Visualization of data analysis platform — Taking QQ music recommendation system as an example

Xinyue Li

The High School Affiliated to Renmin University of China, Beijing, China, 100089

lxy20050219@163.com

Abstract. With the rapid development of big data technology, people's demand for personalized music recommendation systems is growing more and more urgent. However, the current music recommendation system still has some problems, such as inaccurate recommendations and too slow recommendation speeds, as well as cold starts and data sparsity caused by massive data. In order to design and implement a music recommendation system for the recommendation system storage caused by the continuous increase of data, insufficient storage, and computing power, this paper improved the QQ music recommendation system based on the collaborative filtering recommendation algorithm of the offline data warehouse technology project. After testing, the music recommendation system designed in this paper has good stability, scalability, and efficiency.

Keywords: Data Warehouse, Text Similarity, Big Data, Recommendation.

1. Introduction

Before the advent of recommendation systems, people could solve the problem of information overload through search engines. The use of search engines by users is an information filtering behavior that they actively participate in and search for desired content according to the input keywords. However, if users are not particularly clear about their needs or cannot find suitable keywords to describe them [1], then they believe that search engines are useless. For this reason, the music recommendation system came into being. The recommender system uses information gleaned from the user's historical using behavior and potential preferences to recommend information that may be of interest to the user, thereby improving the user's satisfaction. The solutions of search engines and recommendation systems complement each other, jointly coping with the problem of information overload so as to better serve users.

The music recommendation system based on the layered realization of a data warehouse is an effective means to solve massive data calculations. In many branches, the music recommendation system is also the focus of research. At present, the music libraries of popular music platforms contain a huge number of music tracks, which are divided into different languages, eras, emotions, radio stations, etc. For users, it is impossible to listen to all the tracks, and the ability and energy to search for a single track is limited. Moreover, music may be a background sound for people's lives. People can listen to music while focusing on doing other things, which leads to the vagueness of users' needs. They only need to hear music that matches their preferences, and the recommendation system will do the rest. The purpose is for users' potential preferences to be displayed and recommended to other users [2].

Today, with the increasing amount of data, more and more data is stored in business systems, and stand-alone systems can not meet the recommended music system storage and computing Hadoop. The combination of Hadoop and recommendation systems enables recommendation algorithms to perform calculations on large amounts of data. Besides, this Hadoop architecture can provide higher processing efficiency for the processes of data storage and computing and maximize the computing efficiency of the recommendation algorithm [3]. Therefore, a music recommendation system can store massive amounts of data, process data quickly, run algorithms quickly, and provide personalized recommendations, which has great practical value.

In order to advance a music recommendation system for the recommendation system storage caused by the continuous increase of data, insufficient storage, and computing power, this paper improved the QQ music recommendation system based on the collaborative filtering recommendation algorithm of the offline data warehouse technology project. The algorithm of the recommendation system is summarized in detail. Related technologies such as Hadoop and Spark are also analysed. The advanced QQ music recommendation system is completed and is based on the Hadoop distributed framework. For the collection and transmission of data, the combination of Flume and Kafka is adopted to ensure the integrity and security of the data. Then, HDFS is used for distributed storage, and Hive is used to build a music data warehouse, and operations such as data cleaning, decontamination, and hierarchical processing are completed. The data preprocessing module is completed, which is beneficial to the calculation of subsequent recommendation algorithms. Secondly, for the data calculation, the Spark and MapReduce distributed computing frameworks are used to accelerate the code operation. It hopes to provide a better music platform for music lovers.

2. Technical introduction

Flume data collection technology

Flume is a distributed, reliable, and high-availability massive log collection, aggregation, and transmission system. It supports customizing various data senders in the log system for data collection; at the same time, Flume provides the ability to simply process data and write to various data receivers (such as text, HDFS, Hbase, etc.). Flume's data flow is run through events. The event is the basic data unit of Flume. It carries log data (in the form of a byte array) and header information. These events are generated by a source outside the agent. When the source captures the event, it will perform a specific format, and then the source will store the event and push it into (single or multiple) channels. The channel can be thought of as a buffer that will hold events until the sink has finished processing them. The sink is responsible for persisting logs or pushing events to another source [4].

1) Reliability of flume. When a node fails, logs can be transferred to other nodes without loss. Flume provides three levels of reliability guarantees, from strong to weak: First, end-to-end: the agent that receives the data first writes the event to the disk, and then deletes it after the data transmission is successful; if the data transmission fails, it can be resent. Second, store on failure: this is also the strategy adopted by scribe; when the data receiver crashes, the data is written to the local, and after recovery, it continues to send. Third, Besteffort: after the data is sent to the receiver, it will not undergo verification.

