ISSN:2790-1688 Volume-13-(2025)

Architecture Design of Recommendation System Based on Large Multimodal Agents

Yang Zhou [1, a], Youwei Li [1, b], Yifan Zhou [1, c], Chaoying Zheng [1, d] and Yangyang Li [1, e]

[1]Academy of Cyber, Beijing, China.

[a] zhyang_626@163.com, [b] liyouwei@1010club.cn, [c] yifanzhou1218@yeah.net,

[d] zhengchaoying0228@foxmail.com, [e] liyangyang@live.com

Abstract. Exploiting the large language models, autonomous agents have been widely developed for tackling decision-making and complex tasks in interactive environments. With the development of LLMs, multimodal agents have emerged with the ability of understanding not only text, but also images, sounds and videos. Therefore, multimodal agents are naturally applicable to generating recommendation information. However, few solutions focus on solving the intelligent recommendation problems on mobile phones, particularly in actively collecting and analyzing users preferences. In this paper, we propose an architecture for autonomous recommendation agents on cellphones, which can actively collect and understand multimodal data without restrictions on application layouts and graphical interface. We also introduce detail methods and workflows of the proposed agent. To this end, we capture a complete view of the proposed system.

Keywords: Large language model; Multimodal agents; Mobile agents; Recommendation system.

1. Introduction

The large language model (LLM)[1] represents a promising technique for understanding and generating text, particularly in the domains of long text understanding and summary generation. Through large-scale data pre-training, an LLM can output accurate content by receiving a simple instruction, known as a prompt. As LLMs are dedicated to processing text, they are inherently limited in their ability to comprehending pictures, voices, or even videos. However, the emergence of multimodal large language models[2] has effectively addressed this limitation. Along with the development of multimodal large models, agents have garnered extensive attention from both industry and academia. An agent is capable of perceiving the environment and taking corresponding actions in order to achieve a specific goal[3]. Driven by multimodal LLMs, multimodal agents are gradually becoming new productive forces in various industries, especially in fields such as GUI automation[4] and role-playing[5]. Similar to multimodal language models, multimodal agents are capable of handling multiple types of data. Nevertheless, current agents typically focus on automatic operations on mobile phones to execute actions, lacking in the monitoring and analysis of user behaviors.

Recommendation systems[6] are able to recommend items to users according to their preferences and behaviors. However, most of these systems lack interactivity and feedback mechanisms, and it is difficult to achieve cold start and cross-domain recommendation. The LLM-based recommendation system can address the above issues, but its automatic intelligent interaction ability remains insufficient.Agent-based recommendation systems usually focus on conversational recommendation and simulation recommendations, without considering information across applications.

In this paper, we propose a detailed construction method applicable to building an autonomous mobile recommendation system based on multimodal agents. By leveraging the advantage of agents, it is inherently more straightforward to solve the problem of automating the execution of actions. The key idea is that we introduce the architecture of our proposed recommendation system, which is designed to automatically analyze user preferences. Another important aspect is that we propose using agents to achieve automatic operating as recommended. Additionally, we provide a detailed

view of the implementation procedures for our proposed system. The main contributions of our paper are presented as follows.

1) we design a recommendation system that is applicable to all mobile phones and has no restrictions on applications or cellphone layouts.

2) we employ a multimodal agent that is capable of analyzing user preferences by combining gestures and browsing contents.

3) by utilizing agents, our system possesses the ability of self-memory and learning.

The rest of the paper is organized as follows. In section Ⅱ, we introduce related works. Section Ⅲ provides an overall view of our proposed system architecture. In section Ⅳ, we illustrate the concrete methodology of implementation. We summarize the entire paper in section Ⅴ.

2. Related Work

Recommender systems have been widely incorporated with LLMs to achieve more accurate recommendations in recent years. Some research focuses on how to improve the results by designing different prompting strategies[7]. Others utilize LLMs to rank and generate recommendations considering users preferences[8], such as clicks, purchases, ratings and other historical behaviors.

