Research on Key Technologies of Personalized Education Resource Recommendation System Based on Big Data Environment

Yunli Cheng¹, Xu Bu²
Guangzhou Nanyang Polytechnic College, Guangzhou 510925, China

Abstract. This paper focuses on the construction of personalized education resource recommendation system and the research of key technologies in the big data environment, designs a personalized recommendation system architecture that can handle PB education resources, and implements an educational resource personalized recommendation system based on this architecture. A hybrid recommendation algorithm based on content and collaborative filtering is designed to meet the recommendation novelty requirements and solve the cold start problem of the recommendation system. Research a flexible, reliable, high-performance personalized recommendation system architecture that can store and process PB-level data and can be recommended in real time, research personalized recommendation engine, data pre-processing and data mining model construction, and implement the recommendation system. It is important to determine the availability of the system by performing related functions and performance verification.

1. Topic background
Nowadays, with the practice and promotion of major Internet giants such as Amazon and Taobao, it is the hotspot and difficulty of competing research on the structure of personalized recommendation system on the basis of big data. The architecture of traditional personalized recommendation system can not realize PB. The storage and processing of the level data, although there are offline batch pre-processing or multiple filtering methods to reduce the amount of information and then calculate the implementation of the recommendation model, but it will inevitably lead to user behavior delays / delayed or distortion, the timeliness of the recommended model Accuracy cannot be guaranteed in a timely and effective manner. So it is urgent to study the need to store and process PB-level personalized recommendation systems. This project intends to use the extension data mining technology for big data as the research object, aiming to lay a theoretical foundation for data analysis technology, data extraction technology and extension data mining technology in the era of big data, and finally build a data based on extension data. An educational resource personality recommendation system for mining technology.

2. Main content framework of the study
The main content of the project research is to design a personalized recommendation system architecture that can handle PB-level educational resources, and implement an educational resource personalized recommendation system based on this architecture. The project mainly designs and implements a personalized recommendation system to meet the educational resources of colleges and universities, and teachers and students can quickly find educational resources that are meaningful to them.
During the construction of the personalized recommendation system, the internal logic structure constructs two subsystems simultaneously with the other two external systems, namely the internal classification tree subsystem, the personalized recommendation Web subsystem, the external resource sharing platform, and the unified login system. System functions. The overall structure of the system function design and application is shown in Figure 1:

![Figure 1: The overall framework of the personalized recommendation system for educational resources](image1)

**3. Analysis of key technologies of big data computing framework**

MapReduce is a distributed computing framework, first proposed by Google, and is well known for its ability to process large amounts of data in distributed clusters of thousands of inexpensive servers. It is widely used in big data processing. MapReduce can complete the reading and writing and processing of large data sets at the same time. The user generates a set of intermediate key-value pairs by defining a Map function to process the key-value data set, and then processes the intermediate set through the Reduce function and finally merges all the values of the same key. This project uses HadoopMapReduce, which is a widely used open source MapReduce implementation. The detailed execution process of MapReduce is shown in Figure 2.

![Figure 2: Hadoop MapReduce execution process](image2)
The Map function reads the HDFS fragment data nearby, and writes the Map processing result to the ring memory buffer first. If the buffer writes a certain threshold, it starts to split into the file. During this period, the data is partitioned according to the number of Reduce and press. The same key value is written into the output partition file after sorting in the partition; when the Reduce function starts, it selects the data partitions that need to be copied from the nodes of the completed Map output results to their own nodes and merges and sorts them. The data is used as input for Reduce processing and finally the result is output to HDFS.

Education Resource Personalization Recommendation System Personalized recommendation engine, Mahout is a Java program framework that can run large-scale machine learning algorithms based on Hadoop. This project uses Mahout's Item recommendation engine part, and the Mahout recommendation engine is from the original independent framework. Taste is implemented.

4. Key research difficulties
(1) Construct a mathematical model based on a three-dimensional matrix. Establish a mathematical model based on three-dimensional matrix. Each attribute of the obtained data is defined as one dimension in space, thereby forming an N-dimensional space.

(2) Analysis method of association rules based on the idea of co-evolution. The improved genetic algorithm and particle swarm optimization algorithm are used to iterate simultaneously for the two populations to make up for the shortcomings of the premature convergence and slow convergence of the genetic algorithm, and effectively search for high-quality association rules.

