

Measuring "Why" in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation

1. Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China
2. Department of Computer Science, Rutgers University



arxiv.org

<https://arxiv.org> › cs

[2202.06466] Measuring "Why" in Recommender Systems

X Chen 저술 · 2022 · 7회 인용 — Title: Measuring "Why" in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation.

Comments: 9 pages, 2 tables, submitted to IJCAI 2022 Survey Track

Subjects: **Information Retrieval (cs.IR)**; Artificial Intelligence (cs.AI)

Cite as: [arXiv:2202.06466](https://arxiv.org/abs/2202.06466) [cs.IR]

(or [arXiv:2202.06466v1](https://arxiv.org/abs/2202.06466v1) [cs.IR] for this version)

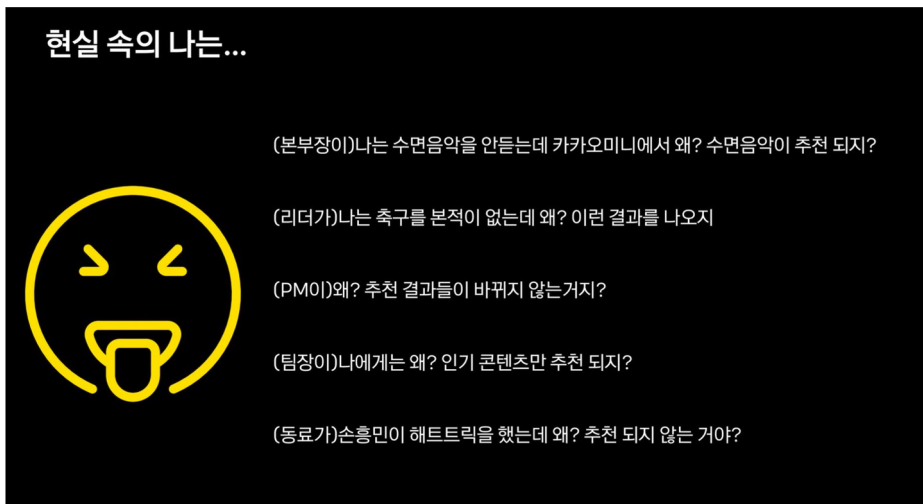
<https://doi.org/10.48550/arXiv.2202.06466> 

목차

1. Explainable Recommender System 이란
 - a. 카카오 사례
2. Abstract
3. Introduction
4. Evaluation Perspective
5. Evaluation Methods
6. Conclusion and Outlooks

Explainable Recommender System 이란

- 어떤 추천 시스템이 사용자의 선호도를 파악해서 상품을 추천할 때, **해당 상품을 추천하는 이유를 함께 제공하는 패러다임**
- 현실 속에서, 서비스 이해관계자들로부터 피드백을 받음
 - 추천팀은 '왜' 이런 추천 결과가 나왔는지 증명하고자 함



Explainable Recommender System 이란

- 서비스가 잘 되면, '왜' 를 궁금해하지 않음
 - 서비스 공급자 관점에서 안되는 이유를 알기 위해서
 - 이유를 못 찾으면 서비스를 접는다.

현실 속의 나는... 소비이력, 서비스이슈, 시스템을 먼지까지 털어
추천 이유를 설명합니다.



(본부장이)나는 수면음악을 안듣는데 카카오키니에서 왜? 수면음악이 추천 되지?

→ 오후 시간대 고양이 수면 음악 시청 이력이 있는데 본부장님 계정을 다른 분이 사용하시나요?

최근에 축구를 본적이 없는데 왜? 이런 결과를 나오지

→ 오늘 손흥민이 헤트트릭을 해서 모든 이들이 손흥민 기사만 봅니다.

왜? 추천 결과들이 바뀌지 않는거지?

→ 이번에 적용된 CTR 예측 모델의 추천 결과 업데이트 주기가 15분 입니다. 추천 결과가 바뀌는 것이 사용자 만족도가 올라 간다면 업데이트 주기를 10분으로 단축시키고, 실시간으로 추천 랭킹을 제공하는 방식도 검토 하겠습니다.

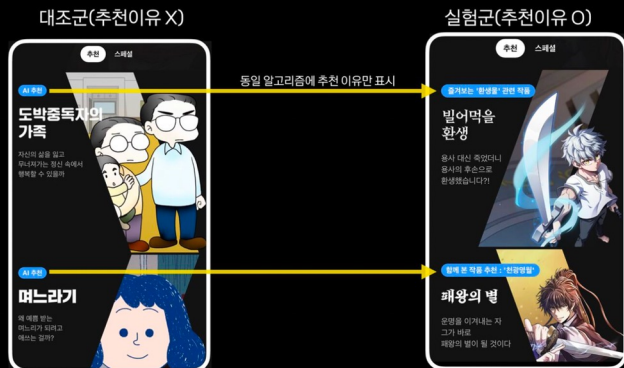
나에게는 왜? 인기 콘텐츠만 추천 되지?

