LLM-Augmented Knowledge-Graph-Based Recommendation System

By

Kartikey Chauhan

501259284

Literature Review &

Exploratory Data Analysis

Master of Science

in the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2024

Contents

List of Figures

List of Tables

1 Introduction

This document covers the Introduction, Literature Review and Exploratory Data Analysis for the first deliverable of Major Research Project (MRP). It begins with a brief background on the topic and datasets, defines the problem, and states the research question. This is followed by a literature review and a detailed exploratory analysis of the dataset.

1.1 Background

Recommender systems have been widely applied to address the issue of information overload in various internet services, exhibiting promising performance in scenarios such as e-commerce platforms and media recommendations. In the general domain, the traditional knowledge recommendation method is *collaborative filtering (CF)*, which usually suffers from the cold start problem and sparsity of user-item interactions. Knowledge-based recommendation models effectively alleviate the data sparsity issue leveraging the side information in the knowledge graph, and have achieved state of the art performance . However, KGs are difficult to construct and evolve by nature, and existing methods often lack considering textual information. On the other hand, LLMs are black-box models, which often fall short of capturing and accessing factual knowledge. Therefore, it is complementary to unify LLMs and KGs together and simultaneously leverage their advantages. This project aims to explore LLM-augmented KGs, that leverage Large Language models (LLM) for different KG tasks such as embedding, completion, construction and also incorporate textual information which could be a way to help overcome these challenges and lead to better recommendation systems.

1.2 Research Objectives

The research objectives of this project are to investigate the use of Large Language Models (LLMs) to enhance the construction, quality, and volume of information in knowledge graphs (KGs). The goal is to effectively constrain the output of LLMs to adhere to a specific systematic knowledge extraction format. Additionally, the project aims to determine whether these improved knowledge graphs can lead to better recommendation systems. Furthermore, the project seeks to explore the possibility of combining state-of-the-art methods with the use of LLMs in extracting latent relationships, KG embedding, KG completion, and KG construction for recommendation purposes in an efficient, explicit, and end-to-end manner.

2 Literature Review

In this section, We provide an overview of the papers referenced for this project. Knowledge Graphs (KGs) as a form of structured knowledge have drawn significant attention from academia and the industry (Ji et al. (2022) [4]). There have been several efforts to construct KGs to facilitate the discovery of relevant information within specific fields. Most of these efforts have focused on extracting information from text.

In recommendation, KGs have been used to enhance the performance of recommendation systems by incorporating high-order connectivities from KGs into user-item interactions. Wang et al. (2019) [6] introduce the **Knowledge Graph Attention Network (KGAT)**, which enhances recommendation systems by leveraging an attention mechanism to discern the significance of various neighbor connections, demonstrating superior performance and interpretability compared to existing models such as Neural FM and RippleNet through extensive experiments on multiple public benchmarks. The model's end-to-end approach efficiently captures and utilizes high-order relations, providing more accurate, diverse, and explainable recommendations.

Guo et al. (2020) [2] present a comprehensive survey of knowledge graph embedding techniques, which have been widely applied in various tasks such as recommendation, search, and question answering. The survey categorizes embedding methods into three groups: translation-based, semantic matching-based, and neural network-based. The authors provide a detailed overview of each category, discussing their strengths, weaknesses, and applications. The survey also highlights the challenges and future directions in knowledge graph embedding research, emphasizing the importance of incorporating textual information to enhance the quality and interpretability of embeddings.

He et al. (2020) [3] propose LightGCN, a lightweight graph convolutional network that simplifies the design of graph neural networks for collaborative filtering. LightGCN eliminates the feature transformation and nonlinear activation functions in traditional GCNs, focusing solely on the graph structure. The model achieves state-of-the-art performance on several recommendation benchmarks, outperforming more complex models such as NGCF and GAT. LightGCN's simplicity and efficiency make it an attractive choice for large-scale recommendation systems, demonstrating the effectiveness of collaborative filtering with graph neural networks.

