Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241620

Evaluation and optimization of intelligent recommendation system performance with cloud resource automation compatibility

Kangming Xu[1a,*], Haotian Zheng[1b], Xiaoan Zhan[2], Shuwen Zhou[3], Kaiyi Niu[4]

[1a]Computer Science and Engineering, Santa Clara University, CA, USA
[1b]Electrical & Computer Engineering, New York University, New York, NY, USA
[2]Electrical Engineering, New York University, NY, USA
[3]Computer Science, The University of New South Wales, Sydney, Australia
[4]Artificial intelligence, Royal Holloway University of London, Egham, UK

[*]kangmingxu87@gmail.com

Abstract. This paper comprehensively explores the integration of cloud computing and advanced recommendation systems, emphasizing their pivotal roles in enhancing user experiences and operational efficiencies across digital platforms. It reviews the evolution of recommendation algorithms, highlighting their application in diverse domains such as e-commerce and media. The study evaluates the performance of advanced models like UniLLMRec against traditional counterparts using datasets from news and e-commerce domains. Additionally, the paper discusses the infrastructure architecture of cloud computing, demonstrating its capability to support scalable and efficient data processing. Through experimental insights and methodology, the research underscores the transformative impact of cloud technologies on optimizing recommendation system performance, thereby advancing digital engagement and competitiveness.

Keywords: Cloud Computing, Recommendation Systems, Artificial Intelligence, Big Data.

1. Introduction

With the rapid advancement of digital transformation and continuous innovation in cloud technology, cloud computing has become an indispensable infrastructure across business, government, and personal services. The cloud computing market is segmented into Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service [1] (SaaS), each playing a distinct role and driving growth. IaaS, as the foundation, offers virtualized computing resources, storage, and network services, providing enterprises with a flexible and scalable platform for application deployment. Cloud platforms facilitate centralized data management and analysis, enabling precise user behavior analysis, real-time recommendation strategy adjustments, and enhancing system intelligence and user satisfaction.

This paper examines how cloud computing can enhance the performance and effectiveness of recommendation systems, thus improving user experiences and enterprise competitiveness.

DOI: 10.54254/2755-2721/87/20241620

2. Related work

2.1. Intelligent recommendation system

A recommendation system is an information filtering system that predicts a user's behavior or preference for an item. [2] Why do we need a recommendation system? Generally speaking, the recommendation system has much data information, far greater than the user's demand. The recommendation system is produced to enable the user to quickly find the information that meets the user's needs from the massive data. Therefore, the recommendation system is used for enormous data information overload, and the amount of data is too small; it is not worth using the recommendation system.

In the early years, users of online shopping platforms needed to go step by step, according to the classification of goods or keyword search, to find their products in the mass of goods. In recent years, on the Double Eleven, ordinary consumers can quickly screen out the goods they want and receive recommendations for goods and live broadcasts that align with their preferences. [3] Later, the rise of AI technology, especially the information flow content recommendation and short video recommendation typical of Toutiao and Douyin, once again helped the recommendation system to improve significantly. Whether it was the product recommendation of shopping platforms, the anchor recommendation of live broadcasting platforms, or the video content recommendation of video platforms, more and more people began to sign that [4]AI knows itself better.AI recommendation systems have also become necessary for Internet companies' business. Data show that on some of the world's large online websites, even if the relevance of the recommended content is only increased by 1%, its sales will increase by billions; the AI recommendation system is undoubtedly a high-value system hidden behind many Internet applications.

2.2. Recommendation system algorithm

The core of the recommendation system is the recommendation algorithm. The recommendation algorithm can be simplified into a black box, and the input of the black box is various attributes and characteristics of the user and the candidate, including the user's age, gender, purchasing power, and the candidate's content, category, and release time. [5] The output of the black box is a list of recommendations for the user, ranked by preference.

Currently, the main recommendation algorithms include popularity-based algorithms, collaborative filtering algorithms, content-based and model-based algorithms, etc. Different recommendation algorithms have different preferences, advantages, and disadvantages. Still, they are all based on extensive data analysis to predict and recommend users and generate lists of items they may be interested in.

Based on the popularity of the algorithm, the simple version of the implementation can be sorted by the heat, such as the user's likes, comments, and forwarding amount to calculate the heat, and then according to the heat value to recommend sorting, this algorithm is relatively simple, the disadvantage is that the heat needs to be constantly optimized and improved, the need to integrate various factors continually changing, to have a good performance. Cloud computing infrastructure architecture

Cloud computing has become an infrastructure for several reasons:

1. Public cloud accelerates the integrated development of hardware and software and truly promotes the process of T service

Software and hardware integration is one of the development trends of T., the public cloud can be used as the "glue" of software and hardware; through the public cloud, the integration and integration between software and hardware becomes easy public cloud is responsible for managing all hardware resources, the software can be through the interface with the cloud, it can achieve the goal of software and hardware integration.