→ 팀장님의 계정에 소비 이력이 없습니다. (테스트 계정으로 로그인됨)

Explainable Recommender System 이란

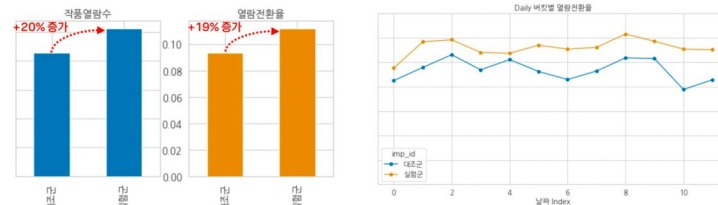
- 고객도 추천 이유를 궁금해할까?
 - NO. 고객이 추천을 만족하지 못하면, 소리없이 서비스를 떠남
 - 그러나, 추천 이유를 알려주면 잔존하는 고객도 생겨남. 즉, 사용자에게 추천 이유를 설명한 것만으로도, 지표 개선이 됨

카카오웹툰에 설명가능한 추천을 적용했습니다.



추천이유 표시 여부 AB테스트 결과

- 추천이유가 명시된 실험군이 대조군 대비 **작품 열람수는 +20%**,
열람전환율은 +19%가 높게 나왔습니다.



Explainable Recommender System 이란

- Explainable recommender system Goal
 - 사용자와 공급자에게 추천 이유를 제공하여, 목표 달성을 돕는다

	서비스 사용자	서비스 공급자
Recommender System Goal	Enjoy, Satisfaction, Entertained	Keep engaged, Revenue, Monetization
Explainable Recommender System Goal	투명성 (Transparency) 설득력 (Persuasiveness) 효과성 (Effectiveness) 신뢰성 (Trustworthiness) 만족도 (Satisfaction)	추천 알고리즘의 진단과 개선 디버깅

Abstract

- Explainable Recommend (ER) 은 추천의 설득력, 유저 만족, 시스템 투명성 등을 개선하는 데 큰 장점을 보여줌
 - 그러나, ER 은 'explanations' 의 근본적인 문제 중 하나는 '설명을 평가하는 방법' 임
- 과거에는 다양한 평가 방법이 제안되었으나, 각기 다른 논문에 흩어져있고, 평가 방법들 사이의 디테일한 비교와 체계성이 떨어짐
- 본 논문에서는, 이전 연구들을 기반으로, **평가 관점 (evaluation perspective)** 및 **평가 방법 (evaluation method)** 에 따른 분류 체계 제공
- 본 논문은 top-tier 컨퍼런스 (IJCAI, AAAI, TheWebConf, Recsys, UMAP, IUI 등) 의 100개 이상의 논문을 리뷰했고, 비교 시트를 작성함
 - <https://shimo.im/sheets/VKrpYTcwVH6KXgdy/MO%20DOC>

Abstract

- 본 논문은 top-tier 컨퍼런스 (IJCAI, AAAI, TheWebConf, Recsys, UMAP, IUI 등) 의 100개 이상의 논문을 리뷰했고, 비교 시트를 작성함
 - <https://shimo.im/sheets/VKrpYTcwVH6KXgdy/MO%20DOC>