Zhang et al. (2021) [9] introduce the Knowledge Graph Embedding Transformer (KGET), a novel model that leverages the transformer architecture to learn embeddings for knowledge graphs.

KGET incorporates a self-attention mechanism to capture complex relational patterns and dependencies in the graph structure. The model outperforms existing embedding methods such as TransE, DistMult, and ComplEx on several knowledge graph completion tasks, demonstrating its effectiveness in capturing long-range dependencies and semantic relationships. KGET's ability to model complex interactions between entities and relations makes it a promising approach for knowledge graph embedding.

However, the traditional KG construction methods often lack the ability to incorporate textual information, which is essential for capturing the rich semantics and context of entities and relations.

The emergence of Large Language Models (LLMs) has revolutionized research and practical applications by enabling complex reasoning and task generalization through techniques like In-Context Learning and Chain-of-Thought. LLMs offer promising solutions to existing recommender system challenges, such as poor interactivity, explainability, and the cold start problem, by generating more natural and cross-domain recommendations and enhancing user experience through stronger feedback mechanisms.

As such, the integration of LLMs with KGs presents a novel direction to overcome the limitations of traditional KGs, such as the challenge of incorporating textual information.

Several recent studies have explored the integration of of language models with knowledge graphs to enhance the quality and efficiency of knowledge representation and recommendation systems.

Xu et al. (2021) [7] propose a novel method for constructing knowledge graphs from text, called Text2KG. Text2KG utilizes a pre-trained language model to extract structured knowledge from unstructured text data, generating entity and relation triples for constructing knowledge graphs. The model achieves competitive performance on knowledge graph construction tasks, outperforming existing methods such as OpenIE and ReVerb. Text2KG's ability to extract high-quality knowledge from text data demonstrates its potential for automating the construction of knowledge graphs from large-scale text corpora.

Zhang et al. (2021) [8] introduce KG-BERT, a pre-trained language model that incorporates knowledge graph embeddings to enhance the representation learning of entities and relations. KG-BERT leverages the pre-trained BERT model to capture contextual information from text data and knowledge graph embeddings to capture structured information from knowledge graphs. The model achieves state-of-the-art performance on several knowledge graph completion tasks, demonstrating its effectiveness in capturing both textual and structured information. KG-BERT's ability to leverage both text and knowledge graph embeddings makes it a promising approach for enhancing the quality and interpretability of knowledge graph embeddings.

Ullah et al. (2021) [5] introduce a novel method for knowledge graph completion using large language models, called LLM-KGC. LLM-KGC leverages the pre-trained language model BERT to predict missing relations in knowledge graphs, capturing complex relational patterns and dependencies. The model outperforms existing knowledge graph completion methods such as TransE, DistMult, and ComplEx on several benchmark datasets, demonstrating its effectiveness in capturing long-range dependencies and semantic relationships. LLM-KGC's ability to leverage large language models for knowledge graph completion makes it a promising approach for enhancing the quality and completeness of knowledge graphs.

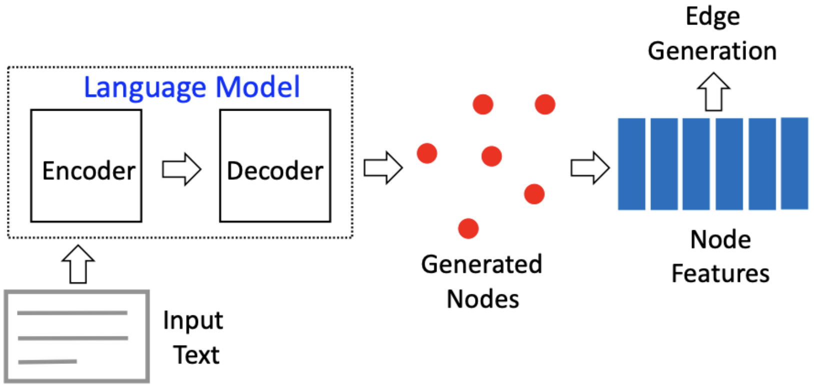

Figure 1: High level overview of constructing Knowledge Graphs from Language models

Gao et al. (2023) [1] propose a novel paradigm called Chat-Rec, which augments large language models (LLMs) to build conversational recommender systems. This system converts user profiles and historical interactions into prompts, making the recommendation process more interactive and explainable. Chat-Rec is effective in learning user preferences and establishing connections between users and products through in-context learning. Moreover, it addresses challenges such as cold-start scenarios with new items and cross-domain recommendations, demonstrating improved performance in top-k recommendations and zero-shot rating prediction tasks.

