Product Recommendation System Based on Deep Learning

Ping Lu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 295075022@qq.com

Pingping Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1341369601@qq.com

***Abstract*—With the development of Internet big data and e-commerce, the widespread popularity of information, information acquisition and personalized recommendation technologies have attracted extensive attention. The core value of personalized recommendation is to provide more accurate content and services around users. The recommended scenarios are not uniform, and different dimensions need to be considered. For example, we are facing enterprises or individuals, different age groups, different levels of education, social life and other aspects. In this paper, the classic DNN (Deep Neural Networks) double tower recommendation algorithm in the recommendation algorithm is used as the ranking algorithm of the recommendation system. It is divided into user and item for embedding respectively. The network model is built using tensorflow. The data processed by the initial data through feature engineering is sent into the model for training, and the trained DNN double tower model is obtained. Recall adopts collaborative filtering algorithm, and applies tfidf, w2v, etc. to process feature engineering, so as to better improve the accuracy of the system and balance the EE problem of the recommendation system. The recommendation module of this system is divided into data cleaning as a whole. Feature engineering includes the establishment of user portraits, the analysis of multiple recall and sorting algorithms, the adoption of multiple recall mode, and the implementation of a classic recommendation system with in-depth learning. This makes the recommendation system better balance the interests of both the platform and users, and achieve a win-win situation.**

Keywords-Deep Learning; Recommendation System; Deep Neural Network

I. INTRODUCTION

At present, the Internet is developing rapidly in China. Information plays an increasingly important role in people's daily life. Many enterprises are also involved in the wave of the Internet and build relevant platforms with the help of the Internet, such as shopping, video, news, music, marriage, social networking and other types. However, the highly developed number of goods also brings a huge problem, that is, how to flexibly help users to accurately meet their needs. At the same time, under the influence of education and social environment, everyone has put forward new requirements for self-expression. Self personality has become an important indicator of consumer demand. In the network environment, the variety of complex goods and services and the difference of people's personalized needs together constitute a dazzling choice. Under the combined effect of these multiple factors, the recommendation system also came into being. If you want to overcome the above needs, personalized recommendation system is essential [1].

The product recommendation system based on deep learning has the following advantages and characteristics: efficiency, improve the efficiency of users to obtain the required goods; Transparency, the user's imperceptibility in the recommendation process will not affect the user's use; Accuracy: the recommended results are consistent with the products required by users; Novelty, recommend products in many ways, and break the inherent search deadlock.

At present, domestic and foreign recommendations mainly include three traditional recommendation methods: content based recommendation, collaborative filtering recommendation, and hybrid recommendation, and four common deep learning recommendation

DOI: 10.2478/ijanmc-2023-0041

models: convolutional neural network (CNN), deep neural network (DNN), recurrent neural network (RNN), and graph neural network (GNN).

At present, due to the continuous development of machine learning technology and the deepening of people's understanding of this technology, the recommendation system has made new progress. Take Alibaba and Toutiao today, two companies with significant influence in China, for example. The two key technologies included in their recommendation systems are machine learning and deep learning. These two technologies together constitute their recommendation systems. However, the traditional collaborative filtering and physical recommendation methods, which were more popular in the past, have faded out of people's vision. Such strategies are often difficult to solve the drawbacks of the long tail content's refined personalized recommendation and popularity penetration. At the same time, the benefits of the new recommendation methods are dwarfed by the traditional ones. In 2017, the National University of Singapore proposed NeutralCF, the famous double tower structure, which uses neural networks to optimize the original collaborative filtering algorithm. Youtobe and Google have successively modified the proposed model. The traditional collaborative filtering and LR mode recommendation algorithm are gradually moving towards the double tower model [2].

This paper designs a product recommendation system based on deep learning. The system conducts model training through the product data and user data in the data set. The main contents include: first, feature engineering is carried out, and basic product data, user data, and user behavior data are processed to build user portraits and item portraits; Then, the model is constructed and trained. The DNN double tower model is constructed using tensorflow, and the data set is imported for model training; The next stage is the commodity recall stage, which specifies the quantity of commodities for use in the subsequent sorting stage; The next step is to sort the goods. The recalled goods are sent to the model for sorting to get the final result; Finally, the result is output, and the result is returned through Python construction[3].

