International Journal of Advanced Smart Convergence Vol.12 No.4 202-207 (2023)
http://dx.doi.org/10.7236/IJASC.2023.12.4.202

IJASC 23-4-25

Design and Implementation of AI Recommendation Platform for Commercial Services

Jong-Eon Lee

Professional, Enterprise Division, LG UPLUS
jongeonlee@lguplus.co.kr;medosecy@gmail.com

Abstract

In this paper, we discuss the design and implementation of a recommendation platform actually built in the field. We survey deep learning-based recommendation models that are effective in reflecting individual user characteristics. The recently proposed RNN-based sequential recommendation models reflect individual user characteristics well. The recommendation platform we proposed has an architecture that can collect, store, and process big data from a company's commercial services. Our recommendation platform provides service providers with intuitive tools to evaluate and apply timely optimized recommendation models. In the model evaluation we performed, RNN-based sequential recommendation models showed high scores.

Keywords: *Recommendation Platform, DNN, RNN, GRU4Rec BERT4Rec, GPT4Rec*

1. Introduction

As AI technology has recently developed, companies are making efforts to improve customer satisfaction with services by providing personalized recommendation lists based on customer data. Before the era of AI technology, service providers provided recommendation services through technologies such as popularity models and machine learning-based collaboration models [1]. In this case, the items recommended to the user are provided from the same recommendation list created by the service provider, rather than a recommendation list that reflects the individual's characteristics.

Recently, the recommendation algorithm is applying personal characteristic data based on customer usage history to deep learning to recommend personally optimized content, and each business provider uses an AI deep learning model that can reflect the characteristics of the service and individual preferences. In this paper, we aim to design and implement a commercial service platform based on AI deep learning technology that can easily learn with the latest recommendation model based on existing customer big data. The structure of the paper is to examine the characteristics of representative recommendation models in Section 2, and to design the recommendation model learning procedure and system architecture in Section 3. Section 4 introduces a case of evaluating the model and applying the model using a customer dataset in an actually implemented system.

Manuscript Received: october. 21, 2023 / Revised: october. 26, 2023 / Accepted: November. 2, 2023
Corresponding Author: jongeonlee@lguplus.co.kr;medosecy@gmail.com
Tel: +82-10-8719-1099
Professional, Enterprise Division, LG UPLUS, Korea

2. Recommendation Models

In the case of a recommendation model, when the amount of data is small, commonly known collaborative filters or Item2Vec are used, and when data accumulates, a deep learning-based model is used, which is advantageous for personalization [2].

As deep learning language models have gradually developed, sequential recommendation models have also developed along with them. LSTM emerged to solve the popularity of RNN and its essential problem, the long-term dependency problem, and sequential recommendation models such as GRU4Rec and BERT4Rec have also been developed. Additionally, GPT4Rec which introduces a generative language model for recommendation, is also worth noting [3].

- **GRU4Rec.** It is one of the variants of LSTM, and is a model with fewer parameters and fewer calculations because there is no separate Cell State and Output Gate. It consists of two gates: Reset Gate and Update Gate. It is a much lighter model with no clear performance difference from LSTM. It is an RNN series and has good performance in continuous data processing. This model predicts the next item based on the user's session data [4].
- **Bert4Rec.** In the case of GRU4Rec, an RNN-based sequential recommendation model, unidirectional recommendation models that only considered the user's previous behavior patterns were mainly used. However, in the case of this one-way recommendation model, performance may be limited because the model is learned only with information about items the user has purchased in the past. BERT4Rec started with the idea of understanding the user's sequence data by learning context from both sides through a two-way mechanism. The BERT4Rec model consists of Embedding Layer, Transformer Layer, and Output Layer as shown in Figure 1. In order to learn the model bidirectionally, like BERT's learning method, a masked token for the user's action sequence is used to identify information from before and after information [5].
- **GPT4Rec.** GPT4Rec is a generative framework for interpreting user preferences and providing personalized recommendations. GPT4Rec first generates a virtual search query through a generative model. The generative model includes items viewed by the user as input, and items for recommendation are obtained through queries generated using a search engine. GPT4Rec is said to create a user interests representation with diversity and granularity through this process and improve the diversity and relevance of recommendation results [6].

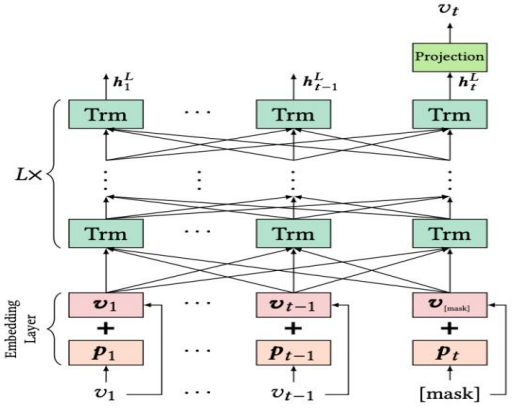

Figure 1. BERT4Rec Model Architecture

3. Design

The core component required for the system is to implement a series of procedures related to the creation and application of recommendation models as service components. In this Section, we introduce the process of learning recommendation models and design a recommendation platform architecture that can accommodate big data.