(3) Based on the three-dimensional matrix model, the clustering algorithm based on undirected hypergraph and the association rule loop redundancy detection based on directed hypergraph are used for data analysis of the project. The application of clustering algorithm based on undirected hypergraph produces higher quality clustering results; the association rule loop redundancy detection method based on directed hypergraph can effectively search for high quality association rules from the survey database and Reduce redundancy and loop rules.

(4) Using Mahout and MLib to construct an offline, near-line and real-time three-layer recommendation structure. We realize flexible big data processing capability through Spark and Hadoop technology, and form a label data source to better organize resource model by constructing discipline classification tree. And the user model, using a hybrid recommendation algorithm based on content and collaborative filtering, solves the cold start problem and improves the accuracy of the algorithm.

5. Research technology route
The research idea of the subject is that the problems that arise in the education research are too complicated, and the existing methods are difficult to accurately analyze the problems. Based on the method of big data, the education problems are studied. The technology and application research of the personalized recommendation system based on big data, summarizes the application aspects and technical characteristics of the existing personalized recommendation technology, analyzes the advantages and disadvantages of different scenarios, and analyzes the differences based on the individualized recommendation fields of educational resources. Research on how to use the big data personalized recommendation system in educational resources. The topic focuses on the construction of the personalized education resource recommendation system and the research on key technologies in the big data environment. Under the premise of combing the existing research results, based on the needs of practical development, the following research methods are adopted, from multiple perspectives. The application of the construction of personalized education resource recommendation system, combined with data preprocessing technology, personality engine recommendation, data mining technology and other issues and countermeasures.

(1) Data Preprocessing and 3D Matrix Model
Firstly, a large amount of redundant data is preprocessed by means of data cleaning, data integration, data selection, etc., mainly by different methods such as binning, clustering, regression, and data
integration, so that the data meets the requirements of the data model.

The mathematical model is based on attributes. For example, in a questionnaire, the attributes correspond to the questions in the questionnaire and the intercepted keywords, and the attribute values correspond to the options and key attitudes of the questions. Then we define each attribute as a dimension in space to form an N-dimensional space. Each person's information is represented by an MXN matrix (M is the maximum number of options for the topic), and its choice in each dimension is called an element. The types of elements include vectors and matrices, with the choice of "1" and unselected with "0". Thus each person's information includes N elements. The single-choice questions, multiple-choice questions, and sorting questions of one of the data sources are used as examples, as shown in Table 1. There are three questions in Table 1, which means that there are three dimensions. The first single choice question is B, and the corresponding position is set to 1. The second multiple choice question selects ACE, which is represented by vector, and is also set to 1 at the corresponding position. The representation of the track sorting problem is slightly different from the first two questions because the sorting questions are in order, assuming the order is CBADFIHG. We need to perform dimension enhancement processing, which is represented by a matrix. The first column of the matrix represents the first option in the sorting problem. In this case, C, the position corresponding to C is 1. The second column indicates the second option in the sorting problem. In this case, B, the position of B corresponds to 1, and so on. Then, suppose that the questionnaire information with the N-question is represented by a matrix of size $M \times N_1$, $N_1 \geq N$ ($N_1$ is the matrix after the dimension addition process). K's N-question questionnaire information can be represented by K matrices of size $M \times N_1$.

| Radio | Multiple | Sorting problem |
|-------|----------|-----------------|
| A     | 0        | 1 0 0 1 0 0 0 0 0 |     |
| B     | 1        | 0 0 1 0 0 0 0 0 0 |     |
| C     | 0        | 1 0 0 0 0 0 0 0 0 |     |
| D     | 0        | 0 0 0 0 1 0 0 0 0 |     |
| E     | 0        | 1 0 0 0 1 0 0 0 0 |     |
| F     | 0        | 0 0 0 0 0 1 0 0 0 |     |
| G     | 0        | 0 0 0 0 0 0 0 0 1 |     |
| H     | 0        | 0 0 0 0 0 0 1 0 0 |     |
| I     | 0        | 0 0 0 0 0 1 0 0 0 |     |
Next, we use the method to convert and represent the data of each data source. After getting several 3D matrices we do data integration. All rectangles are combined according to certain rules to form a three-dimensional matrix containing all the information, that is, the mathematical model we built. The three-dimensional matrix mathematical model is shown in Figure 4.