Most recently, agents designed for recommendation systems incorporated with LLMs have been widely researched. Some of them commit to simulate user behaviors and items by agents. Agent4Rec[9] is an LLM-based users preferences and behaviors simulator in the movie recommendation scenario. AgentCF[10] simulates both users and items as agents, and utilizes a collaborative learning method to achieve multiple agents optimization. ToolRec[11] is a framework that utilizes LLMs to emulate a real user and his/her decision-making process. It selects appropriate attribute-oriented tools via tool learning according to the simulated user's preferences. Other researches aim to optimize the planning and memory modules, or workflows of agents to achieve accurate recommendations. A paradigm Rec4Agentverse is designed in [12] to depict three roles in personalized recommendation, users, agent items and agent recommenders. RecMind[13] aims at improving the planning ability by designing a self-inspiring algorithm. By utilizing LLMs as brains, InteRecAgent[14] proposes an interactive recommendation system incorporating three key components, memory, task planning and reflection. More efforts have been spent demonstrating how to collaboratively work for multiple agents. MACRec[15] tackled a specific recommendation task with multiple agents working together which have different abilities. [16] proposed a multimodal system consists of three agents, product recommendation agent, image analysis agent and market analysis agent. In this work, the multimodal data used to analyze user preferences is actively provided by the user. However, these schemes have not taken passively monitoring to cross-application data to enable user preference analysis on mobile phones into account.

3. System Architecture

In this section, we describe the architecture of our proposed system, which is divided into 4 layers shown as figure 1. The detailed functions of each layer is described as follows.

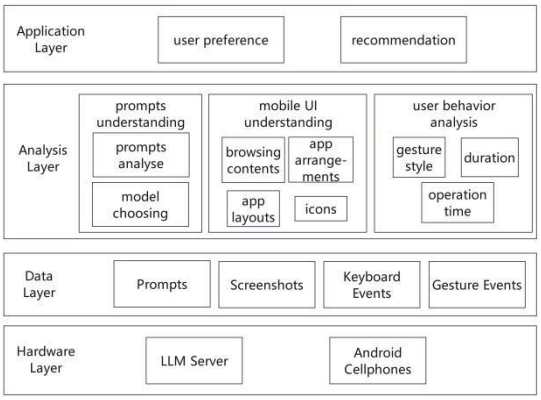

Fig. 1 An Illustration on System Architecture

Hardware Layer. This layer is comprised of two types of hardware: LLM servers and Android cellphones. The LLM servers are utilized to deploy LLMs. These LLMs have the capability to process and analyze multimodal data, including texts, screenshots and so on. Android cellphones serve as operating terminals for users and agents.

Data Layer. The data layer collects various kinds of multimodal information from the user's mobile phone, including input commands, screenshots, keyboard input, and gesture data. To enhance the precision of fusion analysis, the data layer now incorporates preliminary processing capabilities for multimodal data. For screenshots, we utilize lightweight image encoding models (such as MobileNet[17]) to extract fundamental features. For keyboard inputs and gesture information, we employ embedding representation and temporal feature extraction methods to uniformly represent user behaviors. At the data layer, each type of data undergoes feature extraction in its corresponding feature space before being transferred to the analysis layer for further fusion.

Analysis Layer. In this layer, we introduce an attention mechanism to enable effective fusion of multimodal information. The features originating from different modalities (for instance, text and images) are mapped into a unified feature space. We use a Transformer-based multimodal fusion model to integrate these features and capture the relationships between modalities. Specifically, we use graph convolutional networks (GCNs) to extract spatial features from gestures and screenshots. An attention mechanism is utilized to accomplish information transfer and integration among different modalities, ensuring temporal alignment and semantic coherence of the data.

Application Layer. In this layer, different algorithms are deployed to address various problems. By deploying user preference analysis algorithms, it becomes possible to analyze multi-dimensional information such as users' operating habits and browsing habits. Based on this, recommendation algorithms can be deployed to achieve content and behavior recommendations for users.

4. Methodology

In this section, we provide a detailed illustration of the principles and implementation processes of the proposed agent.

4.1 Agent architecture

Firstly, we present a comprehensive view of the architecture of our proposed agent, as shown in figure 2.

The proposed agent framework is mainly composed of four modules.

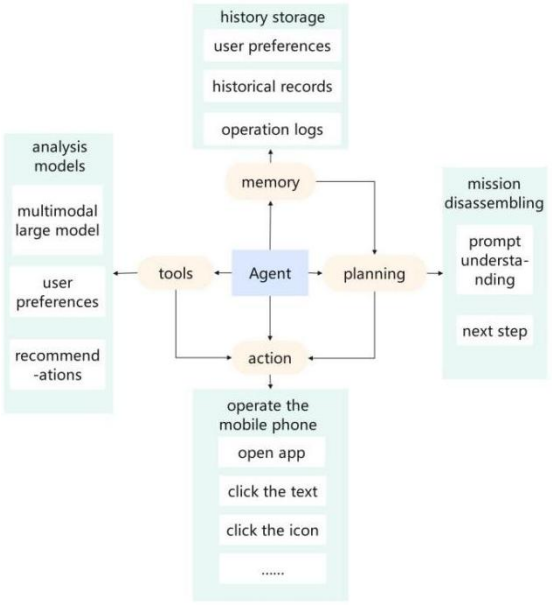

Fig. 2 Architecture of the Proposed Agent

Planning. Understanding problems and reliably finding solutions are important for our agent. It responds to user requests by decomposing the problems into necessary steps or subtasks.

