Counterfactual Analysis in Knowledge Graph Based Recommender System

by

Celine Hovanessian Far

GRADUATION REPORT

Submitted to
American University of Armenia

ABSTRACT

Insert your abstract here. Original Hanze format states one can use a maximum of 200 words. Though, it is advised is to discuss with your supervisor(s) if you need more.

Keywords: Keyword, Keyword, Keyword.

Contents

Chapter 1

Introduction

1.1 Background and Motivation

The rapid growth of the Internet as a center for digital and commercial interactions has highlighted the essential role of recommender systems in today's technology landscape. These systems, which are increasingly ~~vital~~ to companies from media giants like Netflix to e-commerce leaders such as Amazon, utilize user feedback to create personalized experiences. This feedback varies from explicit ratings to implicit behavioral data, such as browsing and purchasing histories, and serves as a critical foundation for these systems. By analyzing past interactions between users and products, recommender systems can predict future preferences and use advanced algorithms to detect patterns in user behavior. These capabilities not only improve user satisfaction by aligning recommendations with individual tastes but also enhance business strategies by promoting customer engagement and retention. As user data grows more complex, these systems evolve by integrating sophisticated models that consider content attributes, user specifications, and contextual information, 'black-box' nature, where the internal reasoning processes remain hidden from both users and developers. This lack of transparency poses significant challenges, especially from a business perspective, where understanding and trust in these systems are crucial. For businesses, the drive to make recommender systems explainable is motivated by several key factors. Transparency builds trust; when users understand how recommendations are generated, they are more likely to trust and accept these suggestions, thereby increasing user satisfaction and loyalty. This is particularly important in sectors like e-commerce or content streaming, where the accuracy of recommendations

directly affects user engagement and retention. Furthermore, explainability aids in complying with regulatory standards, such as the European Union's General Data Protection Regulation (GDPR), which requires that users must understand the decisions made by automated systems that affect them. This regulation is critical for businesses to ensure that their AI systems can provide clear explanations for their outputs. Additionally, from a business innovation standpoint, explainability enables the refinement and optimization of recommender systems. By understanding the decision-making processes, developers can identify and address biases or errors, leading to more accurate and fair recommendations. This not only improves the system's performance but also increases its fairness and accountability, aligning with ethical AI practices. The development of explainable AI (XAI) seeks to clarify AI operations, allowing both users and developers to understand, trust, and effectively manage AI outputs. Recent trends toward this field ~~highlight~~ the growing recognition of the need for AI systems, particularly recommender systems, to be both effective and understandable. As such, the business case for explainable recommender systems is strong, promising not only to enhance user experience and comply with regulations but also to foster technological advancement and ethical AI practices (Vultureanu-Albisi & Badica, 2021) Explainable recommender systems can be understood from different angles. One way for their classification is their approach to explainability.

Types of Explainable Recommender Systems

- **Intrinsic Explainability:** This approach involves developing interpretable models where the decision-making process is transparent. As a result, it's easier to provide natural explanations for the decisions made by the model.
- **Model-Agnostic Approach:** Also known as the post-hoc explanation approach, this allows the decision mechanism to remain a black box. Instead of making the decision process itself clear, it focuses on developing a separate explanation model that generates explanations after a decision has been made.

The underlying philosophy of these approaches aligns closely with theories in human cognitive psychology. Sometimes, we make decisions through careful, rational reasoning and can clearly explain why we made those decisions. At other times, we make decisions first and then rationalize them afterward to support or justify our actions. There are multiple approaches to achieving intrinsic explainability in recommender systems. These include factorization-based, topic modeling, graph-based, specialized deep learning methods, knowledge-based, and rule mining approaches. Knowledge graphs, in particular, play

a critical role in advancing the field of Explainable Artificial Intelligence (XAI) by providing a structured and semantically rich framework that enhances the interpretability of complex AI models. As Tiddi and Schlobach (2022) describe, knowledge graphs are typically structured as directed, edge-labeled graphs that describe entities and their relationships, often organized within an ontological schema covering a variety of topics. This organization reflects human cognitive processes of understanding and reasoning, making it a transparent layer in AI systems where the decisions of otherwise

opaque models can be traced and understood through clear, logical pathways. The primary value of knowledge graphs in explainability stems from their ability to connect AI outputs with comprehensible and verifiable pieces of information, offering deeper insights into the reasoning behind AI decisions. Therefore, knowledge graphs not only enhance the trustworthiness and clarity of AI systems but also play a significant role in bridging the gap between human understanding and machine reasoning, crucial for the broader adoption and ethical deployment of AI and path-based methods. Embedding-based approaches capitalize on the rich semantic information available within knowledge graphs to improve the representation of users and items. These methods may focus on creating item-specific graphs that enrich item representations with attributes from the knowledge graph or develop user-item graphs that include both entities and utilize the defined relationships and attributes to predict user preferences. Conversely, path-based methods leverage the connectivity patterns within the graph, utilizing paths to define and measure the similarity and interaction context between users and items. These methods often employ predefined meta-paths to capture semantic similarities and refine representations, thereby enhancing the recommendation process (Q. Guo et al., 2020). While these explanations significantly improve transparency, they often do not fully meet the human cognitive need for a deeper understanding. Users frequently question how recommendations might change if product attributes were altered, such as brand changes or descriptors like "animal cruelty-free" or "plant-based." This study aims to extend the level of explainability of path based kg recommendation through counterfactual analysis. Counterfactuals represent hypothetical alternatives to actual events, offering insights into what could have happened under different circumstances. These constructs are crucial in various domains of artificial intelligence (AI), especially in enhancing the interpretability and trustworthiness of AI systems—a fundamental aspect of explainable AI (XAI). Counterfactuals serve as a critical tool in XAI by providing "what if" analyses that elucidate the decision-making processes of AI models, showing how alternative inputs might alter an AI application of counterfactuals in XAI are present in both intrinsic and post hoc approaches. Intrinsic methods integrate

