Knowledge-aware Recommendations: Exploring the Interplay between Utility, Explanation Quality, and Fairness in Path Reasoning Methods

Discussion Paper

Giacomo **Balloccu***, Ludovico **Boratto**, Christian **Cancedda**, Gianni **Fenu** and Mirko **Marras**

Department of Mathematics and Computer Science, University of Cagliari, V. Ospedale 72, 09124 Cagliari, Italy

Abstract

Adopting Knowledge Graphs (KGs) in recommender systems has engendered the emergence of sophisticated techniques, such as path reasoning, designed to navigate KGs and model complex relationships. KGs enable the representation of intricate connections, while path reasoning approaches adeptly learn to traverse these graphs, constructing detailed user-product relationships by discerning reasoning paths linking recommended products with those previously experienced by users. These identified paths are subsequently converted into well-articulated textual explanations, facilitating a deeper and more comprehensive understanding for the users. Despite its potential, the field is hindered by disparate and insufficient evaluation protocols, complicating efforts to assess the impact of existing methodologies. In this paper, we summarize our previous work on replicating and evaluating three state-of-the-art path reasoning recommendation approaches, originally presented at prestigious conferences, using a standardized protocol based on two publicly available datasets and benchmarking them against other knowledge-aware techniques. Our analysis encompasses recommendation utility, explanation quality, and fairness considerations for both consumers and providers. This investigation offers a comprehensive overview of the progress in the field, emphasizing key challenges and potential avenues for future exploration. Source code is available at https://github.com/giacoballoccu/rep-path-reasoning-recsys.

Keywords

Recommender Systems, Knowledge Graphs, Replicability, Evaluation

1. Introduction

Recommender systems (RS) have become a prevalent approach for facilitating personalized user experiences. Traditional RSs are trained using historical data, such as browsing activity and ratings, as well as product characteristics, such as their textual description. To augment product information, Knowledge Graphs (KGs) have been employed as an additional data source [1, 2]. KGs encompass entities (e.g., users, movies, actors) and relations between entities

(e.g., an actor starring in a movie). The integration of KGs into RSs has resulted in improved recommendation utility [3, 4], particularly in the context of sparse data and cold-start situations [5]. More importantly, incorporating KGs in RSs is crucial for rendering RSs more explainable and fostering a transparent social recommendation process [6, 7].

Path reasoning methods stand out among KG-based recommendation techniques, as they utilize high-order relations between users and products to guide RS training and generate explanations [8, 9, 10, 11, 12, 13]. By identifying relevant paths between experienced and recommended products, these methods create explanations through templates [14, 15] or text generation. For instance, in the movie domain, the path "$user_1$ watched $movie_1$ directed $director_1$ directed^{-1} $movie_2$" might lead to the template-based explanation "$movie_2$ is recommended to you because you watched $movie_1$ also directed by $director_1$". In contrast, embedding-based methods weigh product characteristics without providing explanations [16, 4, 3].

Numerous KG- and path-reasoning-based methods have been proposed for recommendation and explanation generation [6]. However, evaluation protocols are inconsistent and limited, focusing primarily on recommendation utility. Prior works compare novel methods to (non) knowledge-aware baselines without (i) thoroughly examining beyond utility goals, (ii) considering the quality of the produced explanations or fairness dimensions. This complex landscape is requiring a common evaluation ground to determine when and how each method can be effectively adopted, as the potential trade-offs between unexplored goals remain unclear.

In this paper, we summarize our prior work [17] on conducting a replicability study of path reasoning methods, focusing on unexplored evaluation perspectives. We initially reviewed top-tier conference and journal proceedings, identifying seven relevant papers, but only three methods were replicable using the released source code. We then established a common evaluation protocol, encompassing two public datasets with two user sensitive attributes, and sixteen metrics across four perspectives. Path reasoning methods were assessed under this protocol and compared to other knowledge-aware methods. Our results indicate that despite similar utility, these methods differ in achieving other recommendation goals. This study highlights the need for broader evaluation and responsible adoption of path reasoning methods.

2. Research Methodology

Paper Collection. To gather existing path reasoning methods, we systematically reviewed proceedings from top-tier information retrieval events and journals published by leading publishers. We employed keywords combining technical and non-technical terms and identified papers that addressed recommendation methods augmented with KGs, and produce reasoning paths. We excluded papers on other domains or tasks, as well as knowledge-aware methods that could not generate reasoning paths. Seven relevant papers were selected for our study.