Memory. This module is responsible for storing user preferences, historical records, operation logs, and so on. It is extremely important for the learning and decision-making processes of the agent. We employ the method of embedding vectors to implement the query of memory content, thereby improving the accuracy and efficiency of the query.

Tools. This module incorporates a combination of lightweight and deep learning models to deal with multimodal data. Firstly, we use MobileNet to process phone screenshots, leading to a more compact feature representation that is suitable for subsequent multimodal analysis, as shown in Fig.3. For text and gesture data, we employ a variant of BERT and LSTM networks for text embedding and temporal modeling respectively. Then, these features from different modalities are integrated into a multimodal Transformer, which can capture relationships among modalities. For instance, user gestures during browsing news(such as swipe direction and frequency) can be input into the Transformer together with browsing content (screenshots and text). Through multiple layers of attention modules, the system can better understand the subtle differences in user preferences. The detailed steps are as follows:

1) Feature Extraction: Utilize MobileNet to extract visual features from screenshots. Employ variants of BERT for text input embedding, and use LSTM to extract temporal features from gesture behavior.

2) Feature Fusion: Input these extracted multimodal features into a Transformer-based multimodal fusion model to capture relationships among modalities, ensuring the fused representation reflects both visual, interactive, and content preferences of users.

3) Preference Modeling: Use GCNs to further process these fused features and build a more comprehensive and consistent user preference model.

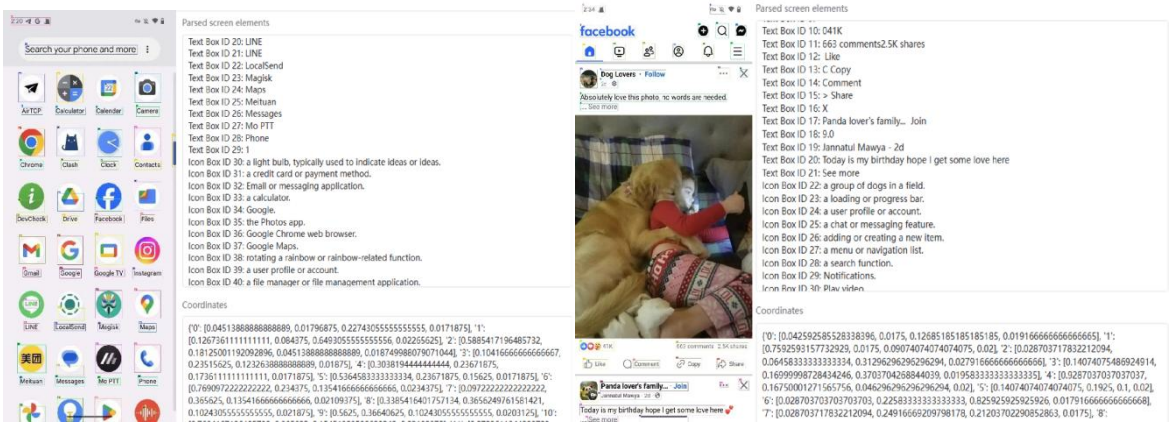

Fig. 3 Illustrations on Screenshots Understanding

Actions. The agent executes specific operation actions on mobile phones by invoking tools according to its planning and memory, for example, opening some app, clicking on text and sending, browsing news, and so on. To achieve recommendation considering user preferences, we integrate fine-tuned multimodal large language models into our system. Different from the PALR method[8] shown in Fig.4, we generate the natural language user profile by taking into account not only user-item interactions but also the user's browsing contents and gesture features. A retrieval model is employed to analyze recommendation candidates. After being fed with a natural language prompt, the large language model for recommendation generates the recommendation results for the user. Given this recommendation information, the agent on the mobile side automatically executes the corresponding operations.

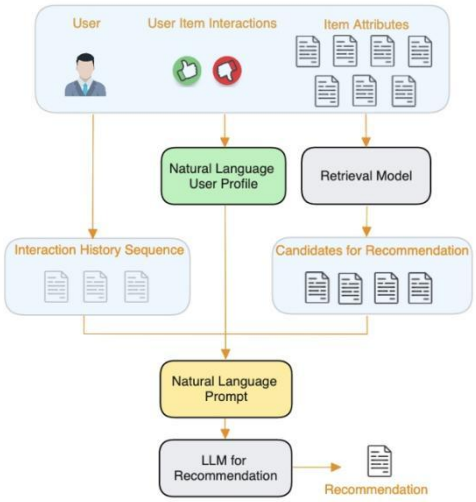

Fig. 4 The Recommendation Model[8]

4.2 Example of workflows

Suppose a user browses products in a shopping app and checks user reviews of similar products on social media.