counterfactual reasoning as part of the AI model's fundamental design, directly influencing its operational framework. Conversely, post hoc techniques generate counterfactual explanations after the model has made a decision. By simulating different scenarios, counterfactuals provide insights into the specific conditions under which the outcomes of AI decisions might change, thereby highlighting the causal relationships within the model's reasoning. These "what if" scenarios do more than alter outcomes; they enrich the AI's decision-making process, allowing users to explore and understand the model's behavior under various hypothetical conditions. This aspect of counterfactuals in XAI not only aids in debugging and improving AI systems but also enhances user trust, making AI decisions more relatable and understandable, thereby fostering a deeper human-AI rapport Byrne (2019).

1.2 Research Problem and Objective

The central challenge addressed by this thesis is the limited explainability in path-based knowledge graph recommender systems. These systems are adept at stating the reasons for selecting a particular product but often lack the capability to delve into hypothetical "what-if" scenarios, such as the impact of different attributes on recommendations. Specifically, the systems do not answer questions like, "What if the recommended product had different attributes? Which attributes would still make it a plausible choice? Through what other attributes the recommended product would be still plausible" This thesis addresses this gap by applying counterfactual analysis to these systems, aiming to explore how hypothetical changes to product attributes and interactions could alter the generated recommendations, thereby enhancing the system's explainability. The primary objectives of this research are twofold:

1. **Development of an Analytical Model:** To create an analytical model that evaluates how changes in the attributes of recommended items affect the recommendation outcomes. This model is designed to identify attributes that would continue to uphold the recommendation's validity, providing deeper insights into the decision-making process of the recommender system.
2. **Marketing and User Experience Insights:** To leverage the insights gained from the analytical model to offer marketing intelligence, foster diversity in recommendations, and formulate user profiles that influence these recommendations. This objective aims to enhance the way businesses understand and cater to diverse

consumer preferences, thus improving user satisfaction and engagement.

By meeting these objectives, the research will not only advance the theoretical understanding of explainable AI in recommender systems but also equip businesses with practical tools to refine their marketing strategies and enhance recommendation processes.

1.3 Methodological Approach

The methodological approach developed in this thesis represents a novel strategy for enhancing the explainability of knowledge graph-based recommender systems through counterfactual analysis. This approach employs techniques in data processing and graph analytics to simulate potential scenarios under various attribute conditions. By manipulating attributes and observing the resultant changes in recommendations, this method offers a detailed view into the decision-making process of the recommender system.

Key Components

The methodology consists of several essential components:

1. **Attribute Extraction:** This process involves identifying and extracting relevant attributes and interactions of recommended products from the knowledge graph. This step is crucial for understanding the specific factors that might influence the recommendations.
2. **Community Identification:** Utilizing community detection algorithms helps manage the complexity and size of the data by grouping similar products and attributes. This organization enhances the relevance and efficiency of the counterfactual analysis by ensuring that the simulations are grounded in realistic and comparable scenarios.
3. **Recommendation Scoring:** This component evaluates each hypothetical scenario by calculating the recommendation scores for altered paths. It allows for a comparison with actual recommendation outcomes to determine if a product with changed attributes would still be recommended. This scoring process is integral in assessing the robustness and flexibility of the recommender system under varied conditions.

1.4 Expected Contributions and Outcomes

Contributions to the Field: This research contributes to the fields of recommender systems and explainable AI by introducing a counterfactual framework that incorporates counterfactual reasoning into the recommendation process. This adds a significant layer of depth to the analytical capabilities of existing systems and enhances the explainability of complex recommender systems, bridging the gap between advanced algorithmic decisions and user comprehensibility.

Anticipated Findings: The expected findings of this research are likely to show that certain attributes and their changes have significant impacts on recommendations, highlighting the importance of these attributes in the recommendation logic. By providing clearer explanations for recommendations, user satisfaction and trust in the system are expected to improve. Additionally, insights from counterfactual analyses can inform more targeted marketing strategies and product development initiatives, leading to better customer targeting and enhanced product offerings. These contributions and findings are anticipated to foster a deeper understanding of recommendation systems, promoting more informed and transparent use in various applications.

1.5 Contribution

This research advances the fields of recommender systems and explainable AI by integrating counterfactual reasoning into the path-based knowledge graph recommendation process. This integration introduces a depth to the analytical capabilities of existing systems and enhances the explainability of complex recommender systems, effectively bridging the gap It aligns algorithmic decision-making more closely with stakeholder needs and expectations, thereby increasing the practical utility and transparency of these systems.