II. INTRODUCTION TO MACHINE LEARNING AND DEEP NEURAL NETWORK

A. *Fundamentals of Machine Learning Theory*

The development of artificial intelligence has produced a wide range of branches, and machine learning is one of the important branches that can not be ignored. If we want to realize artificial intelligence, we must not neglect the technology of machine learning. Under machine learning, the algorithm that allows the machine to modify and develop itself is adopted. At the same time, the computer obtains the corresponding rules in continuous data analysis, and the obtained rules can be used to speculate on new sample data. If we use formal language to describe its content. Its definition should be: for a certain type of task T and performance measurement P, if a computer program's performance measured by P on T is self improving with experience E, it can be called learning from experience E.

In machine learning, LR is one of the core algorithms [4]. Logistic Region is also recognized by all types of enterprises because of its simplicity, parallelization, and strong interpretability. The essence of logistic regression is to assume that the data obey this distribution, and then use maximum likelihood estimation to estimate the parameters. Logistic distribution is a continuous probability distribution. Its distribution function and density function are shown in (1) and (2):

$$F\left(x\right) = P\left(X \le x\right) = \frac{1}{1 + e^{-\frac{x-\mu}{\gamma}}} \tag{1}$$

$$f\left(x\right) = F'\left(X \le x\right) = \frac{e^{-\frac{x-\mu}{\gamma}}}{\gamma\left(1 + e^{-\frac{x-\mu}{\gamma}}\right)^2} \tag{2}$$

Logistic regression essentially uses the predictive value of the linear regression model to approximate the logarithmic probability of the real mark of the classification task. Its benefits lie in that: it directly models the probability of classification without realizing the hypothetical data distribution, which prevents the disadvantages caused by inaccurate distribution (different from the generative model); Not only the category can be predicted, but also the probability of the prediction can be obtained[5]. When encountering some kind of tasks that use probability to assist decision-making, it can often play a huge role; The logarithmic probability function is a convex function with arbitrary order derivability, and many kinds of numerical optimization algorithms can calculate the optimal solution. Both gradient descent and Newton iteration can be used to solve the logic regression, and the gradient descent method is briefly introduced:

$$J\left(w\right) = -\frac{1}{n}\left(\sum_{i=1}^{n}\left(\begin{array}{l} y_i \ln p\left(x_i\right) + \\ \left(1 - y_i\right)\ln\left(1 - p\left(x_i\right)\right)\end{array}\right)\right) \quad (3)$$

$$g_i = \frac{\partial J\left(w\right)}{\partial w_i} = \left(p\left(x_i\right) - y_i\right)x_i \quad (4)$$

$$w_i^{k+1} = w_i^k - \alpha\ g_i \quad (5)$$

Gradient descent is to find the descent direction through the first derivative of J (w) to w and update the parameters in an iterative manner, where k is the number of iterations.

B. *Theoretical Basis of Artificial Intelligence*

In 1943, the neuron model was successfully drawn by psychologist Warren McCulloch and mathematician Walter Pits. It can be seen that there are many inputs on the left side of the figure below, similar to the dendrites of neurons [6]. After a nuclear processing, that is, the weighted sum part in the figure below, and then through the activation function, an output is finally obtained. When several single neurons are combined, the output of some neurons will be used as the input of others, thus forming a neural network. If these neurons are regarded as a whole, the structure has input and output. The layer receiving data input is called the input layer, and the layer outputting the results is called the output layer[7]. The middle layer composed of middle neurons is called the hidden layer or hidden layer. The number of neural network layers is mainly determined by the number of hidden layers. As shown in Figure 2.3, it is a single-layer neural network. This neural network has 4 neurons in the input layer, 5 neurons in the hidden layer, and 2 neurons in the output layer [8].

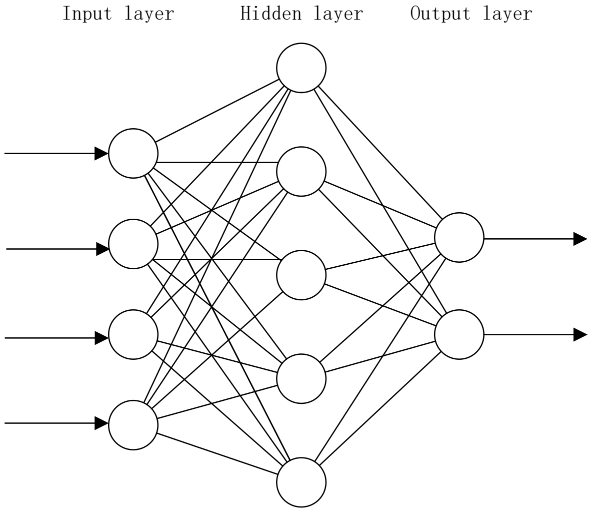

Figure 1. Neural network structure diagram.