Each of these papers contributes to the field of knowledge graph construction, embedding, and recommendation systems by introducing novel methods and techniques to enhance the quality and efficiency of knowledge representation and recommendation. By leveraging the power of large language models and graph neural networks, these models demonstrate the potential to improve the performance and interpretability of recommendation systems, paving the way for more effective and scalable knowledge-based recommendations.

3 Exploratory Data Analysis

3.1 Exploratory Data Analysis

This section aims to provide a comprehensive understanding of the dataset used for the research project. It contains information on the data source and files, as well as a description of the data's basic features.

By performing exploratory data analysis (EDA), we aim to gain a deeper understanding of the data, including the relationship between variables and identifying trends. This analysis will inform the subsequent steps in the research and help address the research questions effectively.

3.2 Data Source and Files

The primary dataset in scope is Amazon Reviews'23. This is a large-scale Amazon Reviews dataset, collected in 2023 by McAuley Lab, and it includes rich features such as:

- User Reviews (ratings, text, helpfulness votes, etc.);
- Item Metadata (descriptions, price, raw image, etc.);
- Links (user-item / bought together graphs).

The datasets are open-sourced and compliant with the MRP requirements.

The reviews span from May'96 to Sep'24 and cover a wide range of categories, including electronics, books, movies, and more. The dataset is designed to facilitate research in recommendation systems, natural language processing, and other related fields.

3.3 Data Description

For each category in the dataset, there are two main files: *User Reviews* and *Item Metadata*. The User Reviews file contains information about the reviews posted by users, including ratings, text, helpfulness votes, and more. The Item Metadata file contains information about the items being reviewed, such as descriptions, prices, images, and more.

3.3.1 For User Reviews

Field	**Type**	**Explanation**
rating	float	Rating of the product (from 1.0 to 5.0).
title	str	Title of the user review.
text	str	Text body of the user review.
images	list	Images that users post after they have received the product. Each image has different sizes (small, medium, large), represented by the small_image_url, medium_image_url, and large_image_url respectively.
asin	str	ID of the product.
parent_asin	str	Parent ID of the product.
user_id	str	ID of the reviewer.
timestamp	int	Time of the review (unix time).
verified_purchase	bool	User purchase verification.
helpful_vote	int	Helpful votes of the review.

Table 1: User Reviews Data Fields

3.3.2 For Item Metadata

3.4 Data Analysis

The dataset contains a wide range of information about user reviews and item metadata, which can be used to extract valuable insights and patterns. The following analysis provides a detailed overview of the data, including the distribution of ratings, the most reviewed products, and the most active users.

Field	**Type**	**Explanation**
main_category	str	Main category (i.e., domain) of the product.
title	str	Name of the product.
average_rating	float	Rating of the product shown on the product page.
rating_number	int	Number of ratings in the product.
features	list	Bullet-point format features of the product.
description	list	Description of the product.
price	float	Price in US dollars (at time of crawling).
images	list	Images of the product. Each image has different sizes (thumb, large, hi_res). The "variant" field shows the position of image.
videos	list	Videos of the product including title and url.
store	str	Store name of the product.
categories	list	Hierarchical categories of the product.
details	dict	Product details, including materials, brand, sizes, etc.
parent_asin	str	Parent ID of the product.
bought_together	list	Recommended bundles from the websites.

Table 2: Item Metadata Fields

We limit our analysis to the Video Games category for the purpose of this document, due to the large size of the dataset and the need to focus on a specific category for detailed analysis.