Paper Title					
A	B	C	D	E	F
Paper Title	Evaluation Perspective	Evaluation Method	Conference/Journal	Year	
1 Explainable Session-based Recommendation with Meta-path Guided Instances and Transparency	Transparency (Case studies)	Case studies - Attention weights and Meta-path	SIGIR	2022	@inproceedings(zheng2022explainable, title=[Explainable Session-based Recommendation with Meta-path Guided Instances and Self-Attention Mechanism
2 Post Processing Recommender Systems with Knowledge Graphs for Recency, Popularity, and Diversity of Explanation	Effectiveness (Quantitative metrics)	Quantitative metrics - recency(LRI), popularity(SEP), di	SIGIR	2022	@inproceedings(balocco2022post, title=[Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanation
3 PEVAE: A Hierarchical VAE for Personalized Explainable Recommendation.	Effectiveness (Quantitative metrics, Case studies)	Quantitative metrics - BLEU, ROUGE, METEOR, Distric Case studies - Fully generated natural language explan	SIGIR	2022	@inproceedings(cai2022pevae, title=[PEVAE: A Hierarchical VAE for Personalized Explainable Recommendation], author=[Cai, Zefeng and Cai, Zerui], book
4 Explanation Guided Contrastive Learning for Sequential Recommendation	Effectiveness (Quantitative metrics)	Quantitative metrics - NDCG (Comparison among mod	CIKM	2022	@inproceedings(wang2022explanation, title=[Explanation guided contrastive learning for sequential recommendation], author=[Wang, Lei and Lim, Ee-Peng
5 PARSRec: Explainable Personalized Attention-fused Recurrent Sequential Recommen	Transparency (Case studies)	Case studies - Attention weights	KDD	2022	@article(gholami2022parsrec, title=[PARSRec: Explainable personalized attention-fused recurrent sequential recommendation using session partial actions],
6 Reinforcement Learning over Sentiment-Augmented Knowledge Graphs towards Acco	Effectiveness (Crowdsourcing)	Case studies - Top sentiment-related words & explana	WSDM	2022	@inproceedings(park2022reinforcement, title=[Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable reco
7 Path Language Modeling over Knowledge Graph for Explainable Recommendation	Effectiveness (Case studies)	Case studies - Recommendation path	WWW	2022	@inproceedings(gang2022path, title=[Path language modeling over knowledge graphs for explainable recommendation], author=[Gang, Shijie and Fu, Zuchui
8 Accurate and Explainable Recommendation via Review Rationalization	Effectiveness (Quantitative metrics)	Quantitative metrics - Explainability Recall Case studies - Rationales extracted from reviews	WWW	2022	@inproceedings(pan2022accurate, title=[Accurate and Explainable Recommendation via Review Rationalization], author=[Pan, Sicheng and Li, Dongsheng an
9 Comparative Explanations of Recommendations	Effectiveness (Quantitative metrics, Crowdsourcing) Persuasiveness (Crowdsourcing)	Quantitative metrics - IDF-BLEU, BLEU, Feature precis Crowdsourcing - Crowdsourcing with public datasets	WWW	2022	@inproceedings(yang2022comparative, title=[Comparative Explanations of Recommendations], author=[Yang, Aobo and Wang, Nan and Cai, Renqin and Der
10 ExpScore: Learning Metrics for Recommendation Explanation	Effectiveness (Quantitative metrics)	Quantitative metrics - ExpScore (proposed in paper), B	WWW	2022	@inproceedings(wen2022expscore, title=[ExpScore: Learning Metrics for Recommendation Explanation], author=[Wen, Bingbing and Feng, Yunhe and Zhang
11 Graph-based Extractive Explainer for Recommendations	Effectiveness (Quantitative metrics, Case studies)	Quantitative metrics - BLEU, ROUGE, Attribute-level pr Case studies - Fully generated natural language explan	WWW	2022	@inproceedings(wang2022graph, title=[Graph-based Extractive Explainer for Recommendations], author=[Wang, Peng and Cai, Renqin and Wang, Hongning
12 ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recomm	Transparency (Case studies)	Case studies - Representative user/item prototypes, Vi	RecSys	2022	@inproceedings(melchiorre2022protomf, title=[ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations], author=[Me
13 Explaining Recommendations in E-Learning: Effects on Adolescents' Trust	Effectiveness (Crowdsourcing)	Crowdsourcing - randomized controlled experiment wit	IUI	2022	@inproceedings(ooge2022explaining, title=[Explaining Recommendations in E-Learning: Effects on Adolescents' Trust], author=[Ooge, Jeroen and Kato, Shc
14 Similarity-Based Explanations meet Matrix Factorization via Structure-Preserving Em	Effectiveness (Quantitative metrics) Transparency (Case studies)	Quantitative metrics - the Mantel Test, Neighborhood C Case studies - Top nearest neighbors to the recomm	IUI	2022	@inproceedings(marinho2022similarity, title=[Similarity-Based Explanations meet Matrix Factorization via Structure-Preserving Embeddings], author=[Marin
15 Explaining Call Recommendations in Nursing Homes: a User-Centered Design Appro	Effectiveness (Quantitative metrics)	Crowdsourcing - Questionnaires	IUI	2022	@inproceedings(gutierrez2022explaining, title=[Explaining call recommendations in nursing homes: a user-centered design approach for interacting with kn
16 Generating Recommendations with Post-Hoc Explanations for Citizen Science	Effectiveness (Online experiments, Crowdsourcing)	Online experiments Crowdsourcing	UMAP	2022	@inproceedings(ber2022generating, title=[Generating Recommendations with Post-Hoc Explanations for Citizen Science], author=[Ben Zaken, Daniel and S
17 Explaining User Models with Different Levels of Detail for Transparent Recommendation	Effectiveness (Case studies) Transparency (Case studies)	Case studies - Explanations with different levels of det Crowdsourcing - Measurements of six personal charac	UMAP	2022	@inproceedings(guesm2022explaining, title=[Explaining user models with different levels of detail for transparent recommendation: A user study], author=[G
18 Entity-Enhanced Graph Convolutional Network for Accurate and Explainable Recomm	Effectiveness (Quantitative metrics) Transparency (Case studies)	Case studies - Attention weights with a real explanation Quantitative metrics - the Overlap between relevant e	UMAP	2022	@inproceedings(wang2022entity, title=[Entity-Enhanced Graph Convolutional Network for Accurate and Explainable Recommendation], author=[Wang, Qing
19 Is More Always Better? The Effects of Personal Characteristics and Level of Detail o	Effectiveness (Crowdsourcing) Transparency (Case studies)	Case studies - Explanations with different levels of det Crowdsourcing - Measurements of six personal charac	UMAP	2022	@inproceedings(chatt2022more, title=[Is more always better? the effects of personal characteristics and level of detail on the perception of explanations in
20 Attribute-aware Explainable Complementary Clothing Recommendation	Transparency (Case studies) Effectiveness (Crowdsourcing)	Case studies - Attention weights (Image) Crowdsourcing - Crowdsourcing with public datasets -	WWWJ	2021	@article(li2021attribute, title=[Attribute-aware explainable complementary clothing recommendation], author=[Li, Yang and Chen, Tong and Huang, Zi], j
21 EX3: Explainable Attribute-aware Item-set Recommendations	Effectiveness (Crowdsourcing) Persuasiveness (Online experiments)	Crowdsourcing - Crowdsourcing with public datasets - Online experiments	Recsys	2021	@inproceedings(xian2021ex3, title=[Ex3: Explainable attribute-aware item-set recommendations], author=[Xian, Yikun and Zhao, Tong and Li, Jin and Ch