3.1 Recommendation Model Learning Procedure

Learning of the recommendation model can be used in the recommendation model based on the usage history and customer information data of each service customer. Recommendation model learning analyzes data features, learns by reflecting the characteristic data in the model, and provides optimized results by tuning the hyper-parameters of the model based on the verification results of the learning model. Figure 2 shows the procedure for learning recommendation models. The creation of learning models is divided into two steps: data feature analysis and model training and tuning. The data feature analysis stage consists of raw data extraction, EDA (Exploratory Data Analysis), and feature selection. In the model training and tuning stage, learning and tuning are repeated to verify and deploy the optimal model

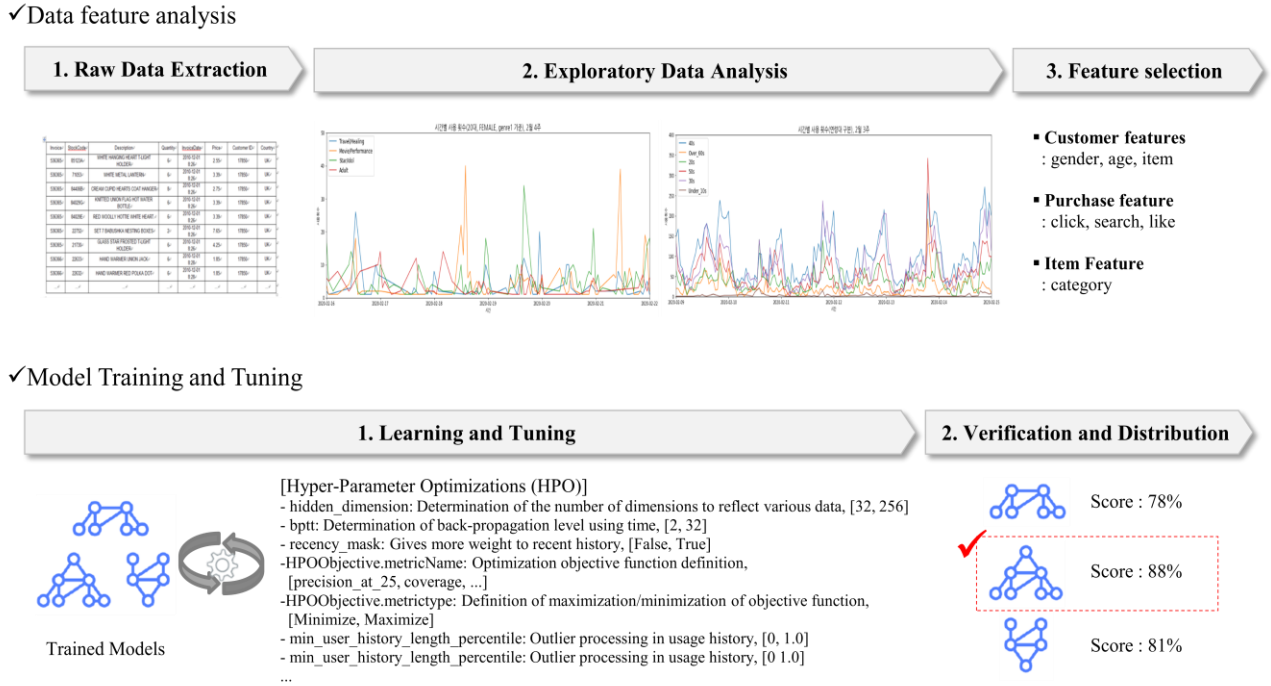

Figure 2. Recommendation Model Learning Procedure

3.2 AI Recommendation Platform Architecture

The architecture of this system was designed to support various commercial services in the real world. The platform configuration can provide recommendation models to each commercial service through the legacy data area, collection area, storage/processing, and service provider IF. Each area has a structure that allows

services, models, and I/F to be flexibly expanded. Figure 3 shows the architecture of the AI recommendation platform.