Methods Replicability. For each relevant paper, we analyzed the rationale of the proposed method and the components of the experimental setting, as summarized in Table 1. We then attempted to replicate each method using the original source code, making necessary changes to accommodate different datasets and extract recommendations and reasoning paths. Out of the seven papers, three (PGPR, UCPR, CAFE) were replicable with reasonable effort, while EKAR, PLM-Rec, MLR and TAPR were not reproducible due to the unavailability of the original source

Table 1
Path reasoning methods deemed as relevant in our study, where Status .

Method	Year	Status[1]	Experimental Setting				
			Data Sets[2]	Split Size[3]	Split Method[4]	Recommendation[5]	Explanation[5]
PGPR [8]	2019	RE	AZ	70-00-30	$Rand$	NDCG, R, HR, P	-
EKAR [9]	2019	$R\overline{E}$	ML, LFM, DB	60-20-20	$Rand$	NDCG, HR	-
CAFE [10]	2020	RE	AZ	70-00-30	$Rand$	NDCG, R, HR, P	-
UCPR [11]	2021	RE	ML, AZ	60-20-20	$Rand$	NDCG, R, HR, P	PPC
MLR [12]	2022	$R\overline{E}$	AZ	70-00-30	$Rand$	NDCG, R, HR, P	-
PLM-Rec [18]	2022	$\overline{R}\overline{E}$	AZ	60-20-20	$Time$	NDCG, R, HR, P	-
TAPR [13]	2022	$\overline{R}\overline{E}$	AZ	60-10-30	$Rand$	NDCG, R, HR, P	-

[1] **Status** RE : Replicable and Extensible; $R\overline{E}$: Replicable but not Extensible; $\overline{R}\overline{E}$: Not Replicable nor Extensible.
[2] **Data Set** AZ : Amazon [19]; ML : MovieLens 1M [?]; LFM : LastFM [?]; DB : DBbook2014 [20].
[3] **Split Size** reports the percentage of data for training, validation, and test, respectively.
[4] **Split Method**. $Rand$: Random based; $Time$: Time based.
[5] **Metrics** R : Recall; HR : Hit Ratio P : Precision; PPC : Path Pattern Concentration

code and, in the case of MLR, due to unavailability of external dependencies.

Evaluation Protocol. We combined replication and reproduction to create a unified evaluation framework for assessing path reasoning methods on two public datasets, MovieLens-1M (ML1M) and LastFM-1M (LFM1M). We relied on original source codes, our data processing, and computed evaluation metrics based on returned recommendations and paths. The datasets were chosen due to the availability of demographic attributes (Gender, Age) for fairness assessment.

We preprocessed and prepared the data, and performed a temporal-based training-validation-test split for each dataset. Additionally, we compared path reasoning methods against non-explainable knowledge-aware models (namely, CKE, CFKG, and KGAT), which were replicated and evaluated under the same protocol. Hyperparameters were fine-tuned using grid search.

The evaluation metrics computed for each model include recommendation utility (NDCG and MRR [21]), beyond utility objectives (COVerage, DIVersity, SERendipity, and NOVelty) and both consumer and provider fairness. For explanation path quality, based on [14], we measured perspectives related to recency (R), popularity (P), and diversity (D) of different path portions, named Linked Interaction (LI), Shared Entity (SE), and Path Type (PT). The detailed procedure, including metric definitions and implementation details, can be found in our original work [17].

3. Experimental Results

Trading Recommendation Goals for Explanation Power (RQ1). First, we compared path reasoning methods (PGPR, CAFE, UCPR) with knowledge-aware but non-explainable methods (KGAT, CKE, CFKG) in terms of recommendation utility, beyond utility objectives, and fairness. The goal is to assess if any trade-off between explainable power (i.e., being able to produce a textual explanation) and other goals exists. The results, reported in Table 2, showed that path reasoning methods achieved comparable recommendation utility to knowledge-aware methods in the ML1M dataset, but lower utility in the LFM1M dataset. However, path reasoning methods generally outperformed knowledge-aware methods in terms of serendipity, diversity, and provider fairness. Overall, path reasoning methods sacrificed recommendation utility and coverage for increased explanation power, particularly in the LFM1M dataset.