Introduction

- **Explainable Recommend** 은 '왜 이 아이템이 추천되었는가' 를 해결함
 - 추천 이유를 제공함으로써, 사람들은 더 많은 정보를 얻어 더 명확한 결정을 내릴 수 있음. 이를 통해 추천의 설득력과 알고리즘의 투명성이 개선될 수 있음
 - 예를 들어, 추천 시스템이 음식을 추천하는 경우, 추천한 음식에 대한 이유와 근거를 설명해 준다면,
 - 사용자는 왜 그 추천이 나왔는지를 이해하고, 그 추천이 자신에게 맞는지 여부를 판단할 수 있음. 이를 통해 추천 결과에 대한 사용자의 신뢰도가 높아지고, 사용자는 추천 시스템이 자신의 취향과 요구에 맞게 동작하는지 알 수 있음
 - 또한, 추천 시스템이 내부적으로 사용하는 알고리즘이나 데이터가 공개되어 설명될 경우, 시스템의 투명성이 개선됨.

Introduction

- In the field of explainable recommendation, **evaluation is a fundamental problem**
 - 다른 evaluation task 와 비교했을 때, Explainable Recommend 은 GT 가 사람의 감정을 기반으로 하고, 측정하기 어려워, 평가하기 어려운 task 임
 - 이를 위해, 과거에 많은 평가 지표가 소개되었지만, 모두 다른 논문에서 독립적으로 소개되었음
- 위 문제를 해결하기 위해, 본 논문에서는 과거 논문을 기반으로. evaluation 전략에 대해 clear summarization 제공
 - 예를 들어, recommendation explanations 는 다른 타겟을 가질 수 있음
 - **users, providers, and model designers**
 - 또한, 이전 연구들로부터 주요한 4가지 evaluation perspective 정리
 - **explanation effectiveness, transparency, persuasiveness and scrutability**
 - 4가지 카테고리의 evaluation method 소개
 - **case studies, quantitative metrics, crowdsourcing, and online experiments**
- 마지막으로, real-world problems 에서 가이드라인을 어떻게 선택해야 하는지를 설명함

Introduction

- Main contributions
 - 토픽 처음으로 explainable recommendation evaluation 을 체계적으로 정리함
 - 100개가 넘는 top-tier 컨퍼런스 논문들의 평가 관점, 평가 지표를 요약 정리
 - 존재하는 평가 방법의 장단점을 정리하고, 어떤 시나리오에서 어떤 평가 지표를 사용해야 할지 가이드라인 제공

Evaluation Perspectives

- 평가 관점
 - Recommendation explanations 는 다양한 목적으로 사용될 수 있음
 - Three Stakeholders
 - Users, providers, and model designers

Table 1: Summarization of the evaluation perspectives.