2) The recoverability of the flume still depends on the channel. It is recommended to use FileChannel. Events are persistent in the local file system (poor performance).

HDFS distributed storage technology

HDFS is the abbreviation of the Hadoop Distributed File System (Distributed file storage and computing platform). HDFS adopts the principle of master-slave structure design, which consists of a worker node named NameNode and slave nodes named DataNode to make. This NameNode master node is mainly responsible for recording and tracking the metadata of all files in the system, which is equivalent to the index directory of all files. Multiple slave nodes, DataNodes, are responsible for storing files. Files are stored in data blocks. The default size of each data block is 128 MB. Hadoop is suitable for processing large files but not for processing small files. Too many small files will waste resources on the NameNode. The processing of HDFS stored data is like in Figure 1 [5].

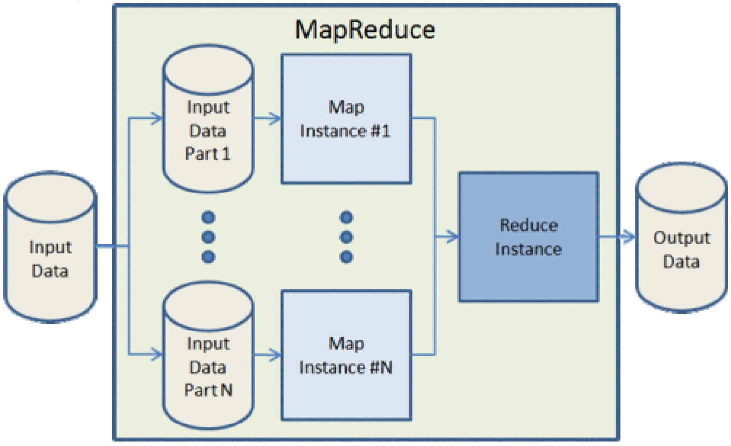

Figure 1. The processing of mapreduce.

Hive data warehouse technology

Hive is a data warehouse tool based on the Hadoop framework for data extraction, transformation, and loading. It is a mechanism for storing, querying, and analyzing large-scale data stored in Hadoop. Hive is very suitable for statistical analysis of data warehouses. The hive is like Figure 2.

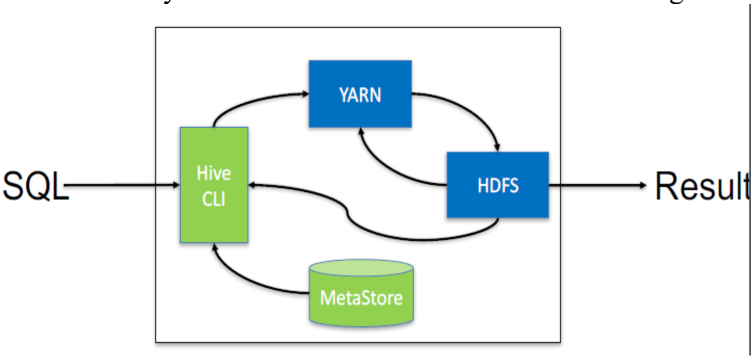

Figure 2. The processing of hive.

Spark distributed computing technology

Spark is a distributed computing framework based on the Hadoop platform. Compared with the MapReduce computing framework, Spark is characterized by the fact that it can perform storage operations in memory. In big data computing work, the biggest bottleneck is disk I/O speed [6]. The advantage of using Spark is that it reduces the I/O process of data on disk, improves data read and write capabilities, and also improves data processing speed. Spark is 10~100 times faster than MapReduce in performance. Spark supports the use of a variety of scenarios, including graph computing, stream computing, batch processing, and machine learning [7]. The main components of the current Spark ecosystem are shown in Figure 2-9, including the core base libraries. Spark Core, Spark SQL, is used to query data in huge datasets. It can interact with Hive and read both local files and files in HDFS. Spark Streaming is a distributed stream processing framework derived from Spark, which is generally used in real-time recommendation systems. MLlib contains some commonly used machine learning algorithm libraries to meet the algorithm requirements. GraphX is Spark's distributed graph computing framework.