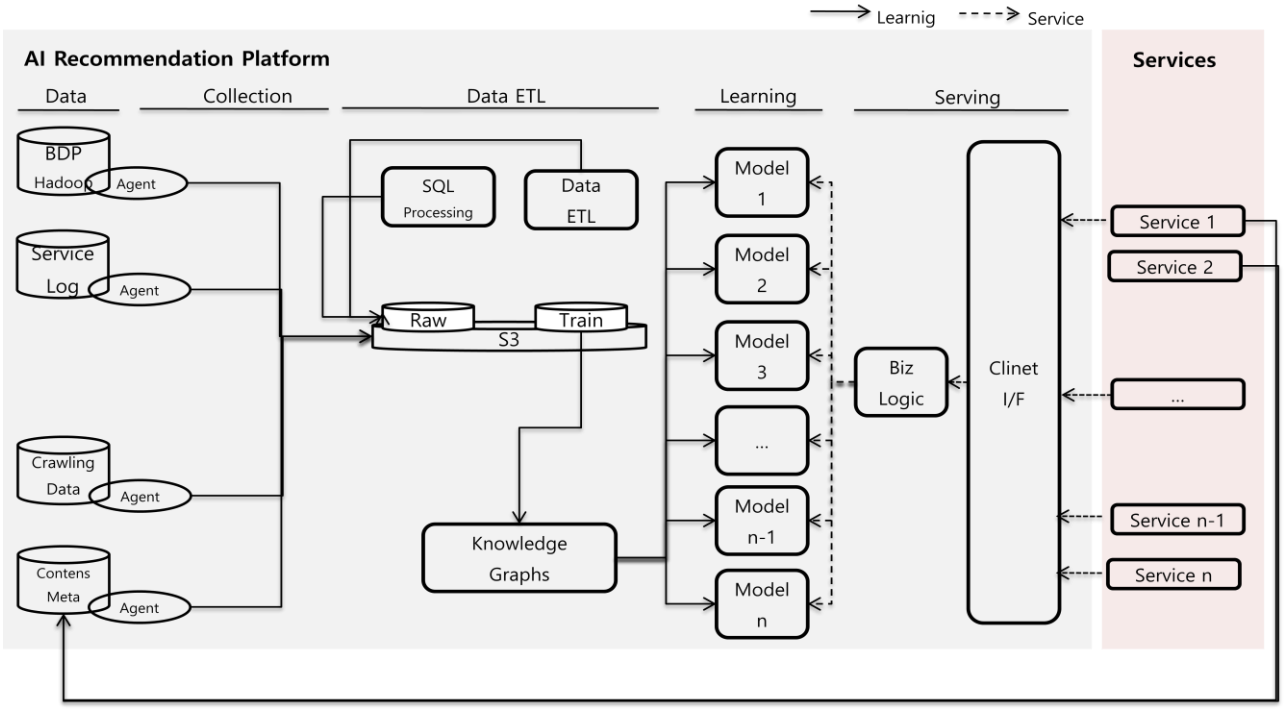

Figure 3. AI Recommendation Platform Architecture

4. Implementation : Evaluate and Deploy Recommendation Models

The system we propose is a dynamic system that can evaluate models according to changes such as data accumulation and apply the most advantageous model. It provides users with tools to evaluate and apply models optimized for service situations. For example, at the beginning of the service, a well-known model is adopted and operated, but when data accumulates and a personalized recommendation model becomes available, the model is evaluated based on user data and the results are reported. Service providers can select and apply the optimal model based on the results.

Table 1 and Figure 4 show the results of extracting, processing, and evaluating the company's actual commercial service purchase list data for three months. Four evaluation indicators were applied: Precision, Recall, Coverage, and F1 Score, and Bert4Rec showed the best results.

Table 1. Recommendation Model Evaluation Results

Model	Precision	Recall	Coverage	F1 Score
GUR4Rec	0.13	0.30	0.11	0.18
Bert4Rec	0.27	0.60	0.09	0.36
GPT4Rec	0.21	0.40	0.07	0.26

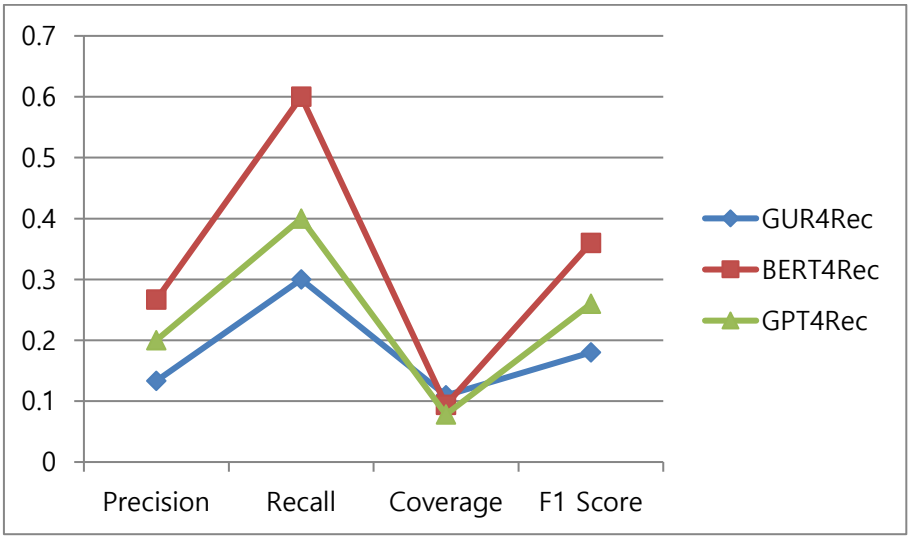

Figure 4. Comparison Graph of Recommendation Model Evaluation Results

Service providers can apply the optimal model with the evaluation results. Figure 5 is an example of a web service that service providers can provide by selecting an AI recommendation model.