The corresponding neural network needs to be designed according to the specific problem to be solved. Generally speaking, the number of neurons in the input layer and output layer is fixed, while the number of layers in the middle hidden layer can be adjusted freely. If compared with machine learning, deep learning is more convenient in these aspects:

1) Stronger fitting ability of deep learning model

Among feature crossing methods, dot product and other methods are relatively simple, so that when the sample data is more obscure and complex, the phenomenon of under fitting often occurs. Deep learning can effectively solve this problem by improving the fitting ability of the model. Take NeuralCF (Neural Network Collaborative Filtering) model as an example, in this model, the point accumulation layer is exchanged with the multi-layer neural network. Theoretically, multi-layer neural network can fit any function, so as long as the number of neural network layers is increased, the problem of insufficient fitting will be solved [9].

2) The structure of deep learning model is more flexible

The structures of deep learning models are different. Most of them can stack network layers with different functions. The simplest is a serial structure, some like a mesh structure, some like a pyramid structure, and so on. The typical cases are Alibaba's DIN and DIE, which adopt a sequence model with attention mechanism and interest evolution simulation in the model structure to achieve a better effect of simulating user behavior.

C. Fundamentals of Deep Neural Networks

The neural network layer inside DNN can be divided into three different types: input layer, hidden layer and output layer. In the absence of special circumstances, the first layer is called the input layer, the last layer is called the output layer, and the number of layers in the middle is called the hidden layer.

Each layer is closely connected. It can be said that a neuron randomly selected from layer i must be connected to any neuron in layer i+1. The small local model is the same as the perceptron, that is, a linear relation input data is used by the neurons in the first layer (not hidden), then the output is provided to the neurons in the next layer, and the output is repeated until the final successful output. The output has the possibility of yes or no (expressed in probability) prediction. Each layer often has one or more neurons, and each neuron calculates a small function, that is, the activation function. The activation function simulates signal transmission to the next connected neuron. The value of the afferent neuron will be compared with the threshold value. When the neuron value is large, the output can pass. The weights formed by the connection between two neurons of the continuous layer are related. The influence of input on the next neuron and the final output is defined by weight. In the neural network, the initial and weight values are randomly generated. Of course, in order to successfully predict the accurate output value, certain weights will be updated in time. Decompose the network mainly by defining some logical building blocks, such as neurons, layers, weights, inputs, outputs, activation functions, and so on. At the end is the learning mechanism (optimizer), which is of great significance in the process of gradual weight optimization of the network (random initialization). The weight value can be optimized to help predict the correct result [10].

III. RESEARCH ON RECOMMENDATION ALGORITHM AND MODEL

A. Recall Algorithm

In the recommendation system, recall is the first step. This algorithm mainly requires to select a small amount of data from the huge item information database quickly and accurately, depending on the user's relevant information and the characteristics of the products they are interested in. After the recall, the next step is the sorting process. In the recall process, huge data volume and high speed requirements are significant features. Therefore, the strategies, models and features adopted by the algorithm should not be too complicated. Data acquisition at the recall stage needs to be carried out quickly. Multi way recall is generally a classic form, such as content-based collaborative filtering recall, user based collaborative filtering recall and popular recall, as shown in Figure 2.

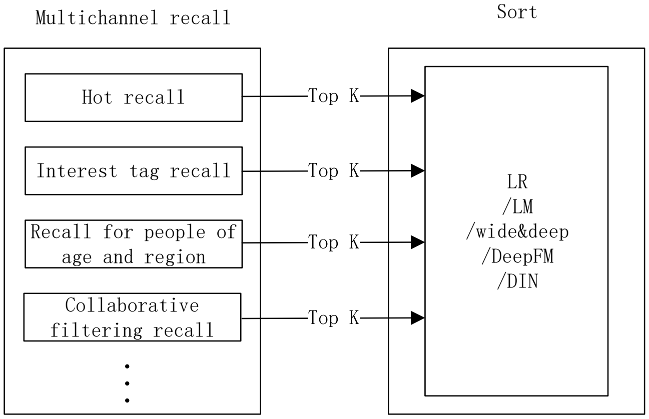

Figure 2. Recall ranking diagram.