1.6 Structure of the Thesis

This thesis is organized into several chapters, each serving a specific purpose to comprehensively cover the research undertaken. Here's a brief outline of each chapter and what it entails:

Chapter 1: Introduction

This opening chapter sets the stage for the research by introducing the topic, stating the research problem, and outlining the objectives. It provides a thorough background

on the importance of explainability in knowledge graph-based recommender systems and introduces the concept of counterfactual analysis as a novel approach to address these challenges.

Chapter 2: Literature Review

The literature review chapter delves into existing studies and theories relevant to knowledge graph recommender systems and explainable AI. It discusses previous approaches to improve transparency in AI-driven recommendations, evaluates their limitations, and highlights the need for innovative solutions like counterfactual analysis. This review establishes the theoretical foundation for the research and identifies gaps that this thesis aims to fill.

Chapter 3: Methodology

In this chapter, the research methodology used in the thesis is detailed. It describes the development of the counterfactual framework, including the processes of attribute extraction, community identification, and the computation of recommendation scores. The chapter elaborates on the tools and algorithms employed, the data collection process, and the techniques used to simulate and analyze various counterfactual scenarios.

Chapter 4: Results

The results chapter presents the findings from the application of the counterfactual analysis framework. It includes detailed discussions on how different attributes and their hypothetical modifications impact the recommendation outcomes. The results are supported by quantitative data and visual representations to illustrate the effects of attribute changes on the recommendations.

Chapter 5: Conclusion and Recommendations

The concluding chapter synthesizes the findings, discussing the implications for both theory and practice. It evaluates the success of the counterfactual framework in enhancing the explainability of recommender systems and suggests areas for future research. Recommendations for practitioners in the field of recommender systems are also provided, focusing on how they can implement similar methodologies to improve the transparency and effectiveness of their systems.

Chapter 2

Literature Review

2.1 Explainable AI In Knowledge Graph Recommender Systesm

Knowledge Graphs (KGs) are pivotal in enhancing the explainability and accuracy of recommender systems. These structured, relational frameworks capture complex interactions among users, items, and their attributes, allowing for more nuanced recommendations coupled with clear, logical explanations. This literature review synthesizes recent advancements in explainable artificial intelligence (XAI) that utilize KGs to illustrate how these technologies not only refine recommendation quality but also enhance user trust and understanding through transparency.

2.1.1 Path-based Modeling

Path-based modeling has emerged as a fundamental innovation in the utilization of KGs for recommender systems. Techniques such as the Knowledge-aware Path Recurrent Network (KPRN) and Path Language Modeling Recommendation (PLM-Rec) illustrate this trend's dynamic nature. KPRN leverages LSTM networks to interpret paths of entities and relationships, emphasizing those connections that are most influential in understanding user preferences. This method enriches the recommendation process by providing a temporal and semantic depth that traditional models lack, allowing for a better prediction of user behavior based on past interactions (X. Wang et al., 2019). On the other hand, PLM-Rec employs a novel approach by integrating natural language processing techniques to extend KG paths. This model treats paths as sentences, using

a language model to dynamically predict and extend these paths within the KG. Such extensions help the system explore new, potentially uncharted areas of the KG, thereby enhancing the system's ability to recommend items that were previously unreachable. This approach addresses the inherent limitations of static KG structures and improves the system's recall capabilities, making it particularly valuable for discovering long-tail items (Geng et al., 2022). Together, these path-based methods signify a shift towards more dynamic and exploratory use of KGs, expanding both the depth and breadth of what recommender systems can achieve.

2.1.2 Integration of User Profiles and Behavior

The integration of user behavioral data into KGs has significantly refined the personalization capabilities of recommender systems. The "Cafe" model by Xian et al. (2020) represents a sophisticated application of this concept, employing a coarse-to-fine strategy where initially broad user profiles help to narrow down and guide the path-finding algorithms in KGs. These profiles are crafted from historical data and are instrumental in focusing the recommendation process on paths most relevant to individual users, thus enhancing both the relevance and personalization of the recommendations. This method mirrors strategies used in other models that combine knowledge-base embeddings (KBE) with collaborative filtering. By embedding user behaviors and item characteristics into a unified representation, these models achieve a granular understanding of user-item relationships. This integration allows for a tailored recommendation experience, where the system's outputs are closely aligned with individual preferences and behaviors, as demonstrated in the work by Ai et al. (2018).

2.1.3 Explainability through Symbolic and Logical Reasoning

The demand for explainability in AI has driven the adoption of models that incorporate transparent, logical reasoning processes. Monotonic GNNs (MGNNs), introduced by Tena et al. (2022), exemplify this trend by ensuring that every transformation within the network adheres to a set of logical rules, akin to traditional rule-based systems. This adherence guarantees that the network's operations are interpretable and justifiable, enhancing user trust by providing comprehensible explanations for the recommendations made. Similarly, the Policy-Guided Path Reasoning (PGPR) model uses reinforcement learning to navigate through the KG, selecting paths that not only lead to relevant recommendations but are also interpretable. This model provides explicit paths that detail

the reasoning behind each recommendation, fulfilling the dual requirements of accuracy and transparency in the recommendation process (Xian et al., 2019).