Evaluation perspective	Evaluation problem	Representative papers	Serving target
Effectiveness	Whether the explanations are useful for the users to make more accurate/faster decisions?	[Hada, 2021; Guesmi, 2021; Wang, 2020; Gao <i>et al.</i> , 2019; Chen <i>et al.</i> , 2013]	Users
Transparency	Whether the explanations can reveal the internal working principles of the recommender models?	[Chen, 2021a; Sonboli, 2021; Li, 2021c; Li, 2020c; Fu, 2020]	Model designers
Persuasiveness	Whether the explanations can increase the click/purchase rate of the users on the items?	[Musto <i>et al.</i> , 2019; Tsai and Brusilovsky, 2019; Balog and Radlinski, 2020; Tsukuda, 2020; Chen and Wang, 2014]	Providers
Scrutability	Whether the explanations can exactly correspond to the recommendation results?	[He <i>et al.</i> , 2015; Tsai and Brusilovsky, 2019; Cheng <i>et al.</i> , 2019; Balog <i>et al.</i> , 2019; Liu, 2020]	Model designers

Evaluation Perspectives

- **Effectiveness** (users)
 - 사용자(User)의 편의를 개선
 - **"Explanations 을 통해 사용자에게 더 나은 서비스를 제공하는가?"** 를 평가하는 것을 목표로 함
 - 추천 시스템의 목적은 사용자가 원하는 제품이나 서비스를 제공하는 것이므로, 설명을 제공하여 사용자가 추천 결과를 이해하고, 적절한 결정을 내릴 수 있도록 돕는 것이 중요함. 이를 통해 사용자의 만족도와 신뢰도를 높일 수 있으며, 추천 시스템이 보다 효과적으로 작동할 수 있음.
- **Transparency** (model designers)
 - 투명도
 - **"Explanations 이 추천 모델 내부 작동 원리를 잘 설명하는가?"** 를 평가하는 것을 목표로 함
 - 이를 통해, model designer 들은 추천 모델을 어떻게 더 잘 구성하고, 디버깅 할 지 알 수 있음

Evaluation Perspectives

- **Persuasiveness (Providers)**

- 설득력
- **"Explanations 이 사용자의 아이템 상호작용 확률을 증가시킬 수 있는지"**를 평가하는 것을 목표로 함
- 즉, 추천 시스템에서 사용자에게 제공되는 설명이 사용자가 아이템에 대한 상호작용을 더 많이 하도록 유도할 수 있는지 여부를 확인하고자 함. 예를 들어, 음악 추천 시스템에서 사용자가 특정 곡을 추천받았을 때, 그 곡에 대한 설명을 제공하여 사용자가 그 곡을 듣고 좋아하게 되어, 해당 아티스트의 다른 곡들에 대한 상호작용 확률을 높일 수 있는지를 평가하는 것
- 따라서, 설명의 "설득력"은 사용자의 상호작용 확률을 증가시킬 수 있는 능력과 밀접한 관련이 있음

Evaluation Perspectives

- **Scrutability** (model designers)

- 정밀 조사 가능성
- "설명이 추천 결과와 정확하게 일치하는지 여부"를 평가하는 것을 목표로 함
 - 즉, 추천 시스템에서 사용자에게 제공되는 설명이 추천 결과와 일치하는지를 확인하고자 함. 예를 들어, 사용자가 특정 상품을 추천 받았을 때, 그 추천 결과와 관련된 특징이나 이유를 설명으로 제공하였을 때, 그 설명이 추천 결과와 일치하는지 여부를 평가하는 것입니다.
- Transparency 가 모델 내부에 관심을 두었다면, scrutability 는 설명과 출력 사이의 관계에 더 관심을 둬
 - 모델이 어떻게 작동하는지에 대한 지식이 없어도 설명이 정확하게 일치하는지를 평가할 수 있고 보다 일반적이고 범용적인 평가 척도로 사용될 수 있음
 - Model agnostic - 모델에 대한 특정한 가정이나 제약이 없이 일반적으로 적용 가능함

Evaluation Methods

- 위의 4가지 관점에서 recommendation explanations 을 평가하기 위해, 다양한 평가 methods 디자인되었음
- 이전 이 연구들은 사려하며 크게 4가지 카테고리 그룹화

Table 2: Summarization of the evaluation methods. In the second and third column, we present the most significant strengths and shortcomings of different evaluation methods. In the last column, we present the perspectives that a method is usually leveraged to evaluate.

Evaluation methods	Strengths	Shortcomings	Representative papers	Evaluation perspectives
Case studies	Better intuitiveness	Bias; Cannot make comparisons	[Xian, 2020; Xian, 2020; Fu, 2020; Liu, 2020; Barkan, 2020; Li, 2020c]	Effectiveness; Transparency
Quantitative metrics	Quantitative evaluation; Easy to benchmark; High efficiency	Deviating from the explanation goals; Less effective approximation	[Li, 2021a; Hada, 2021; Li, 2020a; Tan, 2021; Tai, 2021]	Effectiveness; Scrutability
Crowdsourcing	Based on real human feelings	High cost	[Li, 2021b; Xian, 2021; Li, 2020b; Chen, 2021b; Wang <i>et al.</i> , 2018]	Effectiveness; Scrutability; Transparency; Persuasiveness
Online experiments	High reliable	High cost	[Zhang <i>et al.</i> , 2014; Xian, 2021]	Effectiveness; Persuasiveness

