topic modeling, graph models, deep learning, knowledge reasoning, association rule mining, and others.

With this taxonomy, each combination of the two dimensions refers to a particular sub-direction of explainable recommendation research. We should note that there could exist conceptual differences between “how explanations are presented (display style)” and “the type of information used for explanations (information source)”. In the context of explainable recommendation, however, these two principles are closely related to each other because the type of information usually determines how the explanations can be displayed. As a result, we merge these two principles into a single classification dimension. Note that among the possibly many classification taxonomies, this is just one that we think would be appropriate to organize the research on explainable recommendation, because it considers both HCI and ML perspectives of explainable recommendation research.

Table 1.1 shows how representative explainable recommendation research is classified into different categories. For example, the Explicit Factor Model (EFM) for explainable recommendation (Zhang et al., 2014a) developed a matrix factorization method for explainable recommendation, which provides an explanation sentence for the recommended item. As a result, it falls into the category of “matrix factorization with textual explanation”. The Interpretable Convolutional Neural Network approach (Seo et al., 2017), on the other hand, develops a deep convolutional neural network model and displays item features to users as explanations, which falls into the category of “deep learning with user/item feature explanation”. Another example is visually explainable recommendation (Chen et al., 2019b), which proposes a deep model to generate image regional-of-interest explanations, and it belongs to the “deep learning with visual explanation” category. We also classify other
Table 1.1: A classification of existing explainable recommendation methods. The classification is based on two dimensions, i.e., the type of model for explainable recommendation (e.g., matrix factorization, topic modeling, deep learning, etc.) and the information/style of the generated explanation (e.g., textual sentence explanation, etc.). Note that due to the table space this is an incomplete enumeration of the existing explainable recommendation methods, and more methods are introduced in detail in the following parts of the survey. Besides, some of the table cells are empty because to the best of our knowledge there has not been a work falling into the corresponding combination.

<table>
<thead>
<tr>
<th>Information/style of the explanations</th>
<th>Neighbor-based</th>
<th>Matrix factorization</th>
<th>Topic modeling</th>
<th>Graph-based</th>
<th>Deep learning</th>
<th>Knowledge-based</th>
<th>Rule mining</th>
<th>Post-hoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual explanation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chen <em>et al.</em>, 2019b</td>
<td></td>
</tr>
<tr>
<td>Social explanation</td>
<td>Sharma and Cosley, 2013</td>
<td>Ren <em>et al.</em>, 2017</td>
<td>Park <em>et al.</em>, 2018</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word cluster</td>
<td>Zhang, 2015</td>
<td>Wu and Ester, 2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Full text available at: http://dx.doi.org/10.1561/1500000066
research according to this taxonomy, so that readers can understand the relationship between existing explainable recommendation methods.

Due to the large body of related work, Table 1.1 is only an incomplete enumeration of explainable recommendation methods. For each “model – information” combination, we present one representative work in the corresponding table cell. However, in Sections 2 and 3 of the survey, we will introduce the details of many explainable recommendation methods.

1.4 Explainability and Effectiveness

Explainability and effectiveness could sometimes be conflicting goals in model design that we have to trade-off (Ricci et al., 2011), i.e., we can either choose a simple model for better explainability, or choose a complex model for better accuracy while sacrificing the explainability. While recent evidence also suggests that these two goals may not necessarily conflict with each other when designing recommendation models (Bilgic et al., 2004; Zhang et al., 2014a). For example, state-of-the-art techniques – such as the deep representation learning approaches – can help us to design recommendation models that are both effective and explainable. Developing explainable deep models is also an attractive direction in the broader AI community, leading to progress not only in explainable recommendation research, but also in fundamental explainable machine learning problems.

When introducing each explainable recommendation model in the following sections, we will also discuss the relationship between explainability and effectiveness in personalized recommendations.

1.5 Explainability and Interpretability

Explainability and interpretability are closely related concepts in the literature. In general, interpretability is one of the approaches to achieve explainability. More specifically, Explainable AI (XAI) aims to develop models that can explain their (or other model’s) decisions for system designers or normal users. To achieve the goal, the model can be either interpretable or non-interpretable. For example, interpretable models (such as interpretable machine learning) try to develop models whose
1.6 How to Read the Survey

Potential readers of the survey include both researchers and practitioners interested in explainable recommendation systems. Readers are encouraged to prepare with basic understandings of recommender systems, such as content-based recommendation (Pazzani and Billsus, 2007), collaborative filtering (Ekstrand et al., 2011), and evaluation of recommender systems (Shani and Gunawardana, 2011). It is also beneficial to read other related surveys such as explanations in recommender systems from a user study perspective (Tintarev and Masthoff, 2007a), interpretable machine learning (Lipton, 2018; Molnar, 2019), as well as explainable AI in general (Gunning, 2017; Samek et al., 2017).

The following part of the survey will be organized as follows. In Section 2 we will review explainable recommendation from a user-interaction perspective. Specifically, we will discuss different information sources that can facilitate explainable recommendation, and different display styles of recommendation explanation, which are closely related with the corresponding information source. Section 3 will focus on a machine learning perspective of explainable recommendation, which will introduce different types of models for explainable recommendation. Section 4 will introduce evaluation protocols for explainable recommendation, while Section 5 introduces how explainable recommendation
methods are used in different real-world recommender system applications. In Section 6 we will summarize the survey with several important open problems and future directions of explainable recommendation research.
References


Bauman, K., B. Liu, and A. Tuzhilin (2017). “Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews”. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 717–725.


References


References


Full text available at: http://dx.doi.org/10.1561/1500000066


References