To be specific, multi-channel recall means that users can select the products they are interested in from the commodity library in different forms at the same time, and the products selected from multiple ways are summarized and sorted by the sorting model, and finally provided to users to achieve personalized recommendations. The name of multi-channel recall proves that the core of multi-channel recall is "multi-channel". Such multi-channel form ensures the maximum degree of diversification of product content, and also meets the needs of users to the maximum extent[11].

This system will also adopt the collaborative filtering algorithm based on items, and apply tfidf, w2v, etc. to process feature engineering. The collaborative filtering based on items mainly depends on the weight of the items, selects the items that users like, and then calculates the similarity between other goods in the item library and the products that users like, and finally pushes them to the similar products that users once liked. In addition, their methods for calculating the similarity weight of items have different focuses, mainly including the following categories:

- Click, purchase or download co cash as reference data to realize weight calculation. It can be said that for a user who has purchased Class A, Class B and Class C products at the same time. For article A, there are two co occurrence pairs (A, B) and (A, C). Another user has purchased Class A and Class B products, so there is another co occurrence pair (A, B). The meaning of co occurrence is similarity to some extent. In this way, for article A, the similarity weight value of article B is 1+1=2, and the similarity weight value of article C is 1. There may be many different calculation methods for co occurrence. But the guiding significance of this thought is significant.
- The article weight is calculated based on Word2Vec method. The sequence of items is the key point of this method. When the user purchases, clicks or downloads, the products that the user browses will generate the corresponding sequence. In this case, it is considered that this order is relevant, and items with similar distances from each other are also highly similar. The modeling of this method depends on the similarity between the order of goods. Among them, the method of modeling the item sequence according to word2vec is often adopted. According to the difference of weight vectors of different items, the Euclidean distance between item weight vectors is used to divide the similarity of items.
- Item based collaborative filtering combined with TF-IDF method. This method refers to taking the user's browsing and purchase records as a word, while specific items appear as letters in the word. The tf idf value of each item is not missed, and the value reflects the user's preference for items. Finally, you can rely on the item collaborative filtering method to find products that are highly similar to those users like.

B. *Sorting Algorithm*

When some high-quality goods are selected from the recall stage, the sorting stage is entered. The requirements of this stage are high, and users are required to accurately recommend the products they like according to various materials. For sorting algorithms, fast tracking is an important requirement that cannot be ignored. Users need quick feedback, but they also need to implement accurate recommendations. In the final sorting stage, ctr estimation is performed on all goods, and then topK data is delivered upwards [12].

The design of this paper in the sorting stage adopts the double tower model, and the model architecture is shown in the figure below.