Evaluation Methods

- **Evaluation with Case Studies**

- 구체적 사례를 대상으로 각 과정과 현상들을 기술한 후, 토의 과정을 통해 사례의 내용을 분석하고 해결책을 찾는 과정
- People usually present some examples to illustrate how the recommender model works and whether the generated explanations are aligned with human intuitions.
- The recommender models are firstly learned based on the training set, and then the examples are generated based on the intermediate or final outputs by feeding the testing samples into the optimized model.
 - 학습된 모델에 테스트 데이터를 입력한 후, 이 테스트의 데이터의 중간 또는 마지막 output 을 기반으로 example 이 생성된다.
 - 이렇게 생성된 example 은 positive, negative 할 수 있는데
 - positive 한 경우 : 모델의 effectiveness 함을 증명 가능
 - negative 한 경우 : 언제, 어떻게 모델이 실패하는지 알 수 있음

Evaluation Methods


- **Evaluation with Case Studies**

Pros	Cons
<ul style="list-style-type: none">• intuitivesness (직관력) : reader 가 explanation 이 얼마나 잘 형성되었는지 확인하기 쉬움	<ul style="list-style-type: none">• biased 될 수 있다. 하나의 샘플은 모든 샘플을 나타내지 않으니까.• 정량적인 평가가 없기 때문에, 다른 모델 간 성능 비교가 어렵고, 정확한 모델을 선정하기가 어려움

Evaluation Methods

- **Evaluation with Quantitative Metrics**

- 본 논문에서는 explainable recommendation problem 은 natural language generation (NLG) task 로 간주함. 즉, the goal is to accurately predict the user reviews.
- 이러한 formulation (공식화) 를 기반으로, 다음 metrics 는 explanation qualities 를 평가하는 데 사용됨

Metric	Description
BLEU scores ROUGE scores	<p>모델이 예측한 결과와 정답 사이의 일치하는 word level 에 기반하여 Natural language generation (기계 번역) 성능 평가</p> <ul style="list-style-type: none">• BLEU : n-gram Precision에 기반한 지표• ROUGE : n-gram Precision 에 기반한 지표 <p>참고 설명</p> <ul style="list-style-type: none">•  [자연어처리][Metric] ROUGE score : Recall-Oriented Und erstudy for Gisting Evaluation

Evaluation Methods

• Evaluation with Quantitative Metrics

Metric	Description
Unique Sentence Ratio (USR)	생성된 리뷰의 diversity (다양성) 평가
Feature Coverage Ratio (FCR)	제품 리뷰, 다른 유형의 텍스트 데이터의 품질을 평가하는데 자주 사용되는 평가 지표
Feature Diversity (FD)	<ul style="list-style-type: none">USR : sentence level 에서의 diversity 평가<ul style="list-style-type: none">리뷰에서 표현된 의견의 다양성을 측정In specific, suppose the set of unique reviews is S, and the total number of reviews is N, then this metric is computed as $USR = S /N$FCR : feature level 에서의 diversity 평가<ul style="list-style-type: none">리뷰에서 언급된 제품의 특징 수와 전체 특징 수 간의 비율 측정제품의 다양한 특징이 리뷰에서 충분히 다뤄졌는지 평가suppose the number of distinct features in all the generated reviews is M, and the total number of features is $N_{\{f\}}$, then $FCR = M/N_{\{f\}}$FD : feature level 에서의 diversity 평가<ul style="list-style-type: none">리뷰에서 다뤄진 특징의 종류 수특정 특징이 반복적으로 언급되지 않고, 다양한 특징들이 다뤄졌는지 평가 <p>feature set in the generated review for user-item pair (u, i) is $F_{u,i}$, then $FD = \frac{1}{N(N-1)} \sum_{(u,i) \neq (u',i')} F_{u,i} \cap F_{u',i'}$, where N is the total number of different user-item pairs.</p>

Metric	Description
Feature-level Precision (FP)	많은 연구 논문에서, explanation 은 실제 유저 리뷰와 비교하며 평가함
Recall (FR)	user-item pair (u, i) 에 대해 predicted feature sets $S_{\{u,i\}}$, 실제 predicted feature set $T_{\{u,i\}}$ 를 가정하면, 다음과 같이 계산할 수 있음
F1 (FF)	$FP_{u,i} = \frac{ S_{u,i} \cap T_{u,i} }{ S_{u,i} }$ $FR_{u,i} = \frac{ S_{u,i} \cap T_{u,i} }{ T_{u,i} }$ $FF_{u,i} = \frac{2 \cdot FP \cdot FR}{FP + FR}$ <p>최종적으로, 모든 user-item pair 에 대해서 평균을 취함 (N=the number of user-item pairs)</p>