The convergence of these methodologies highlights a crucial trend towards enhancing both the predictive accuracy and the interpretability of KG-based systems. Through the integration of dynamic path exploration, personalized user profile analysis, and logical reasoning, these approaches offer a more profound understanding of the intricacies involved in making recommendations. They collectively emphasize a shift towards recommender systems that are not only effective in their predictions but also provide transparent and understandable explanations, aligning with the growing user demand for transparency and accountability in AI systems.

2.2 Counterfacutal Reasoning in XAI

Counterfactuals in Explainable AI (XAI) play a crucial role in ensuring AI transparency and adherence to regulations such as GDPR introduced by Keane and Smyth (2020). Addressing the challenge of generating sparse and plausible counterfactual explanations, this study introduces a novel case-based reasoning (CBR) approach that utilizes a curated case-base of effective counterfactuals. In this system, each explanation case serves as a minimal yet impactful alteration of another, resulting in a different class outcome. The method involves identifying and pairing cases within datasets to build an explanation case-base. For new queries, counterfactuals are generated by retrieving and adapting the most similar case from this base, ensuring minimal changes and high plausibility. The generated counterfactuals are systematically validated and integrated throughout the experiments to assess the model's explanatory competence. By using the CBR model, this approach effectively enhances the quality and applicability of explanations, meeting the critical need for actionable counterfactuals in AI systems and enhancing their interpretability and trustworthiness.

Further enriching the field of counterfactual explanations in graph-based models, the CF-GNNExplainer model (Lucic et al., 2021) iteratively perturbs the adjacency matrix of a graph, typically through edge deletions, to generate counterfactual scenarios that influence the predictions made by Graph Neural Networks (GNNs). These counterfactuals help evaluate the minimal perturbations needed to effect a change in predictions, employing a loss function that delicately balances the change in prediction with the extent of the perturbation. This nuanced approach facilitates precise modifications to GNN outputs,

deepening the understanding of model behavior in graph-based data.

Counterfactuals have been used in explainable artificial intelligence (XAI) to enhance understanding and transparency in decision-making processes. As demonstrated by Jaimini and Sheth (2022) with their Causal Knowledge Graph (CausalKG) framework. This framework integrates a Causal Bayesian Network (CBN) and a hyper-relational graph representation using RDF-star to effectively model complex causal relationships. The central neural model, the Causal Bayesian Network, facilitates the manipulation of variables to generate counterfactuals by hypothesizing interventions in the CBN and observing the resultant effects on connected variables. This approach allows the system to predict outcomes under various scenarios. Integrating these counterfactuals into experiments not only aids in evaluating the causal impacts of different decisions but also ensures that AI outcomes are contextually relevant and more comprehensible, underscoring the value of counterfactual reasoning in making AI systems more transparent and understandable.

The CoCoX model, introduced by Akula et al. (2020), represents a notable advancement in enhancing the transparency and understandability of decisions made by deep convolutional neural networks (CNNs). This model generates both conceptual and counterfactual explanations by exploiting cognitive fault-lines—semantically significant features that are critical for classification. CoCoX identifies minimal changes, such as the addition or removal of specific features like stripes or bumps, that could influence the CNN's output. Based on the VGG-16 architecture, CoCoX utilizes Grad-CAM for effective feature extraction and K-means clustering to pinpoint explainable concepts. These concepts help define the positive and negative fault-lines which are crucial for developing counterfactuals. These counterfactuals are approached as an optimization problem that aims to modify the minimum number of features necessary to change the classification outcome.

For graph-based models, the CF2 framework addresses the unique challenge of enhancing graph-based model explainability by generating counterfactuals that identify necessary and sufficient conditions for predictions made by graph neural networks (GNNs). This framework solves this by integrating factual and counterfactual reasoning into an optimization problem, aiming to pinpoint the minimal sub-graph components whose alteration would change the outcome. Employing a novel loss function and relaxation techniques, CF2 balances the explanation's complexity and effectiveness, providing iterative refinement to focus on critical graph components that significantly impact the GNN's predictions. These insights are crucial for advancing the transparency and reliability of GNNs, ultimately enriching the field of explainable AI (Tan et al., 2022). Combined

with the aforementioned CoCoX model, these developments underscore the breadth of methodologies being pursued in the realm of explainable AI, reflecting a robust and multi-dimensional approach to understanding and improving the decision-making processes of complex models.

The CLEAR framework innovatively generates counterfactual explanations for graph-level prediction models by utilizing a graph variational autoencoder (VAE). This approach involves encoding the node features and graph structure into a latent space, and then decoding from this space to construct counterfactual graphs that minimally alter the original while achieving a specific prediction outcome. By mapping graphs into a latent space, CLEAR allows for both effective optimization and generalization across graphs, while an auxiliary variable enhances the model's ability to adhere to underlying causal relationships. This method significantly advances the generation and application of counterfactual explanations in graph data, outperforming existing techniques in validity, proximity, and causality (Ma et al., 2022).