The industry has recognized the trend of IT as a service for many years; Saas (software as a service), PaaS(platform as a service), PaaS (infrastructure as a service), everything is a service. As software and hardware integration continues to deepen[6], T will be presented to users as services without

DOI: 10.54254/2755-2721/87/20241620

distinguishing between software and hardware services. No matter what kind of service, you need a carrier, and the public cloud is that carrier.

2. T technology rooted in the public cloud, its application and innovation must rely on the public cloud

Big data, artificial intelligence, and other new technology developments have been rooted in the public cloud; strictly speaking, if you leave the public cloud, there is no significant data and artificial intelligence. Artificial intelligence is based on extensive data development; without the network and data, artificial intelligence will no longer exist. Therefore, artificial intelligence development must also rely on the public cloud [7].

3. Public cloud is the platform of "convergence" and the interface of all "connectivity."

In the era of digital economy, integration is an inevitable trend; hardware and software should be integrated, T products and services should be integrated, industrialization and information technology should be integrated, all links of the industrial chain should be integrated, and industries should be integrated. [10] The development of all kinds of networks provides a channel for integration.

4. Infrastructure framework

The cloud computing infrastructure architecture takes distributed multi-cloud as the core, builds the "one cloud and multiple computing" converged base, relies on the unified management of heterogeneous resources and the distributed task collaboration framework, builds a new service system with AI running through it, supports the integrated carrying capacity of general computing, intelligent computing, supercomputing, and network convergence services, and ensures the availability of full-link services. The hierarchical system of traditional cloud architecture is retained in terms of overall architecture.

The cloud network resource construction emphasizes the distributed optimal layout of multiple types of resource pools. [8] Diversity is noted in the software and hardware resource layer, divided into CPU-based general computing infrastructure and intelligent computing infrastructure dominated by AI-accelerated chips such as GPU[9]. The distributed cloud platform manages multi-dimensional heterogeneous resources in a unified manner and implements efficient collaborative task scheduling. Based on infrastructure architecture, cloud service forms show a trend of generalization and intelligent development, carrying multiple business types and providing rich industrial digital capabilities.

In conclusion, the evolution of intelligent recommendation systems has revolutionized user engagement across various digital platforms, driven by sophisticated algorithms like collaborative filtering and content-based recommendations. [10] These systems have become indispensable for personalized user experiences in e-commerce, content streaming, and social media. Meanwhile, cloud computing has emerged as a robust infrastructure, offering scalable resources and efficient data processing capabilities crucial for supporting these advanced systems. As we move forward, the integration of cloud resource automation stands out as a pivotal factor in enhancing the agility and performance of recommendation systems. The next phase of our exploration will evaluate how automated cloud solutions can optimize these systems, ensuring seamless scalability, reliability, and operational efficiency.

3. Methodology

3.1. Experimental design

1. Data set

The experiment used two datasets related to detailed information about news recommendations and film and television reviews, respectively - MIND and Amazon Review. The former contains News articles and user behavior logs from the Microsoft News website; The latter is collected from Amazon's e-commerce platform and includes user reviews, ratings, and product information.

2. Evaluate indicators

Regarding evaluation indicators, we mainly focus on the performance of Recall and Re-ranking tasks. Indicators such as Recall, Normalized Discounted Cumulative Gain (NDCG) [11], and Intra-List Average Distance (ILAD) were used to evaluate the model's performance in the recommendation task.

DOI: 10.54254/2755-2721/87/20241620

3.2. Model comparison

The UniLLMRec framework is also compared to a series of baseline models; These include Popularity-based recommendation (Pop), Factorization Machines (FM), Deep FM, NRMS, SASRec, and LLM-Ranker.

Pop: Rank an item based on its overall user popularity in the user base and recommend the most popular item to the user.

Pros: Simple, easy to implement, and effective for widely popular content.

Disadvantages: Lack of personalization, consideration of user-specific preferences, and inability to satisfy users with specific tastes.

FM: The ability to use factorization parameters to estimate interactions between variables and can handle problems with high sparsity. Because it can combine auxiliary information to overcome the difficulties of cold start and sparse data in a recommendation system, it is a very practical recommendation model

Deep FM: Combines a shallow factorization model with a deep neural network, leveraging both strengths to improve the model's predictive power. This makes it excellent at dealing with complex interactions and sparsity in recommendation systems, especially in scenarios like [12]CTR prediction.