Evaluation Methods

- Evaluation with Quantitative Metrics

Metric	Description
Feature Matching Ratio (FMR)	<p>사용자가 원하는 상품 특징과 추천된 상품 특징 간의 일치 정도 측정</p> <p>the predicted results. It is formally computed as $\hat{FMR} = \frac{1}{N} \sum_{u,i} \mathbf{1}(f_{u,i} \in S_{u,i})$, where $f_{u,i}$ is the input feature, $S_{u,i}$ is the predicted review, N is the number of user-item pairs, and $\mathbf{1}$ is the indicator function.</p> <ul style="list-style-type: none">FMR은 추천 시스템이 제공한 제품의 특징 중 사용자가 원하는 특징과 일치하는 특징의 수를 측정한 후, 해당 수를 사용자가 원하는 특징의 총 수로 나눈 것. 따라서 FMR이 높을수록 추천 시스템이 사용자가 원하는 제품 특징을 잘 파악하고 추천을 수행하고 있다는 것을 의미예를 들어, 사용자가 휴대폰 충전기를 구매하려고 하며, 사용자가 원하는 특징으로는 대용량, 빠른 충전 속도, 경량 등이 있다고 가정할 때, 추천 시스템이 제공한 제품의 특징 중 해당 특징과 일치하는 것이 몇 개인지 세어 FMR을 계산할 수 있음

Metric	Description
Probability of Necessity (PN)	<p>If the explanation features of an item had been ignored, whether this item will be removed from the recommendation list?</p> <ul style="list-style-type: none">계산 방법 <p>list?" Formally, Suppose A_{ij} is the explanation feature set of item j when being recommended to user i, let $PN_{ij} = 1$, if item j no longer exists in the recommendation list when one ignores the features in A_{ij}, otherwise $PN_{ij} = 0$. The final PN score is computed as $\sum_{i,j} \frac{PN_{ij}}{\mathbf{1}(A_{ij} >0)}$.</p>
Probability of Sufficiency (PS)	<p>If only the explanation features of an item are remained, whether this item will be still in the recommendation list?</p> <ul style="list-style-type: none">모델의 복잡한 구조와 관계없이 추천 아이템의 해석 가능한 특성만을 고려하여 추천 리스트를 재정렬한 뒤, 해당 아이템이 여전히 추천 리스트 상위에 위치하는지 여부를 평가. 이를 통해 모델이 추천 아이템의 해석 가능한 특성을 잘 파악하고 추천을 수행하는지를 측정할 수 있음.이 값이 높을수록 모델이 추천 아이템의 해석 가능한 특성을 잘 파악하고 추천을 수행하는 것으로 판단할 수 있음

Evaluation Methods

- Evaluation with Quantitative Metrics

Metric	Description
Performance Shift (PS)	Original recommendation performance 가 r , key history 를 제거 했을 대 r' 으로 변경되었다면, $RS = (r-r')/r$
Mean Explainability Precision (MEP) Mean Explainability Recall (MER)	<p>plainability Recall (MER). For a user u, suppose the set of items which can be explained is N_u, the recommended item set is M_u, then $MEP = \frac{ N_u \cap M_u }{ M_u }$ and $MER = \frac{ N_u \cap M_u }{ N_u }$.</p> <ul style="list-style-type: none">MEP는 추천 모델이 사용자에게 제시한 추천 결과 중, 설명 가능성이 높은 아이템의 비율을 나타내는 지표. 즉, 사용자가 선택한 아이템들 중에서 얼마나 많은 아이템들이 설명 가능성이 높은 아이템들인지를 측정MER은 설명 가능성이 높은 아이템들 중, 사용자가 선택한 아이템의 비율을 나타내는 지표. 즉, 모델이 설명 가능성이 높은 아이템들을 추천한 경우, 사용자가 이 중에서 얼마나 많은 아이템들을 선택했는지를 측정예를 들어, 추천 모델이 10개의 아이템을 추천했고, 그 중 설명 가능성이 높은 4개의 아이템이 포함되어 있음. 이 중 사용자가 2개의 아이템을 선택했다면, MEP는 0.2가 되고, MER은 0.5. 이를 통해 모델이 설명 가능성이 높은 아이템을 잘 추천했는지, 사용자가 이를 얼마나 만족스럽게 평가했는지를 측정할 수 있음

Evaluation Methods

- **Evaluation with Crowdsourcing**

- Crowdsourcing : Crowd (대중) + Outsourcing (외부 자원 활용) 의 합성어로, 전문가 대신 비전문가인 고객과 대중에게 문제의 해결책을 아웃소싱 하는 것
- 기업활동의 전 과정에 소비자 또는 대중이 참여할 수 있도록 일부를 개방하고 참여자의 기여로 기업활동의 성과가 향상되면, 그 수익을 참여자 즉 소비자와 공유하는 방법