This paper addresses the challenge of enhancing knowledge graph reasoning by employing counterfactual scenarios to identify crucial relationships within the graph. To tackle this issue, the approach involves generating counterfactuals by modifying relationships in factual reasoning paths and observing the resultant impacts on reasoning outcomes. These modifications help determine the importance of each relationship, which are then assigned weights. This weighted information is integrated as prior knowledge into a reinforcement learning-based reasoning model. Specifically, it utilizes a policy network that combines these weights with neural outputs to guide decision-making. The integration of counterfactual-derived weights with the neural model significantly enhances the reasoning capabilities of the system, as demonstrated through robust experimental validation across multiple large datasets. This method not only boosts performance but also aids in maintaining consistent reasoning across varied path lengths, thereby enriching the explainability and reliability of knowledge graph-based systems (Z. Wang et al., 2021)

Incorporating counterfactual reasoning with knowledge graph completion (KGC), the novel task of CounterFactual Knowledge Graph Reasoning (CFKGR) explores hypothetical alterations within a knowledge graph. The study introduces a neural model called COULDD (COUnterfactual Reasoning with KnowLedge Graph EmbeDDings), which adeptly refines existing knowledge graph embeddings to effectively manage hypothetical scenarios. Through the extraction of logical rules from the knowledge graph, counterfactuals are generated, creating scenarios that involve adding new edges and potentially

removing contradictory ones. These scenarios are then assimilated by updating the embeddings with counterfactual information and re-training the model to discern the validity of these new configurations. This innovative methodology not only enables the model to evaluate the legitimacy of counterfactual changes but also boosts its ability to navigate potential scenarios, thereby maintaining precision (Zellinger et al., 2024)

2.3 Counterfactual Methods in Different Types of Recommender Systems

Counterfactual reasoning in recommender systems has emerged as a pivotal technique within the domain of explainable artificial intelligence (XAI), enhancing both the transparency and fairness of recommendations. By modeling alternative scenarios where specific variables are modified, this approach provides insights into the potential impacts of different data configurations, helping to elucidate the inner workings and dependencies within these systems.

The introduction of the KGCR model (Y. Wei et al., 2023) marks a significant advancement in embedding causal inference within graph-based recommender systems. Utilizing Graph Convolutional Networks, this model enriches user, item, and attribute embeddings, which allow for a more nuanced understanding of user preferences. By constructing a causal graph and applying do-calculus interventions, the KGCR model effectively mitigates biases introduced by previous user interactions, offering a refined approach to understanding how bias influences recommendation outcomes.

In a similar vein, Tran et al. (2021) developed the ACCENT framework, which facilitates the generation of actionable counterfactual explanations in neural recommender systems. This framework leverages extended influence functions to explore how changes in user-item interactions could affect recommendation outputs, significantly enhancing computational efficiency through Fast Influence Analysis. This methodology underscores the minimal adjustments in user behavior that could lead to different recommendations, thereby aiding in the creation of more transparent recommendation mechanisms.

Addressing selection bias, (Liu et al., 2022) implemented counterfactual policy learning to recalibrate recommendation fairness and effectiveness. Their approach utilizes Inverse Propensity Scoring to weigh observed interactions, allowing the system to simulate outcomes under different recommendation policies. By integrating these counterfactual outcomes into the learning process, the model achieves an improved balance, enhancing

both the performance and equity of recommendations across various user groups and item categories.

The Prince method (Ghazimatin et al., 2020), emphasizes the importance of trust and understanding in recommendation systems through counterfactual reasoning within heterogeneous information networks. By identifying key user actions and employing Personalized PageRank, Prince efficiently predicts the impact of these actions on recommendation outcomes. This approach not only avoids exhaustive computations but also outperforms traditional heuristic methods in providing understandable and trust-enhancing explanations.

Yang et al. (2021) utilize causal inference through Structural Equation Models (SEMs) to address data sparsity in recommender systems. By generating counterfactual training samples, they enrich the dataset with diverse user responses that are otherwise not observed but plausible. This approach not only enhances the performance of the recommender systems but also strengthens their capacity to handle scenarios marked by data imbalance.

Finally, the Counterfactual Explainable Recommendation (CountER) model (Tan et al., 2021) focuses on identifying minimal attribute changes that could reverse a recommendation decision. Through a structured optimization process, CountER iteratively adjusts item attributes to discover the least extensive yet impactful changes required for altering outcomes. This model utilizes novel metrics to evaluate the necessity and sufficiency of these changes, demonstrating enhanced precision in providing actionable insights into recommendation decisions.

While counterfactual reasoning enhances the transparency and fairness of recommender systems, it also introduces potential vulnerabilities that can be exploited. A notable exploration of this issue is presented in the paper by Chen et al. (2023), titled "The Dark Side of Explanations: Poisoning Recommender Systems with Counterfactual Examples." This study reveals how counterfactual explanations (CFs) can be manipulated to deceive recommender systems. By employing a novel poisoning technique named H-CARS (Horn-Clause Attacks to Recommender Systems), which utilizes a neural model called Neural Collaborative Reasoning (NCR), the paper illustrates a method where logical reasoning through Horn clauses simulates the decision-making processes of recommender systems. Here, CFs are generated by pinpointing the minimal adjustments necessary in user-item interactions to alter the outcome of recommendations. These counterfactuals are critical in training a surrogate model, which then crafts targeted item embeddings and simulates user interactions to deceive the system. This approach not only highlights the

application and generation of CFs but also emphasizes their potential for both analysis and exploitation, particularly within environments sensitive to security.