![Figure 4: Schematic diagram of the three-dimensional matrix mathematical model](image)

(2) **Association Rules Mining Based on Co-evolution**

The task of association rule mining is to find a strong association rule with the minimum support and minimum confidence given by the user in the transaction database D. The item set $X \cup Y$ corresponding to the strong association rule $R : X \rightarrow Y$ must be a frequent item set, and the confidence level of the association rule $R : X \rightarrow Y$ derived by the frequent item set $X \cup Y$ can be calculated by the support of the frequent item sets $X$ and $X \cup Y$. Therefore, association rule mining can be broken down into two steps:

1. Find all frequent itemsets in the database that satisfy the minimum support given by the user;
2. Generate all association rules that satisfy the minimum credibility given by the user on the basis of frequent itemsets.

The basic model of association rule mining is shown in Figure 5.

![Figure 5: Basic model of association rule mining](image)
Traditional genetic algorithms are not able to use feedback information in time and the local search ability of the algorithm is poor, which makes the simple genetic algorithm more time-consuming, and the search efficiency is lower in the late evolution. Because there is no choice, crossover and operation, the PSO algorithm has a relatively simple algorithm structure and fast running speed. However, if a particle finds a current optimal position during the running of the algorithm, other particles will quickly approach it. If the optimal position is the local best, this will mislead all individuals to approach the position, so that the whole is trapped in the local optimal solution, and premature convergence occurs.

(3) Hypergraph and its establishment in three-dimensional matrix model

In order to solve the problem of association rule redundancy. This project conducts in-depth research on the redundancy of association rules of hypergraphs and proposes a processing method.

The directed hypergraph adds a direction to the superedge of the undirected hypergraph, indicating the order between the superverting vertices. On this basis, we make some necessary extensions to some properties of the undirected hypergraph to the directed hypergraph, and at the same time find the unique properties of the directed hypergraph. Of course, the main purpose of studying directed hypergraphs is to solve the problems in practical applications. This project will use directed hypergraphs to represent association rules, and use the nature of directed hypergraphs to solve the redundancy and loops in association rules mining. Road problem.

Figure 6: Representation of the directed hypergraph $H$
A directed hypergraph is an $H=(V,E)$ binary pair, $v$ is a vertex set, and $E$ is a directed superedge set. The directed super-edge $e \in E$ is defined as an ordered pair $(T(e), H(e))$, where $T(e)$ and $H(e)$ are both subsets of $V$, and $T(e) \cap H(e) = \emptyset$ and $T(e)$ are called tail nodes of the directed super-edge, and $H(e)$ is called the head node of the directed super-edge.

Definition: reachable to super-map A path of length $q$ in super-graph $H=(V,E)$, $Q_{st}$ is defined as the interleaved sequence of vertices and directed super-edges, where

$$s = T(e^1), \quad t \in H(e^q) \quad \text{and} \quad v_j \in H(e^j) \cap T(e^{j+1}), \quad j = 2, \ldots, q.$$ 

In path $Q_{st}$, $s$ and $t$ are respectively referred to as the start and end points of the path, and $s$ to $t$ are said to be reachable. In particular, if $t \in T(e^1)$, $Q_{st}$ is treated as a circle.

Definition: Adjacency matrix of directed hypergraphs The directed hypergraph is $R$, $V$ is the set of vertices, and $E$ is the set of directed super-edges. Adjacency matrix

$$A = \begin{pmatrix} a_{i1} & \ldots & a_{im} \\ \vdots & \ddots & \vdots \\ a_{ni} & \ldots & a_{nm} \end{pmatrix}, \quad a[i, j] \in \{0, 1, 2, \ldots\}$$

represents the number of directed super-edges from the start point $v_i$ to the end point $v_j$, where $v_i$ and $v_j$ are the vertices of $H=(V,E)$.

The size of the directed hypergraph is defined as the cardinal sum of the super-edges, and the rank (lower rank) is also defined as the maximum (small) value of the base of each super-edge. The adjacency matrix of the directed hypergraph and other matrix methods, due to the nature of the cluster existing in the hypergraph itself, it is difficult to establish a one-to-one correspondence between the directed hypergraph and its adjacency matrix. We can achieve the correspondence of the directed hypergraph to the adjacency matrix.
6. Project features and innovations
At present, the most common is the use of asymmetric key technology.