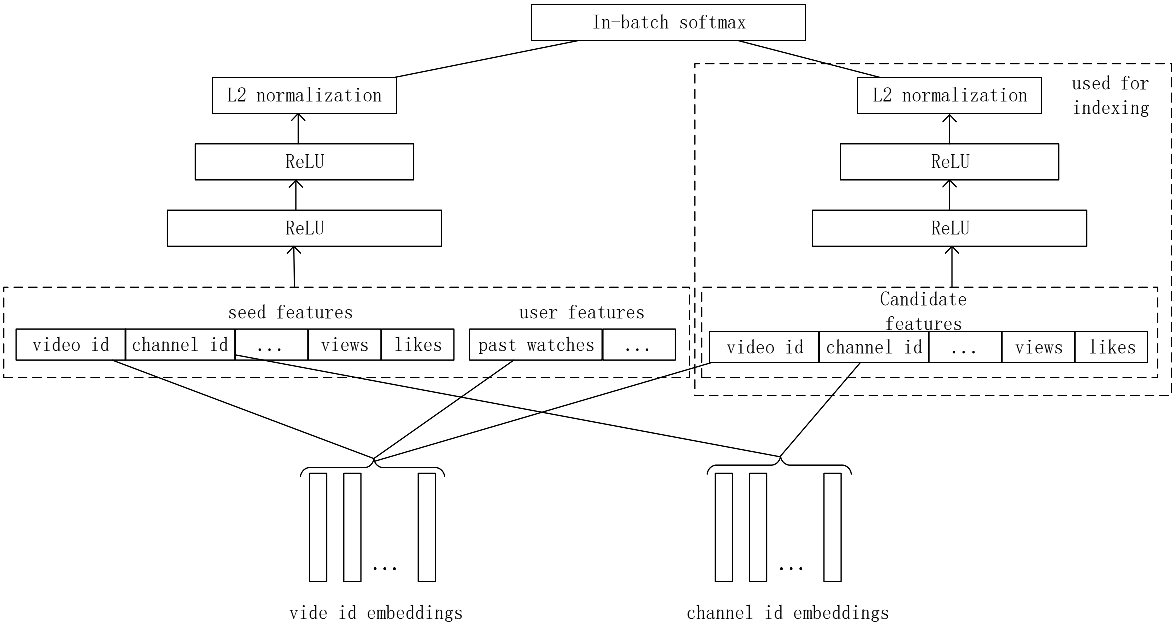

Figure 3. DNN double tower model architecture diagram.

The double tower model in this paper constructs the embedding on the request side and the embedding on the commodity side. The output of the two towers, and the final model is the output of their respective embedding vectors, which is the result of the inner product of the two embeddings. The inner product formula of the vector is as follows:

$$s\left(x, y\right) = u\left(x,\theta\right), v\left(y,\theta\right) \quad (6)$$

Based on the output of the softmax function and user preference weight, the loss function is in the form of a weighted logarithmic likelihood function, as follows:

$$L_T\left(\theta\right) := -\frac{1}{T}\sum_{i\in[T]} r_i \cdot \log\left(\mathcal{P}\left(y_i \ / \ x_i;\theta\right)\right) \quad (7)$$

When there are a large number of commodities in the commodity library, that is, (M is very large), it is very difficult to accurately calculate the softmax function above. In order to overcome this problem, it is obviously an appropriate method to sample a full set of commodities. The traditional form is to sample the negative samples required for training from a fixed set, however, the more efficient method for the traditional form is to sample a batch of data in the real-time stream and obtain the negative sample for training, which is also the negative sample in this batch. However, this method may cause errors, making it possible for some popular samples to become negative samples. Therefore, it is necessary to perform logQ correction on the inner product calculated by the two embedding vectors above, that is, to perform the embedding inner product correction.

$$s^c\left(x_i, y_j\right) = s\left(x_i, y_j\right) - \log\left(p_j\right) \quad (8)$$

Where, p_j represents the sampling probability of commodity j randomly selected into batch. Then use SGD to update the parameters. The specific model training algorithm is as follows:

- Obtain a batch sample from the real-time data stream;
- The sampling probability pi is obtained based on the sampling probability estimation algorithm to be mentioned below;
- Calculate the modified loss function described above.

IV. PRODUCT RECOMMENDATION SYSTEM DESIGN

The functional requirements of the system are the tasks that must be established before the system development, which specify the functions that the system must achieve and ultimately need to meet the requirements. Through the previous research and analysis, the modules of this system are determined as user interface module, search function module, recommendation function module, and data processing module [13].

During the development of this program, pycharm is used as the IDE to write Python code. At the same time, Tensorflow gpu and Keras are used as the development tools for deep learning algorithms.

A. Functional Requirement

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

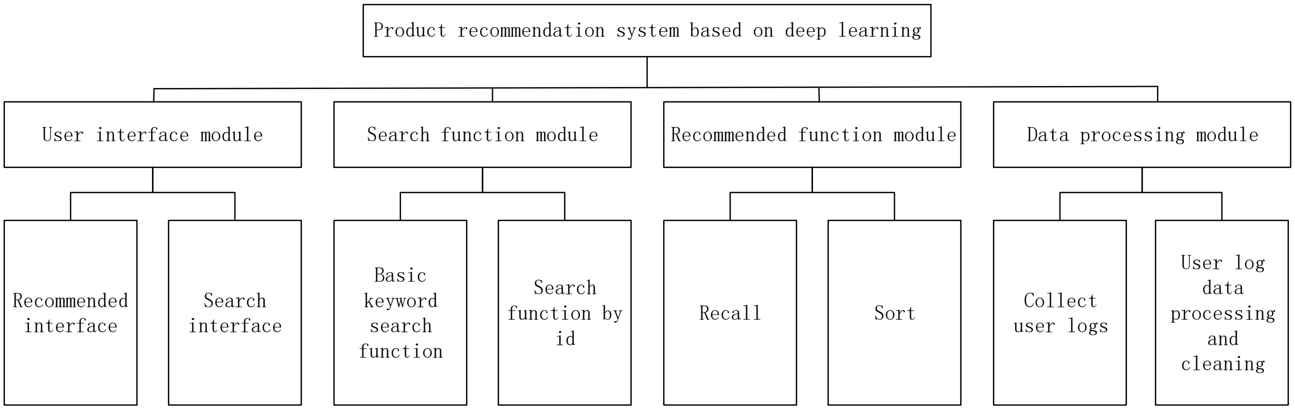

Figure 4. Function module diagram.