The Counterfactual Explainable Recommendation (CERec) system by X. Wang et al. (2024) represents a significant advancement in the domain of explainable AI within recommendation systems. CERec employs reinforcement learning to navigate and optimize paths in a knowledge graph, focusing on attribute-based counterfactual explanations. It identifies counterfactual paths where minimal attribute adjustments lead to notable shifts in recommendation outcomes. An essential feature of CERec is its adaptive path sampler, which incorporates a two-step attention mechanism to efficiently manage the extensive search space of the knowledge graph. This selective exploration ensures computational efficiency and enhances the relevance of the explanations provided. Counterfactual scenarios are crafted by simulating changes in item attributes and observing the consequent alterations in recommendation outputs. These insights are then reintegrated into the recommendation model, improving its accuracy and reducing the incidence of undesirable recommendations. This makes CERec a pivotal tool in enhancing both transparency and decision-making in recommendation systems, contributing to more accurate and user-satisfying outcomes.

In conclusion, counterfactual reasoning offers a robust framework for enhancing the explainability and fairness of recommender systems by providing a deeper understanding of the implications of various data interactions and policies. These innovative approaches not only clarify the decision-making processes but also foster more equitable and user-centric recommendation practices.

2.4 Application of Counterfactuals for Fairness and Bias Mitigation

Counterfactual reasoning plays a pivotal role in the domain of explainable artificial intelligence (XAI), especially for mitigating biases in automated decision-making systems. This method involves hypothesizing alternative scenarios where key variables are altered, allowing for the exploration of how such changes impact outcomes. This not only uncovers hidden biases but also ensures fairness in AI operations. Broadly applied in various AI frameworks, from graph neural networks to recommender systems, counterfactual reasoning enhances transparency and equity in AI outcomes, establishing it as an essential tool for ethical AI development.

2.4.1 Mitigating Bias Across Different AI Frameworks

The use of counterfactual reasoning in graph-based models like those studied Z. Guo et al. (2023) demonstrates a rigorous approach to maintaining consistency in model predictions across varying sensitive attributes. By implementing Graph Variational Autoencoders (GraphVAE), they not only perturb attributes but also train the network to minimize discrepancies in outputs between the original and counterfactual nodes. This methodology effectively addresses biases at a fundamental level, ensuring the fairness of the model's outcomes. Medda et al. (2024) extend this approach within graph neural network-based recommender systems. Their innovative use of counterfactual reasoning to adjust user-item interactions on a bipartite graph includes strategically adding or removing connections, which serves to simulate various scenarios where demographic disparities can be analyzed and mitigated, ensuring a more equitable distribution of utility among users. The field of recommender systems frequently grapples with biases such as popularity and exposure, which can distort user preferences. T. Wei et al. (2021) dissect these issues through the Model-Agnostic Counterfactual Reasoning (MACR) framework, which explicitly separates the influence of item popularity from actual user preferences. By adjusting input data to simulate a scenario where item popularity is neutralized, MACR provides a recalibrated basis for recommendation, aligning more closely with unbiased user preferences. Meanwhile, Xu et al. (2020) focus on exposure bias by employing a counterfactual approach that involves a minimax adversarial model. This model simulates worst-case scenarios to test the resilience of the recommendation system, ensuring that it can withstand and adapt to a range of user exposure conditions, thus promoting a more fair and balanced recommendation landscape.

2.4.2 Enriching Data and Ensuring Equitable Outcomes

Addressing data sparsity and imbalance, Yang et al. (2021) utilize causal inference via Structural Equation Models (SEMs) to generate counterfactual scenarios that enrich training datasets. This not only addresses the immediate issue of insufficient data but also simulates a broader spectrum of user interactions, which helps in developing a more robust and responsive recommender system. On a more focused level, Chiappa (2019) pioneers the use of Path-Specific Counterfactual Fairness (PSCF) within decision-making processes. This approach manipulates causal pathways, particularly those that might be influenced by sensitive attributes such as race or gender, to ensure that resulting decisions are free from the undue influence of these attributes, thus promoting fairness in critical

decision-making contexts.

2.4.3 Leveraging Knowledge Graphs for Fair Recommendations

Expanding the utility of knowledge graphs, Balloccu et al. (2022) integrate counterfactual reasoning within the Policy-Guided Path Reasoning (PGPR) model to optimize recommendation systems. By re-ranking items and explanations based on various fairness-oriented criteria, such as recency, popularity, and diversity, PGPR enhances the quality and equity of recommendations. This approach not only improves the relevance of the recommendations but also significantly increases user trust and satisfaction by ensuring that recommendations cater equitably to diverse user groups.