Advantages: The interaction between features can be learned automatically without manually designing the feature interaction. In addition, due to its profound learning nature, Deep FM scales well to large-scale data sets and can adapt to changing data distributions.

NRMS uses a multi-head self-attention mechanism to enhance the performance of the news recommendation system. The model is divided into two main parts: news encoder and user encoder. News encoders use multi-head self-attention to learn the representation of words in news articles. In contrast, user encoders utilize the exact mechanism to capture behavioral patterns and preferences in a user's reading history. In this way, the model can understand and match user interests and relevant news content more accurately.

SASRec is a sequence-based recommendation system that employs a self-attention mechanism to assign weights to past items dynamically in each time step. Its adaptive nature prioritizes long-term dependencies in dense data sets and focuses on recent activity in sparse data sets, contributing to its superior performance.

LLM-Ranker takes advantage of the rich semantic and contextual understanding of LLMS, such as GPT or BERT, to improve ranking tasks in search and recommendation systems, which makes the model more efficient at handling complex user queries and diverse content.

3.3. Performance comparison

From the direct comparison of the performance of UniLLMRec and traditional recommendation model in MIND and Amazon Review data sets and two indicators, respectively, whether GPT-3.5 or GPT-4 is used as a backbone, UniLLMRec can outperform many traditional models (especially the one with GPT-4 as the backbone) when the proportion of training sets is small, which indicates the advantage of UniLLMRec's zero sample learning cost, and also reflects the effectiveness and relevance of its retrieval and recommendation items.

DOI: 10.54254/2755-2721/87/20241620

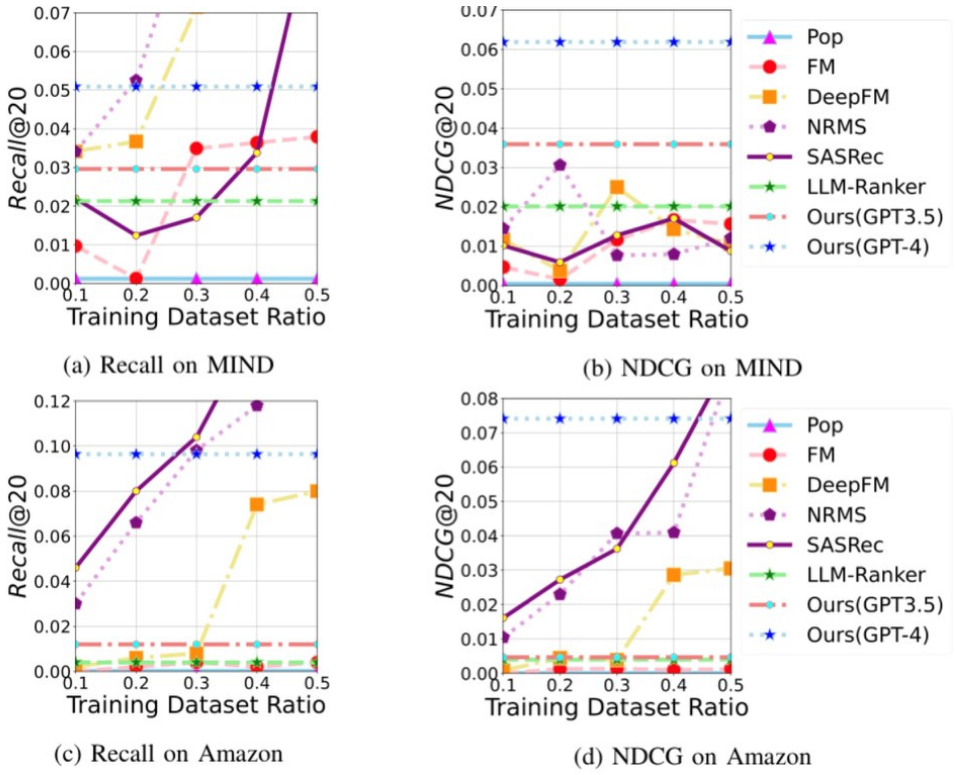

Figure 1. Performance Comparison of Recall and NDCG value on MIND and Amazon datasets.

3.4. Cloud computing deployment

When cloud computing is introduced into the experimental design of a recommendation system, the performance and scalability of the system can be significantly improved. By leveraging the elastic computing resources and efficient storage services provided by cloud computing platforms, we can more efficiently process large-scale data sets, such as MIND and Amazon Review data. During the experiment, the elastic resource characteristics of the cloud computing platform can also automatically adjust the computing resources according to the real-time load, ensuring that the recommendation system can maintain stable performance at peak times. To sum up, the introduction of cloud computing technology can not only optimize the operational efficiency and performance of the recommendation system but also improve the security and reliability of the system, providing a broader and controllable platform for the research and experiment of recommendation algorithms.