Java web technology

The web resources on the Internet for external access are divided into: Static web resources (such as html pages) refer to the data in the web pages for people to browse that is always unchanged. Dynamic web resources refer to the data in the web page for people to browse being generated by the program, and the content seen by visiting the web page at different time points is different. Static web resource development technology: HTML, CSS. Dynamic web resource development technology: JavaScript, JSP/Servlet, ASP, PHP, etc. In Java, dynamic web resource development technologies are collectively referred to as Java Web [8].

3. QQ Music Platform Analysis

QQ music data characteristics

According to the latest data in 2022 provided by Baidu Index [9], the age of the main users of QQ Music is concentrated among 20 - 29 years old, and it gradually declines with age. This may be related to the simple and youthful design of QQ Music, which is more popular with young people. At the same time, female users of QQ Music account for 34.94% of the total, and male users account for 65.04% [9]. There is a big inconsistency with the overall distribution of male and female ratios across the network. The lack of a QQ music community atmosphere may be one of the reasons for this result. In terms of the geographical distribution of users, the vast majority of QQ Music users live in first-and second-tier cities, and there are fewer users in third- and fourth-tier cities and below. From the perspective of application preferences, exclude the "communication and social networking, news, audio and video, and online shopping" of the four giants of mobile APPs and choose the "life, mobile tools, reading and financial management, education, business travel, work" and other preferences that are highly concerned. Looking at it, we can preliminarily summarize some user characteristics as follows: Pursue quality of life, prioritize efficiency, enjoy reading, value education, and have certain investment and financial capabilities [10].

Data analysis algorithm and research of QQ music

Through the basic analysis of user data, it can be inferred that the main users of QQ Music are the following groups: Young college students receiving education in first- and second-tier cities, Young white-collar workers in the workplace who pay attention to work and life efficiency. Middle-aged professionals who pursue quality of life and have strong spending and investment capabilities. For users who listen to music most of the time, the music is a background sound, and there are few active behaviors. Music platforms are different from film and television platforms in that they can significantly rate movies or TV dramas they have watched.

It implicitly collects the user's behavior,calculating the user's score for each song according to the weight of the behavior. For example, if a user has collected a certain song, it can be considered that the user is interested in this type of song, and the next time The weight of such songs in the recommended list will be increased; the user has carried out an evaluation of a certain song or singer. If it is blocked, similar songs or singers will not be recommended on the recommendation page. A user's contribution to the song, such as switching, searching, downloading, favorite, like, share, comment, block, buy, following, etc., will be useless [10]. According to the behavior of the above users, the score design is carried out, and the user designs the following points for the status of a song. When the user has multiple behaviors for a song, the highest rating value is selected. Through the display of the music search function, users can quickly obtain the music they want. The search function is implemented through the ElasticSearch server, which can realize precise queries and fuzzy queries. The search function interface is shown. Enter "moon" in the search box to get "moon" songs by different singers. For the recommendation result, each user's behavior will be recorded as data: in the background log file, the data is transmitted to the recommendation service module through big data components such as Flume and Kafka, the data is processed, and the hybrid recommendation algorithm is used for calculation, and finally to the recommended list. Finally, the recommendation list is returned to the front-end page for display. For the user, the final interface is the music recommendation list provided by the system.

4. Visualization technology

Here we used technologies including the following: Java, ElasticSearch, Kafka, and Flume to implement search and recommendation functions. Through these technologies, we realize the visual display of the front end.

Requirements of visualization technology

We use JavaScript to implement the basic functions of user registration and login in the system. The user login system includes new users and old users. New users need to use the registration function. User information is stored in the MySQL user information table, and old users can log in directly.

Design of visualization technology

4.1.1. Design of collection module. Use Flume to monitor and collect data in HDFS, collect data in HDFS into data warehouse layering, divide it into ODS layer, DWD layer, DWS layer, and ADS layer based on data warehouse layering, and calculate indicator results in the ADS layer. The calculation results are stored in HDFS.

4.1.2. Design of storage module. In the data warehouse layering, the data is divided into four layers, which are divided into ODS layer, DWD layer, DWS layer, and ADS layer, and the index results are calculated in the ADS layer, and the calculation results are stored in HDFS.

4.1.3. Design of data analysis module. The implementation steps of the recommendation algorithm on the Hadoop platform are as follows:

(1) Convert data from user tables, music tables, behavior data tables, and other tables to csv or text format.

(2) Set the environment variables and add the installation paths for software such as Python, Pyspark, Spark, etc.

(3) The hybrid recommendation algorithm trains the model by calling the command script and passing in the relevant model parameters for model training.