Producing Explanations for All Recommended Products (RQ2). We then investigated

Table 2

Metric scores for recommendation utility and beyond utility goals [RQ1].

Method	ML1M							LFM1M						
	NDCG ↑	MMR ↑	SER ↑	DIV ↑	NOV ↑	PEXP[1] $\downarrow_0$	COV ↑	NDCG ↑	MMR ↑	SER ↑	DIV ↑	NOV ↑	PEXP[1] $\downarrow_0$	COV ↑
CKE	**0.29**	**0.23**	0.26	0.10	**0.93**	**0.19**	<u>0.70</u>	**0.40**	**0.34**	<u>0.82</u>	0.18	**0.88**	**0.18**	**0.91**
CFKG	<u>0.26</u>	0.21	0.11	0.11	0.92	0.25	0.16	0.13	0.10	0.04	0.27	0.86	0.34	0.02
KGAT	**0.29**	**0.23**	0.29	0.10	**0.93**	**0.19**	**0.75**	<u>0.37</u>	<u>0.31</u>	0.79	0.19	**0.88**	**0.18**	<u>0.89</u>
PGPR	0.28	0.21	**0.78**	0.42	**0.93**	0.27	0.42	0.31	0.25	0.81	0.54	0.82	0.32	0.20
UCPR	<u>0.26</u>	0.20	0.53	<u>0.42</u>	**0.93**	0.22	0.25	0.34	0.27	**0.94**	<u>0.57</u>	<u>0.87</u>	0.22	0.41
CAFE	<u>0.26</u>	0.18	<u>0.63</u>	**0.44**	**0.93**	0.36	0.21	0.15	0.09	0.75	**0.58**	0.84	0.36	0.11

For each dataset: best result in **bold**, second-best result <u>underlined</u>. [1] **Metrics** $PEXP$: Provider Exposure.

Table 3

Explanation fidelity analysis across cut-offs $k = \{10, 20, 50, 100\}$ [RQ2].

Method	ML1M				LFM1M			
	10	20	50	100	10	20	50	100
PGPR	1.00	0.99	0.99	0.78	0.98	0.74	0.31	0.15
CAFE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
UCPR	0.61	0.34	0.14	0.07	0.99	0.98	0.68	0.35

the ability of path reasoning methods (PGPR, UCPR, CAFE) of producing explanations for recommended products across different list sizes. To this end, as a metric, we consider fidelity [6], i.e., the percentage of explainable items among the recommended items. Results are reported in Table 3. We found varying patterns of fidelity across methods: CAFE maintained high fidelity across datasets and list sizes, while PGPR and UCPR showed different fidelity patterns depending on the dataset. This suggests that CAFE is the best choice when list size is known in advance, constant, or up to a certain limit. While we conjecture that the ability to explain of the other methods is highly influenced by properties of the data, such as KG sparsity.

Differences on Explanation Quality (RQ3). Lastly, we explored how the quality of selected paths and resulting explanations varies based on path characteristics in path reasoning methods (PGPR, UCPR, CAFE). We considered seven explanation path quality perspectives and found that the methods often yield substantially different paths in terms of recency, popularity, and diversity. However, no remarkable disparate impacts on explanation quality were observed. This analysis highlights the importance of understanding the specific characteristics of each path reasoning method when selecting one for a particular application.

4. Conclusions

In this paper, we compared path reasoning methods and knowledge-aware non-explainable baselines in terms of utility, beyond utility perspectives, and fairness objectives. Results indicate that path reasoning methods slightly divergence in terms of utility and coverage but result in higher serendipity and diversity. We also investigated the ability of path reasoning methods to produce explanations across various recommended list sizes and found that model design choices can influence this capability. Lastly, we examined the quality of reasoning paths and empirically show that not all goals can be met simultaneously, aligning to [7, 22].

In the next steps, we plan to explore in detail the impact of KG characteristics on the considered perspectives, as well as devise novel path reasoning methods robust to the KG structure and effective on multiple objectives. especially on the quality of the provided textual explanations.

References

[1] Y. Cao, L. Hou, J. Li, Z. Liu, Neural collective entity linking, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 675–686.