Evaluation Methods

- **Evaluation with Crowdsourcing**
 - Human feelings 를 평가에 포함하자
 - 크게 3가지 전략으로 정리

no	구분
1	Crowdsourcing with public dataset
2	Crowdsourcing by injecting annotator data into public dataset
3	Crowdsourcing with fully constructed dataset

Evaluation Methods

- **Evaluation with Crowdsourcing - Crowdsourcing with public dataset**
 - Public dataset 으로 recommender model 학습 후,
많은 어노테이터들이 생성된 explanations 을 평가. 평가 시, 두 가지 포인트가 있음
 - **[Point 1] Annotation quality control**
 - Annotation quality 를 향상시키기 위해, voting 메커니즘이나, statistical quantities 를 계산함
 - **[Point 2] Annotation question designs**

Evaluation Methods

- **Evaluation with Crowdsourcing - Crowdsourcing with public dataset**
 - **[Point 2] Annotation question designs**
 - Annotation 질문이 다양한 논문에 다양하게 존재함을 확인하였고, 이를 3개의 카테고리로 분류함
 - Model-agnostic
 - The problems do not dependent on any specific model
 - 어노테이터들이 가능한 모든 모델에서의 explanation 답변을 확인하고, 그 중에서 GT 를 선정함.
 - Single-model
 - The problems focus on only one explainable model
 - 하나의 모델이 생성한 답변에 대해서 up-vote, down-vote 를 수행함
 - Pair-model
 - The problems focus on the comparison between a pair of explainable models
 - 어노테이터는 제안 모델과 baseline 모델 중의 어떤 모델이 더 설명을 잘 하는지 레이블링
 - The annotation questions are: between A and B, whose explanations do you think can better help you understand the recommendations? and whose explanations can better help you make a more informed decision?

Evaluation Methods

- **Evaluation with Crowdsourcing - Crowdsourcing with public dataset**
 - **[Point 2] Annotation question designs**
 - 위 3가지 방법 중에서 첫 번째 카테고리가 가장 생성하기 어렵지만, 생성한 데이터는 Model-agnostic 하기 때문에 다른 모델에 대해서도 재사용할 수 있음
 - 두 번째, 세 번째 카테고리는 생성하기는 쉽지만, 재사용될 수 없음

Evaluation Methods

- **Evaluation with Crowdsourcing - Crowdsourcing by injecting annotator data into public dataset**
 - 위 방법은 어노테이터가 실제 유저와 또 다른 선호도를 가지고 있을 수 있다는 문제가 있음
 - 이를 해결하기 위해 새로운 방법 고안 : Combining the annotator generated data with the public dataset
 - 예를 들어서, 어노테이터들이 yelp.com 에 실제 15 개의 리뷰를 작성하고, 해당 데이터로 모델 학습
 - 평가 단계에서, 어노테이터들은 모델 explanations 결과에 대해 1~5 점 사이로 평가
 - 학습된 데이터가 어노테이터가 실제로 만든 real user feedback 이기 때문에 모델 평가도 부합하다고 볼 수 있음

Evaluation Methods

- **Evaluation with Crowdsourcing - Crowdsourcing with fully constructed dataset**
 - 전체 데이터셋을 어노테이터가 구축함 (completely based on the annotator generated data.)
 - 일반적으로 4가지 과정을 거쳐 평가함
 - 다양한 백그라운드의 어노테이터 고용
 - 어노테이터 선호도 수집
 - 어노테이터들에게 추천과 추천에 대한 explanation 제공
 - 어노테이터 별로 다른 평가 결과 피드백
 - 일반적으로 위의 2개 카테고리 보다는 훨씬 많은 어노테이터가 필요함

Evaluation Methods

- **Evaluation with Crowdsourcing - Evaluation with Online Experiments**
 - Evaluate recommendation explanations with online experiments
 - Online user 가 3개의 그룹으로 나뉘어서, CTR 측정하여 explanations 평가
 - Proposed model / Baseline / No explanations

Evaluation Methods

- **Guidelines for selecting the evaluation methods**

- 다양한 특징의 Evaluation method 에 기반하여, real-world 문제에 적용해보자
 - Recommender model 이 high-stake task (결과가 중요한 영향을 미치는 작업) 을 수행할 때, reliable evaluation 평가 방법을 사용해야 함
 - General recommendation tasks 를 수행할 때, 다양한 평가 방법의 trade-off 를 생각하여 선택

Conclusion and Outlooks

- 본 논문에서, explanation 평가 방법에 초점을 맞춰 explainable recommendation 논문을 정리함
- In specific, introduce the main evaluation perspectives and methods in the previous work. For each evaluation method, the paper detail its characters, representative papers, and also highlight its strength and shortcomings

감사합니다.