3.5. Experimental conclusion

Although UniLLMRec can complete the entire recommendation process in most cases, the researchers also found some problems during the experiment, such as even if the output format is clearly defined in the prompt template, LLM sometimes does not output items according to the instructions, resulting in the items not being correctly indexed (intention recognition problem); In the user interest modeling stage, LLM can capture and summarize user interest, but in the leaf node retrieval and diversity perception rearrangement stage, there is a risk of including examples in the wrong prompt words into the retrieval process (illusion problem).

The researchers made an interesting observation when comparing GPT-3.5 and GPT-4 on the Amazon dataset. When GPT-3.5 and GPT-4 reach the wrong child node, GPT-3.5 will usually continue to complete the subsequent process. At the same time, GPT-4 may proactively give a hint that all candidate answers do not meet the requirements (e.g., "Based on the user's interest in UFC and combat sports," None of the Character & Series subcategories provided are relevant "). This suggests GPT-4 is more accurate than GPT-3.5 in capturing user preferences.

DOI: 10.54254/2755-2721/87/20241620

4. Conclusion

Based on the extensive exploration of cloud computing and advanced recommendation systems in this study, it is evident that integrating cloud infrastructure significantly enhances the performance and scalability of recommendation systems across digital platforms. The deployment of cloud resources revolutionizes the processing of large-scale datasets, particularly in dynamic domains like news and e-commerce, where data volumes are immense and constantly evolving. Cloud infrastructure offers scalable computing power and storage capabilities that traditional on-premises systems struggle to match. This scalability enhances the speed and efficiency of data processing and ensures that recommendation systems can handle peak loads without compromising performance.

In evaluating advanced models such as UniLLMRec against traditional methods, significant advantages emerge regarding recall and recommendation accuracy. UniLLMRec leverages state-of-the-art AI technologies like large language models (LLMs) to analyze user behaviours and preferences more effectively. These models excel in understanding nuanced patterns in user interactions, delivering more personalized recommendations that align closely with individual interests and needs.

Furthermore, the study highlights the critical role of cloud computing as a foundational infrastructure supporting AI-driven recommendation algorithms. By leveraging elastic computing resources and efficient storage services, cloud platforms enable dynamic adjustments to real-time loads, ensuring consistent system performance during peak usage. This research underscores the strategic importance of cloud-enabled recommendation systems in driving digital engagement and competitiveness across various industries, paving the way for future innovations in user-centric technologies.

References

[1] Hoque, M.S., Mukit, M.A. and Bikas, M.A.N., 2012. An implementation of an intrusion detection system using a genetic algorithm. arXiv preprint arXiv:1204.1336.

[2] Bace, R.G. and Mell, P., 2001. Intrusion detection systems.

[3] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J. and Ahmad, F., 2021. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies, 32(1), p.e4150.

[4] Zhan, T., Shi, C., Shi, Y., Li, H., & Lin, Y. (2024). Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3)—arXiv preprint arXiv:2405.09770.

[5] Li, Huixiang, et al. "AI Face Recognition and Processing Technology Based on GPU Computing." Journal of Theory and Practice of Engineering Science 4.05 (2024): 9-16.

[6] Yuan, J., Lin, Y., Shi, Y., Yang, T., & Li, A. (2024). Applications of Artificial Intelligence Generative Adversarial Techniques in the Financial Sector. *Academic Journal of Sociology and Management*, *2*(3), 59-66.

[7] Guo, L., Li, Z., Qian, K., Ding, W., & Chen, Z. (2024). Bank Credit Risk Early Warning Model Based on Machine Learning Decision Trees. *Journal of Economic Theory and Business Management*, *1*(3), 24-30.

[8] Li, Zihan, et al. "Robot Navigation and Map Construction Based on SLAM Technology." (2024).

[9] Fan, C., Ding, W., Qian, K., Tan, H., & Li, Z. (2024). Cueing Flight Object Trajectory and Safety Prediction Based on SLAM Technology. Journal of Theory and Practice of Engineering Science, 4(05), 1-8.

[10] Ding, W., Tan, H., Zhou, H., Li, Z., & Fan, C. Immediate Traffic Flow Monitoring and Management Based on Multimodal Data in Cloud Computing.

[11] Lin, Y., Li, A., Li, H., Shi, Y., & Zhan, X. (2024). GPU-Optimized Image Processing and Generation Based on Deep Learning and Computer Vision. *Journal of Artificial Intelligence General Science (JAIGS) ISSN: 3006-4023*, *5*(1), 39-49.

[12] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421, 9(1), 1.