[2] S. Oramas, V. C. Ostuni, T. Di Noia, X. Serra, E. Di Sciascio, Sound and music recommendation with knowledge graphs, ACM Trans. Intell. Syst. Technol. 8 (2017) 21:1–21:21.

[3] X. Wang, X. He, Y. Cao, M. Liu, T.-S. Chua, Kgat: Knowledge graph attention network for recommendation, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 950–958.

[4] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, M. Guo, Ripplenet: Propagating user preferences on the knowledge graph for recommender systems, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 417–426.

[5] C. Huang, Z. Gan, F. Ye, P. Wang, M. Zhang, KNCR: knowledge-aware neural collaborative ranking for recommender systems, in: IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, DASC/PiCom/CBDCom/CyberSciTech 2020, Calgary, AB, Canada, August 17-22, 2020, IEEE, 2020, pp. 339–344.

[6] Y. Zhang, X. Chen, Explainable recommendation: A survey and new perspectives, Foundations and Trends® in Information Retrieval 14 (2020) 1–101.

[7] N. Tintarev, J. Masthoff, A survey of explanations in recommender systems, in: Proceedings of the 23rd International Conference on Data Engineering Workshops, ICDE 2007, IEEE Computer Society, 2007, pp. 801–810.

[8] Y. Xian, Z. Fu, S. Muthukrishnan, G. de Melo, Y. Zhang, Reinforcement knowledge graph reasoning for explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 285–294.

[9] W. Song, Z. Duan, Z. Yang, H. Zhu, M. Zhang, J. Tang, Ekar: An explainable method for knowledge aware recommendation, CoRR abs/1906.09506 (2022). `arXiv:1906.09506`.

[10] Y. Xian, Z. Fu, H. Zhao, Y. Ge, X. Chen, Q. Huang, S. Geng, Z. Qin, G. de Melo, S. Muthukrishnan, Y. Zhang, Cafe: Coarse-to-fine neural symbolic reasoning for explainable recommendation, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1645–1654.

[11] C.-Y. Tai, L.-Y. Huang, C.-K. Huang, L.-W. Ku, User-Centric Path Reasoning towards Explainable Recommendation, Association for Computing Machinery, New York, NY, USA, 2021, p. 879–889.

[12] X. Wang, K. Liu, D. Wang, L. Wu, Y. Fu, X. Xie, Multi-level recommendation reasoning over knowledge graphs with reinforcement learning, in: Proceedings of the ACM Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 2098–2108.

[13] Y. Zhao, X. Wang, J. Chen, Y. Wang, W. Tang, X. He, H. Xie, Time-aware path reasoning on knowledge graph for recommendation, ACM Trans. Inf. Syst. (2022). Just Accepted.

[14] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Reinforcement recommendation reasoning through knowledge graphs for explanation path quality, Knowledge-Based Systems 260 (2023) 110098.

[15] C. Upadhyay, H. Abu-Rasheed, C. Weber, M. Fathi, Explainable job-posting recommendations using knowledge graphs and named entity recognition, in: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021, pp. 3291–3296. doi:`10.1109/SMC52423.2021.9658757`.

[16] F. Zhang, N. J. Yuan, D. Lian, X. Xie, W.-Y. Ma, Collaborative knowledge base embedding for recommender systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 353–362.

[17] G. Balloccu, L. Boratto, C. Cancedda, G. Fenu, M. Marras, Knowledge is power, understanding is impact: Utility and beyond goals, explanation quality, and fairness in path reasoning recommendation, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 3–19.

[18] S. Geng, Z. Fu, J. Tan, Y. Ge, G. de Melo, Y. Zhang, Path language modeling over knowledge graphsfor explainable recommendation, in: Proceedings of the ACM Web Conference 2022, WWW '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 946–955.

[19] G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet Comput. 7 (2003) 76–80.

[20] Y. Cao, X. Wang, X. He, Z. Hu, C. Tat-seng, Unifying knowledge graph learning and recommendation: Towards a better understanding of user preference, in: WWW, 2019.

[21] N. Craswell, Mean Reciprocal Rank, Springer US, 2009, pp. 1703–1703.

[22] G. Balloccu, L. Boratto, G. Fenu, M. Marras, Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 646–656.