(1) Using Mahout and MLib to construct offline, near-line and real-time three-layer recommendation structure, realize flexible big data processing capability through Spark and Hadoop technology, and form a tag data source to better organize resource model by constructing discipline classification tree. And the user model, using a hybrid recommendation algorithm based on content and collaborative filtering, solves the cold start problem and improves the accuracy of the algorithm.

(2) Based on Hadoop distributed file system and MongoDB technology, the secure and efficient storage of big data is realized. The real-time data model update is realized by combining SparkStream technology, which improves the effectiveness of personalized recommendation.

(3) Analysis method of association rules based on the idea of co-evolution. The improved genetic algorithm and particle swarm optimization algorithm are used to iterate simultaneously for the two populations to make up for the shortcomings of the premature convergence and slow convergence of the genetic algorithm, and effectively search for high-quality association rules.

7. Conclusion
This topic focuses on the construction of the personalized education resource recommendation system and the key technology research under the big data environment. Under the premise of combing the existing research results, based on the practical development needs, the personalized education resource recommendation system from multiple perspectives The application of the construction, combined with data preprocessing technology, personality engine recommendation, data mining technology and other issues and countermeasures. A personalized recommendation system architecture that can handle PB-level educational resources is designed. The data mining technology for big data is used as the research object, aiming at data analysis technology, data extraction technology and extension data mining in the era of big data. The technology lays the theoretical foundation and finally builds an educational resource personality recommendation system based on extension data mining technology.

Acknowledgement
[Fund Project 1] This article is the Guangdong Provincial Department of Education's Guangdong Province Characteristic Innovation Project (Humanities and Social Sciences) -- "Research on the Innovative Talents Training Model under the Background of New Engineering" (2018GWTSCX054), phased research results;

[Fund Project 2] This paper is a phased study of the Cloud Computing and Big Data Professional Committee of the Higher Education Institute of Guangdong Province, "The Research on the Construction of Big Data Major and the Training Mode of Innovative Talents in Higher Vocational Colleges under the Background of New Engineering (GDYJSKT19-13)" Results.

[Fund Project 3] This paper is the mid-stage research result of the project of “Big Data and Intelligent Computing Innovation Research Team (NY-2019CQTD-02)”, the third batch of innovative strong school project of Guangzhou Nanyang Polytechnic College in 2019.

[Fund Project 4] This paper is the phased research result of the professional construction project of Computer Application Technology (NY-2019CQZY-04), the third batch of innovative strong school engineering brand in Guangzhou Nanyang Polytechnic College.

Author
Cheng Yunli (1980-), female, master, Communist Party member, associate professor, senior engineer, research direction: computer application technology, digital campus.

References
[1] Personalized adaptive online learning analysis model based on big data and its implementation[J]. Jiang Qiang, Zhao Wei, Wang Pengjiao, Wang Liping. China Electro-education Education.
2018(01)

[2] Access to education big data: acquisition and sharing of learning experience data based on xAPI specification [J]. Gu Xiaoqing, Zheng Longwei, Jian Jing. Modern Distance Education Research. 2017(05)

[3] Personalized learning evaluation model and system design based on educational big data in e-books[J]. Journal of Distance Education. 2019(05)

[4] Learning Dashboard: A New Learning Support Tool in the Age of Big Data[J]. Zhang Zhenhong, Liu Wen, Han Zhi. Modern Distance Education Research. 2017(03)

[5] Education big data in e-books and its application [J]. Zhang Yi, Qi Zhijia. Technology and Publishing. 2018(05)

[6] A fully distributed crawler clustering method based on Kademlia [J]. HUANG Zhimin, ZENG Xuewen, CHEN Jun. Computer Science. 2018(03)

[7] The use of big data in basic education in the United States [J]. Teng Biao, Zhu Xiaoling. People's Education. 2017(01)

[8] Big Data and Wisdom Education [J]. Ke Qingchao. China Education Information. 2018(24)

[9] Analysis of open source recommendation system for big data [J]. Mi Kefei, Zhang Yong, Xing Chunxiao, Wei Xin. Computer and Digital Engineering. 2018(10)

[10] Personalized news recommendation with context trees. Florent G, Christos D, Boi F. Proceedings of the 7th ACM conference on Recommender systems. 2016