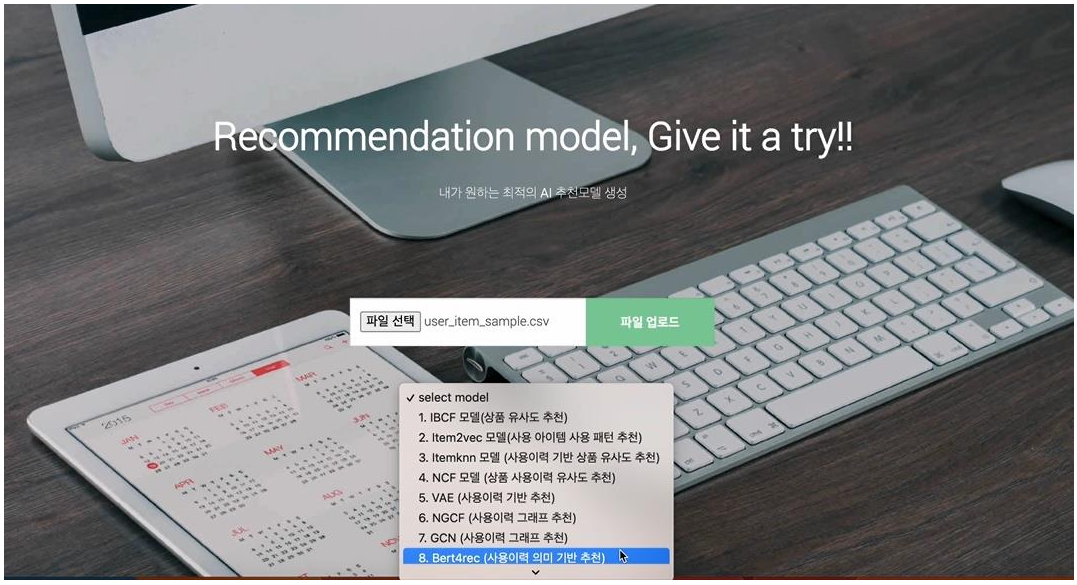

Figure 5. Service Provider's Recommendation Model Selection Interface

4. Conclusion

To apply the optimal commercial service recommendation model, we designed and implemented a component that can generate and evaluate a recommendation model. The AI recommendation platform has a legacy data area, collection area, storage/processing, and service provider I/F area, and each area has a structure that allows for easy and flexible addition of services, models, and I/F. Additionally, this system can provide service providers with the optimal model according to the service point through intuitive UX.

References

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based collaborative filtering recommendation algorithms," Proceedings of the 10th international conference on World Wide Web, May 2001. https://dl.acm.org/doi/10.1145/371920.372071

[2] Oren Barkan and Noam Koenigstein, "Item2Vec: Neural Item Embedding for Collaborative Filtering," 2016 IEEE Workshop on Machine Learning for Signal Processing, November 2016. https://arxiv.org/abs/1603.04259

[3] H Sak, AW Senior, and F Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," INTERSPEECH (2014), Feb 2014. https://static.googleusercontent.com/media/research.google.com/ko//pubs/archive/43905.pdf

[4] Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk,, "SESSION-BASED RECOMMENDATIONS WITH RECURRENT NEURAL NETWORKS," ICLR 2016, March 2016.B. Sklar, Digital Communications, Prentice Hall, pp. 187, 1998. https://arxiv.org/abs/1511.06939

[5] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang, "BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer," Proceedings of the 28th ACM International Conference on Information and Knowledge Management", November 2019 https://arxiv.org/abs/1904.06690

[6] Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni, "GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Interpretation," April 2023 https://arxiv.org/abs/2304.03879

[7] DE Rumelhart, GE Hinton, and RJ Williams, "Learning representations by back-propagating errors," Nature, October 1986. https://www.nature.com/articles/323533a0

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," NIPS, September 2014. https://arxiv.org/abs/1409.3215

[9] Heiga Zen and Hasim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," ICASSP, April 2015. https://ieeexplore.ieee.org/abstract/document/7178816

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, October 2019. https://arxiv.org/abs/1810.04805