(4) Use Java to dynamically pass parameters to Python, then call Python to execute the script and output the experiment's final result.

(5) Analysis and recommendation of the results; storing the returned results in the MySQL database; performing format conversion; and displaying the recommendation list on the system page.

5. Realization of visualization of QQ music data analysis platform

(1) User registration and log-in: Registration and log-in are the basic functions of each system. Although some systems can directly enter the system without logging in, they cannot associate and store the user's basic information and behavior records, and the recommendation service will fail due to missing data. For accurate recommendation, the music recommendation function and personalized recommendation will not be possible. The system login and registration interface are shown in Figures 3. User login includes new users and old users. New users need to use the registration function so that user information can be stored in the database, and old users can log in directly. User login functions include user registration and user login.

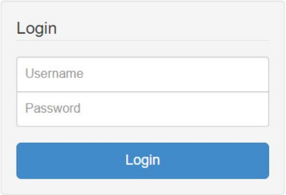

Figure 3. User login interface.

(2) Display of music search function: The search function is implemented through the MySQL server, which can realize precise and fuzzy queries.

(3) Recommendation function display: For the recommendation results, each user's behavior will be digitized and recorded in the background log file, and the data will be transmitted to the recommendation service function module through big data components such as Flume and Kafka for data processing. The time-weighted hybrid recommendation algorithm is used to calculate the final recommendation list. Finally, the recommendation list is returned to the front-end page for display. When evaluating the performance of the algorithm, the recommendation results are comprehensively sorted, and the best top N will be selected. For users, the interface they finally see is the music recommendation list provided by the system. Among them, the list length is 20 songs. The screenshot shows the top 10 recommended songs.

6. Conclusion

In conclusion, this paper mainly focuses on advancing the QQ music recommendation system. After testing, the music recommendation system designed in this paper has good stability, scalability, and efficiency.

Due to the influence of many factors in reality, the system can be further improved, mainly including:(1) Due to the limited experimental conditions, the calculation speed needs to be improved. For Hadoop, the more servers, the better the computing power and storage capacity. (2) The music content can be further analyzed, the audio rhythm of the music can be analyzed, and more features of the music can be extracted, so as to improve the diversity of the recommended content.

References

[1] N D Almalis, G A Tsihrintzis and N Karagiannis. A content based approach for recommending p ersonnel for job positions [C]. The 5th International Conference on Information, Intelligence, Systems and Applications, 2014:45-49.

[2] Xue, Feng,He.Deep Item-based Collaborative Filtering for Top-N Recommendation[J].ACM tra nsactions on information systems.2019,37(3).33.1~33.25.doi:10.1145/3314578.

[3] Ibrahim, Othman, Nilashi.A recommender system based on collaborative filtering using ontolog y and dimensionality reduction techniques[J]. Expert Systems with Application.2018,92(Feb.).507-520.

[4] Deger Ayata,Yusuf Yaslan,Mustafa E. Kamasak.Emotion Based Music Recommendation Syste m Using Wearable Physiological Sensors[J].IEEE Transactions on Consumer Electronics.20 18,64(2).196-203.

[5] Cano, Erion, Morisio,.Hybrid recommender systems: A systematic literature review[J].Intelligen t data analysis.2017,21(6).1487-1524.

[6] Yong Wang, Jiangzhou Deng, Jerry Gao,.A hybrid user similarity model for collaborative filteri ng[J].Information Sciences: An International Journal.2017.418/419102~118.doi:10.1016/j.in s.2017.08.008.

[7] Sattar Asma, Ghazanfar Mustansar Ali, Iqbal Misbah.Building Accurate and Practical Recomme nder System Algorithms Using Machine Learning Classifier and Collaborative Filtering[J]. Arabian journal for science & engineering.2017,42(8).3229-3247.doi:10.1007/s13369-016-2

410-1.

[8] Nilashi, Mehrbakhsh, Jannach.Clustering-and regression-based multi-criteria collaborative filtering with incremental updates[J]. Information Sciences: An International Journal.2015.293

[9] Yue Shi, Martha Larson, Alan Hanjalic.Collaborative Filtering beyond the User-Item Matrix : A Survey of the State of the Art and Future Challenges[J].ACM computing surveys.2014,47(1).

[10] Min-Ling Zhang,Zhi-Hua Zhou.ML-KNN: A lazy learning approach to multi-label learning[J].Pattern Recognition.2007,40(7).2038-2048.doi:10.1016/j.patcog.2006.12.019.