Chapter 3

Research Design and Methodology

This chapter delineates the methodology employed to achieve the primary objective of this study: extending the explainability of path-based knowledge graph recommender systems to explore "what if" scenarios. The methodology is structured as follows:

1. **Initial Recommendation**: The process commences with the system generating a product recommendation for a user. In a path-based recommendation system, the inherent explanation typically comprises a sequence of entities and relationships that link the user to the recommended product.
2. **Counterfactual Analysis**:
   - **Extraction of Relevant Information**: Initially, the analysis extracts pertinent information related to the entities along the recommendation path. This includes the specific attributes that influenced the selection of the final product, as well as other potentially relevant attributes associated with the recommended product.
   - **Scenario Construction**: Utilizing the extracted information, a collection of hypothetical scenarios is constructed. These scenarios are crafted to test various alterations in attributes and their impact on the recommendation outcome.
3. **Recommendation System Utilization**: For the recommendation engine, we employ CAFE (Coarse-to-Fine Neural Symbolic Reasoning for Explainable Recommendation). This system is particularly suitable for our purposes due to its path-based nature and its capability to evaluate the plausibility of different paths by assigning probability scores to the steps that connect users to products.

This methodology both facilitates a deeper understanding of the decision-making processes inherent in the recommender system and also allows us to simulate and evaluate

how changes in product attributes or user-product relationships might alter the system's recommendations.

3.1 Recommender System Details

CAFE (Coarse-to-Fine Neural Symbolic Reasoning for Explainable Recommendation) is used as the foundational framework for our recommender system. This section provides an overview of its implementation and core functionalities.

3.1.1 Data and Implementation

The CAFE model is implemented using the Amazon review dataset, the beauty category, which includes comprehensive user and product interactions. It leverages predefined embeddings train in the model developed by Ai et al. (2018) described in the previous section, as input to their symbolic network.

3.1.2 Knowledge Graph Composition

The knowledge graph at the heart of this recommendation system is intricately structured, comprising several types of entities and relationships:

- **Users** are linked to the words they have used and the products they have purchased.
- **Products** are associated with descriptive words, their brand, category, and other related products. Relationships with related products include those that have been 'bought together', 'also viewed', and 'also bought'.
- **Brands and Categories** form additional nodes, creating multiple pathways that connect different aspects of the data.
- **Related categories** mentioned above.

3.1.3 Path-Based Recommendation Mechanics

The system operates on predefined metapaths that represent meaningful relationships leading a user to a product. These metapaths are substantial for understanding the logic behind the recommendations. Todo: That paths . The recommender system assigns a probability score to each step along the path, determining the strength of the connection between the user and the potential product recommendations. It then selects the top 10 paths with the highest cumulative scores, and the products associated with these

paths are recommended to the user. This scoring and selection process ensures that the recommendations are both relevant and tailored to the user's preferences and behavior patterns. This structured approach allows the CAFE system to recommend products effectively but also provide insights into the reasons behind each recommendation, providing explainability to the system.

3.2 Counterfactual Analysis

subsectionCommunity Detection and Graph Analysis Node FilteringThe counterfactual analysis begins with the detection of communities within the knowledge graph of interactions. Communities are identified using Louvain Method. This helps to cluster entities that share significant similarities and interactions. The Louvain Method is a an efficient algorithm designed for detecting communities in large-scale networks by optimizing modularity, a measure that quantifies the density of links within communities relative to those between them introduced in (Blondel et al., 2008). The algorithm operates in two iterative phases: initially, it optimizes modularity locally, evaluating potential gains by moving individual nodes into different communities. Nodes are shifted to the community that maximizes this gain, and the process is repeated until no further improvement is possible, achieving a local maximum of modularity. In the second phase, the method aggregates these identified communities into new nodes of a reduced network, and the process is reapplied. This hierarchical approach allows the algorithm to uncover community structures at multiple levels effectively. Notable for its speed, the Louvain Method can handle networks with up to millions of nodes efficiently, making it well-suited for modern datasets of substantial size. To refine the analysis further, we calculate the degree centrality for each type of pair within the graph. Nodes that do not provide significant insight are filtered out based on their z-score; specifically, nodes with a z-score exceeding [specific threshold] are removed.

3.2.1 Attribute Selection

Following the predictions provided by the recommender system, a threshold is set for the minimum score required for a product path to be recommended to a user, which is the path score of the last product recommended in the top 10 recommended products. For the analysis of a recommended path, we retrieve the first-level attributes and their

related products. This forms the initial layer of attribute selection, predicated on the hypothesis that first-level connected items possess more relevant attributes. These selected attributes are then evaluated to determine whether they fall within the community of the recommended product. If they do, and their z-score is within an acceptable range, they are considered for further counterfactual analysis.

3.2.2 Performing Counterfactual Analysis

For each attribute, an appropriate metapath is selected based on the type of the attribute. Using the recommender engine, we calculate the score for a user-product combination, which incorporates all the products previously purchased by the user and the counterfactual attribute in question. This approach is grounded in the assumption that the system discerns the user's preferences through their purchase history. If the recalculated score for a product, when considering a counterfactual attribute, exceeds the set minimum score, the attribute is considered a positive influence for a product similar to the recommended one. This isolated attribute analysis not only aids in understanding the influence of specific attributes on product recommendations but also provides marketing insights. Such insights can further be used to enhance the diversity and precision of the recommender system.