The operational workflow of our system would be as follows:

1) Data Collection: The system gathers screenshots from the shopping app and observes user inputs and gestures (e.g., liking, swiping) on social media.

2) Feature Extraction and Fusion: The system extracts features from the screenshots by using MobileNet, processes text input with BERT, and analyzes gesture behavior through LSTM. Then, all these features are fused using the Transformer model.

3) Preference Analysis: The system analyzes the fused features to determine the user's growing interest in a particular type of product. A GCN is employed for further correlation analysis to identify potential product preferences.

4) Recommendation and Execution: Generate a list of recommended products and automatically open the pages of relevant products in the shopping app. The user is given the option to proceed with a purchase.

5. Conclusions

In this paper, we propose an architecture for a recommendation system based on large multimodal agents. We provide a detailed description of our proposed system and the agent. Unlike existing studies on analyzing users' preferences, we take into consideration multimodal information of users, including gestures, inputs, browsing contents, and so on. Our analyzing method is applicable to any application installed on the mobile phone. Additionally, our system can automatically operate the mobile phone according to the recommendations.

References

[1] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, L. Hendricks, J. Welbl, A. Clark et al., "Training compute-optimal large language models. arxiv 2022," arXiv preprint arXiv:2203.15556, vol. 10, 2022.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[3] M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," The knowledge engineering review, vol. 10, no. 2, pp. 115–152, 1995.

[4] J. Wang, H. Xu, J. Ye, M. Yan, W. Shen, J. Zhang, F. Huang, and J. Sang, "Mobile-agent: Autonomous multi-modal mobile device agent with visual perception," arXiv preprint arXiv:2401.16158, 2024.

[5] G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for" mind" exploration of large language model society," Advances in Neural Information Processing Systems, vol. 36, pp. 51 991–52 008, 2023.

[6] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," Egyptian informatics journal, vol. 16, no. 3, pp. 261–273, 2015.

[7] H. Lyu, S. Jiang, H. Zeng, Y. Xia, Q. Wang, S. Zhang, R. Chen, C. Leung, J. Tang, and J. Luo, "Llm-rec: Personalized recommendation via prompting large language models," arXiv preprint arXiv:2307.15780, 2023.

[8] F. Yang, Z. Chen, Z. Jiang, E. Cho, X. Huang, and Y. Lu, "Palr: Personalization aware llms for recommendation," arXiv preprint arXiv:2305.07622, 2023.

[9] A. Zhang, Y. Chen, L. Sheng, X. Wang, and T.-S. Chua, "On generative agents in recommendation," in Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval, 2024, pp. 1807–1817.

[10] J. Zhang, Y. Hou, R. Xie, W. Sun, J. McAuley, W. X. Zhao, L. Lin, and J.-R. Wen, "Agentcf: Collaborative learning with autonomous language agents for recommender systems," in Proceedings of the ACM on Web Conference 2024, 2024, pp. 3679–3689.

[11] Y. Zhao, J. Wu, X. Wang, W. Tang, D. Wang, and M. de Rijke, "Let me do it for you: Towards llm empowered recommendation via tool learning," in Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 1796–1806.

[12] J. Zhang, K. Bao, W. Wang, Y. Zhang, W. Shi, W. Xu, F. Feng, and T.-S. Chua, "Prospect personalized recommendation on large language model-based agent platform," arXiv preprint arXiv:2402.18240, 2024.

[13] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, X. Huang, Y. Lu, and Y. Yang, "Recmind: Large language model powered agent for recommendation," arXiv preprint arXiv:2308.14296, 2023.

[14] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian, and X. Xie, “Recommender ai agent: Integrating large language models for interactive recommendations,” arXiv preprint arXiv:2308.16505, 2023.

[15] Z. Wang, Y. Yu, W. Zheng, W. Ma, and M. Zhang, “Macrec: A multi-agent collaboration framework for recommendation,” in Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2760–2764.

[16] P. Thakkar and A. Yadav, “Personalized recommendation systems using multimodal, autonomous, multi agent systems,” arXiv preprint arXiv:2410.19855, 2024.

[17] B. Koonce and B. Koonce, “Mobilenetv3,” Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization, pp. 125–144, 2021.