This topic is to design and complete a product recommendation system with complete functions. The program can recommend products to users in real time, and provide simple product search function and simple display function. The functional module is shown in Figure.

- User interface module: log in and register, search, recommend, browse favorites and purchase;
- Search function module: basic search function, keyword search;
- Recommendation function module: recall, sorting, popular recommendation, personalized recommendation;
- Data processing module: provide log service, transmit the user's information to the background in real time, and clean the data.

B. Technical Scheme for Recall Sequencing

1) Recall part (machine learning)

The recall part is a multi-channel recall. Currently, collaborative filtering is used, including item based, user based, and model based recalls. Matrix decomposition and failure are used.

2) Sorting part (deep learning)

First of all, the recall part will filter all commodity IDs, and the IDs arriving at the sorting module are only topN. The sorting module is responsible for sorting the topN commodities. (The value of N is controlled by the user). It is recommended to use tensorflow to build some parts and use the python environment of conda.

C. System Architecture Design

The recommended architecture adopts big data lambda architecture, and the front and rear end adopt golang web framework gin for connection, and each service uses grpc microservice framework for connection. This topic uses MVC mode, and the system architecture is shown in Figure 5.

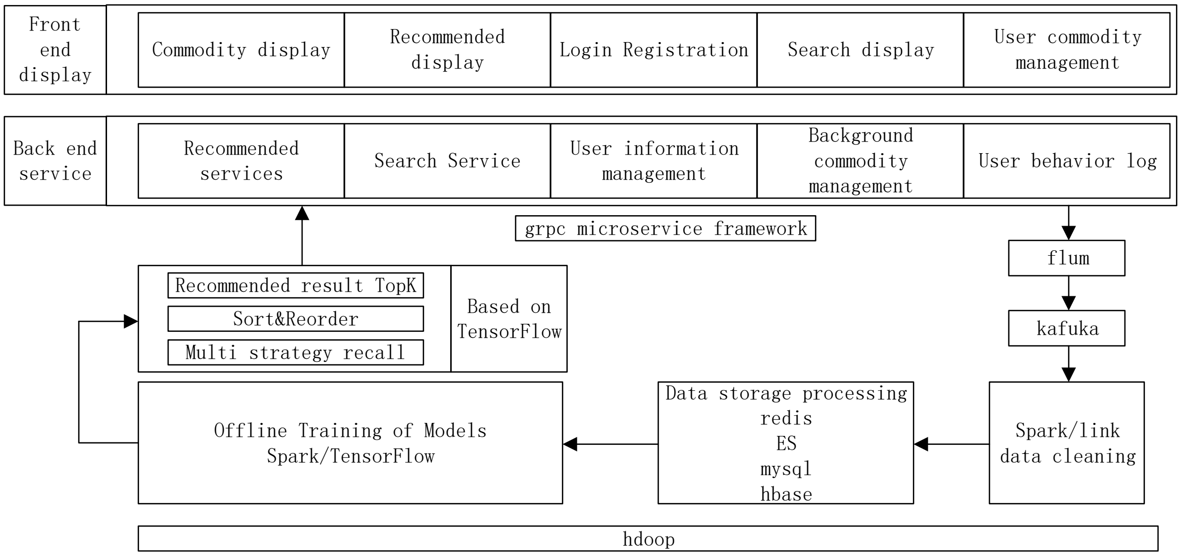

Figure 5. System architecture diagram.

D. *Implementation of Commodity Recommendation System*

First, it introduces the program development environment, and then analyzes the technical feasibility of program development. This program is developed on the PC side using the MAC environment and Python language. At the same time, it involves in-depth learning and network training, so it needs to use high-performance GPU for computing acceleration [14].

The specific development environment is as follows: the development platform is PC; The software environment operating system is MAC; The development language is Python; The IDE is pycharm vscod.

Next, we will introduce various tools and frameworks used in the development of this program. During the development of this program, pycham is used as the IDE to write Python code. At the same time, Tensorflow gpu and Keras are used as the development tools because of the deep learning algorithm [15].