3.3 Case Study

Chapter 4

Results

Chapter 5

Conclusion

References

Ai, Q., Azizi, V., Chen, X., & Zhang, Y. (2018). Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms*, *11*(9), 137.

Akula, A., Wang, S., & Zhu, S.-C. (2020). CoCoX: Generating conceptual and counterfactual explanations via fault-lines. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(3), 2594–2601.

Balloccu, G., Boratto, L., Fenu, G., & Marras, M. (2022). Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 646–656.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 6276–6282.

Chen, Z., Silvestri, F., Wang, J., Zhang, Y., & Tolomei, G. (2023). The dark side of explanations: Poisoning recommender systems with counterfactual examples. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2426–2430.

Chiappa, S. (2019). Path-specific counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 7801–7808.

Geng, S., Fu, Z., Tan, J., Ge, Y., De Melo, G., & Zhang, Y. (2022). Path language modeling over knowledge graphsfor explainable recommendation. *Proceedings of the ACM Web Conference 2022*, 946–955.

Ghazimatin, A., Balalau, O., Saha Roy, R., & Weikum, G. (2020). PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems.

Proceedings of the 13th International Conference on Web Search and Data Mining, 196–204.

Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., & He, Q. (2020). A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3549–3568.

Guo, Z., Li, J., Xiao, T., Ma, Y., & Wang, S. (2023). Towards fair graph neural networks via graph counterfactual. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 669–678.

Jaimini, U., & Sheth, A. (2022). CausalKG: Causal knowledge graph explainability using interventional and counterfactual reasoning. *IEEE Internet Computing*, *26*(1), 43–50.

Keane, M. T., & Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In *Case-based reasoning research and development* (pp. 163–178, Vol. 12311). Springer International Publishing.

Liu, Y., Yen, J.-N., Yuan, B., Shi, R., Yan, P., & Lin, C.-J. (2022). Practical counterfactual policy learning for top-k recommendations. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1141–1151.

Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M., & Silvestri, F. (2021). CF-GNNExplainer: Counterfactual explanations for graph neural networks.

Ma, J., Guo, R., Mishra, S., Zhang, A., & Li, J. (2022). CLEAR: Generative counterfactual explanations on graphs.

Medda, G., Fabbri, F., Marras, M., Boratto, L., & Fenu, G. (2024). GNNUERS: Fairness explanation in GNNs for recommendation via counterfactual reasoning. *ACM Transactions on Intelligent Systems and Technology*, 3655631.

Tan, J., Geng, S., Fu, Z., Ge, Y., Xu, S., Li, Y., & Zhang, Y. (2022). Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. *Proceedings of the ACM Web Conference 2022*, 1018–1027.

Tan, J., Xu, S., Ge, Y., Li, Y., Chen, X., & Zhang, Y. (2021). Counterfactual explainable recommendation. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1784–1793.

Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, *302*, 103627.

Tran, K. H., Ghazimatin, A., & Saha Roy, R. (2021). Counterfactual explanations for neural recommenders. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1627–1631.

Vultureanu-Albisi, A., & Badica, C. (2021). Recommender systems: An explainable AI perspective. *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 1–6.

Wang, X., Wang, D., Xu, C., He, X., Cao, Y., & Chua, T.-S. (2019). Explainable reasoning over knowledge graphs for recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 5329–5336.

Wang, X., Li, Q., Yu, D., Li, Q., & Xu, G. (2024). Reinforced path reasoning for counterfactual explainable recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 1–17.

Wang, Z., Li, L., Zeng, D., & Wu, X. (2021). Incorporating prior knowledge from counterfactuals into knowledge graph reasoning. *Knowledge-Based Systems*, *223*, 107035.

Wei, T., Feng, F., Chen, J., Wu, Z., Yi, J., & He, X. (2021). Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1791–1800.

Wei, Y., Wang, X., Nie, L., Li, S., Wang, D., & Chua, T.-S. (2023). Causal inference for knowledge graph based recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *35*(11), 11153–11164.

Xian, Y., Fu, Z., Muthukrishnan, S., De Melo, G., & Zhang, Y. (2019). Reinforcement knowledge graph reasoning for explainable recommendation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 285–294.

Xian, Y., Fu, Z., Zhao, H., Ge, Y., Chen, X., Huang, Q., Geng, S., Qin, Z., De Melo, G., Muthukrishnan, S., & Zhang, Y. (2020). CAFE: Coarse-to-fine neural symbolic reasoning for explainable recommendation. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1645–1654.

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., & Achan, K. (2020). Adversarial counterfactual learning and evaluation for recommender system.

Yang, M., Dai, Q., Dong, Z., Chen, X., He, X., & Wang, J. (2021). Top-n recommendation with counterfactual user preference simulation.

Zellinger, L., Stephan, A., & Roth, B. (2024). Counterfactual reasoning with knowledge graph embeddings.

Appendix

You are encouraged to put in appendices in your final report. In an appendix you can include things such as large tables or background information. Anything that is useful to know for the reader, but prevents the reader to read your main text in a fluent manner. Each appendix should have a number and a self-explanatory title.