3.4.1 Ratings Distribution

The ratings distribution of the user reviews provides insights into the overall sentiment of the users towards the products. The distribution of ratings can help identify the most popular products and the products that need improvement. The following histogram shows the distribution of ratings in the dataset.

The ratings distribution shows that the majority of the reviews have high ratings, with a peak at 5.0. This indicates that users generally have positive sentiments towards the products they review. However, there are also reviews with lower ratings, indicating that some products may need improvement.

3.4.2 Most Reviewed Products

Identifying the most reviewed products can help understand the popularity and demand for different products in the dataset. The following table shows the top 10 most reviewed products based on the number of reviews.

The most reviewed products are smart speakers from the Echo Dot series, indicating the popularity of smart home devices among users. The number of reviews for each product provides insights into the demand and user engagement with these products.

3.4.3 Most Active Users

Identifying the most active users can help understand the user engagement and contribution to the dataset. The following table shows the top 10 most active users based on the number of reviews posted.

The most active users have contributed a significant number of reviews to the dataset, indicating their engagement and participation in reviewing products. These users play a crucial role in providing feedback and insights on the products, which can help other users make informed decisions.

4 Conclusion

This document provides a comprehensive overview of the literature review and exploratory data analysis conducted for the first deliverable of the Major Research Project. The literature review covers recent advancements in knowledge graph construction, embedding, and recommendation systems, highlighting the integration of large language models with knowledge graphs to enhance the quality and efficiency of knowledge representation and recommendation. The exploratory data analysis provides insights into the Amazon Reviews'23 dataset, including the distribution of ratings, the most reviewed products, and the most active users. The analysis aims to provide a deeper understanding of the dataset and inform the subsequent steps in the research project.

The literature review and exploratory data analysis lay the foundation for the research project, setting the stage for further investigation into the use of large language models to enhance knowledge graph construction and recommendation systems. The insights gained from the analysis will guide the development of novel methods and techniques to improve the performance and interpretability of recommendation systems, paving the way for more effective and scalable knowledge-based recommendations.

5 Future Work

The next steps in the research project will focus on developing a novel method to integrate large language models with knowledge graphs for recommendation systems. The research will explore the use of large language models for knowledge graph embedding, completion, and construction to enhance the quality and efficiency of knowledge representation. The goal is to develop an end-to-end approach that leverages the power of large language models and graph neural networks to improve the performance and interpretability of recommendation systems.

6 References

[1] Yunfan Gao et al. *Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System*. 2023. arXiv: 2303.14524.

[2] Qingyu Guo et al. *A Survey on Knowledge Graph-Based Recommender Systems*. https://arxiv.org/abs/2003.00911. 2020. arXiv: 2003.00911 [cs.IR].

[3] Xiangnan He et al. *LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation*. https://arxiv.org/abs/2002.02126. 2020. arXiv: 2002.02126 [cs.IR].

[4] Shaoxiong Ji et al. *A Survey on Knowledge Graphs: Representation, Acquisition, and Applications*. IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 494-514. 2022. DOI: 10.1109/TNNLS.2021.3070843.

[5] Md. Rezaul Karim Ullah et al. *LLM-KGC: Long Text Generation for Knowledge Graph Completion*. https://arxiv.org/abs/2106.01866. 2021. arXiv: 2106.01866 [cs.CL].

[6] Xiang Wang et al. *KGAT: Knowledge Graph Attention Network for Recommendation*. https://arxiv.org/abs/1905.07854. 2019. arXiv: 1905.07854 [cs.LG].

[7] Yunpu Xu et al. *Text2KG: Generating Knowledge Graphs from Text via Tree Parsing.* https://arxiv.org/abs/2106.01866. 2021. arXiv: 2106.01866 [cs.CL].

[8] Weijie Zhang et al. *KG-BERT: BERT for Knowledge Graph Completion.* https://arxiv.org/abs/2004.12092. 2021. arXiv: 2004.12092 [cs.CL].

[9] Yunpu Zhang et al. *KGET: Knowledge Graph Embedding Transformer.* https://arxiv.org/abs/2106.03218. 2021. arXiv: 2106.03218 [cs.IR].