The program development mainly uses the following development tools and frameworks:

1) Tensorflow gpu

Tensorflow is a symbolic mathematical system based on data flow programming, which is widely used in the programming implementation of various machine learning algorithms. Its predecessor is Google's neural network algorithm library DistBelief. Its gpu version can improve the training speed.

2) Keras

Keras is an open source artificial neural network library written by Python. It can be used as the high-level application program interface of Tensorflow to design, debug, evaluate, apply and visualize in-depth learning models.

3) CUDA

NVIDIA has introduced a general GPU parallel computing architecture. Because the deep neural network will involve a lot of matrix operations when training and using, the use of CUDA to operate on the GPU can greatly improve the efficiency of the model.

In recent years, the development of deep learning has driven the development of the field of artificial intelligence. Research on recommendation systems through deep learning has emerged in an endless stream, and there are many open source resources. Therefore, it is technically feasible to develop a commodity recommendation system [16].

V. CONCLUSIONS

This paper first introduces the development background and research significance of this topic. By analyzing the status quo of recommendation technology and deep neural network, and referring to the development status at home and abroad, this paper explains the significance of the development of this system. Then it explains the technology used in the system development. This system uses Python language and Tensorflow architecture. In the requirement analysis part, the feasibility is analyzed in many aspects, and then the functional requirements and non functional requirements of the system are introduced, and the three functions of the system are analyzed. It describes how different functions are progressing, clearly and concisely shows the system construction method, how different modules collaborate, and shows the completion of the system and the core code.

In a word, the design of this system can meet the needs of the development of the times, so that people can enjoy the application of the recommendation system in the commodity industry. It is an efficient and novel system.

REFERENCES

[1] Elkahky A M, He X. A multi-view deep learning approach for cross domain user modeling in recommendation systems//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015:278-288.

[2] Huang P S, He X, Gao J, et al. Learning deep structured semantic odels for web search using clickthrough data//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. Boston, USA, 2016:29-34.

[3] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems.Boston, USA, 2016:7-10.

[4] Guo H,Tang R, Ye Y, et al. DeepFM:A factorization-machine based neural network for CTR prediction//Proceedings of the 26th International Joint Conference on Artificial Intelligence.Melbourne,Australia, 2017:1725-1731.

[5] Chen C, Meng X, Xu Z, et al. Location-aware personalized news recommendation with deep semantic analysis. IEEE Access, 2017:173-182.

[6] Rendle S. Factorization machines//Proceedings of the 2010 IEEE 10th International Conference on Data Mining. Sydney, Australia, 2010:995-1000.

[7] LYU L, MEDO M, YEUNG C, et al. Recommender systems ［J］ . Physics Reports, 2012, 519(1):1－50.

[8] Shan Y, Hoens T R, Jiao J, et al. Deep crossing: Web-scale modeling without manually crafted combinatorial features. SIGKDD 2016: 255-262.

[9] Song W, Shi C, Xiao Z, et al. Autoint: Automatic feature interaction learning via self-attentive neural networks. CIKM 2019: 1161-1170.

[10] Chen Q, Zhao H, Li W, et al. Behavior sequence transformer for e-commerce recommendation in Alibaba. Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data. 2019: 1-4.

[11] Hao-Jun Michael Shi;Dheevatsa Mudigere;Maxim Naumov; Jiyan Yang Compositional Embeddings Using Complementary Partitions for Memory-Efficient Recommendation Systems [C], 2020.

[12] Antonio Ginart; Maxim Naumov; Dheevatsa Mudigere; Jiyan Yang; James Zou Mixed Dimension Embeddings with Application to Memory-Efficient Recommendation Systems. [J] IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2019.

[13] Hao-Jun Michael Shi; Dheevatsa Mudigere; Maxim Naumov; Jiyan Yang Compositional Embeddings Using Complementary Partitions for Memory-Efficient Recommendation Systems. [J] IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2019.

[14] Yongxin Yang and Timothy Hospedales. 2016. Deep multi-task representation learning: A tensor factorisation approach. arXiv preprint arXiv:1605.06391 (2016).

[15] RAMEZANI M, MORADI P, AKHLAGHIAN F. A pattern mining approach to enhance the accuracy of collaborative filtering in sparse data domains [J]. Physica A: Statistical Mechanics and its Applications, 2014, 408:72–84.

[16] ZHANG F, YUAN N J, LIAN D, et al. Collaborative Knowledge Base Embedding for Recommender Systems[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 353–362.