Design of University Archives Business Data Push System Based on Big Data Mining Technology

Zhongke Wang, Jun Li
Chengdu Technological University
Chengdu, Sichuan Province, 611730, China

***Abstract*—Aiming at the problems of low accuracy, recall, coverage and push efficiency of university archives business data, a university archives business data push system based on big data mining technology is designed. Firstly, the overall architecture and topological structure of the university archives business data push system are designed, and then the functional modules of the system are designed. Using big data mining technology to mine user behavior, modeling according to user behavior sequence, and designing a model to predict user behavior sequence based on hidden Markov model theory. Finally, the user behavior sequence is analyzed, and the factors such as user collaboration, similarity of user behavior sequence and data timeliness are comprehensively considered to push university archives business data for users. The experimental results show that the proposed method has high data push accuracy, recall, coverage and push efficiency, and can effectively push the required business data for users.**

***Keywords*—Big data mining technology; system design; business data pus; hidden markov model; similarity**

I. Introduction

With the rapid development and popularization of information technology, the student file management information system in colleges and universities mainly manages the information related to student files. Relying on the basic platform of campus network, it plays an increasingly important role in teaching and management applications. As a powerful tool, big data mining technology can help university archives departments to better manage and analyze massive data, and discover hidden laws and knowledge from it. The traditional student file management information system generally adopts the client/server architecture mode or browser/server architecture mode, and adopts the centralized management mode. All the information is stored in a database, and the scale is getting larger and larger in colleges and universities. Many colleges and universities have multiple campuses. This mode obviously cannot meet the requirements of improving the efficiency of college student file management. With the rapid development of Internet technology, the data produced and faced by Internet users are increasing, which makes people face the dilemma of "information ocean". Therefore, the recommendation system came into being and became the first choice to help users filter effective information from massive information [1], [2]. When the recommendation system is applied to the data push process of archives business in colleges and universities, users' experience is not good because of the sparse data of archives business, one-sided similarity calculation of users or items and poor real-time recommendation results [3]. Therefore, it is of great significance to study and design an effective data push system for university archives business. The design goal of the push system is to provide an efficient, flexible and intelligent data push platform to help university archives departments manage and utilize data better. The system analyzes and mines the archives business data of colleges and universities through big data mining technology, thus providing valuable information and insight. We will also customize the push content for each user according to individual needs to ensure that users can get the most relevant information in time.

Pan, H et al. [4] proposed a five-layer push system framework, which is divided from bottom to top: the perception layer, the filter layer, the sorting layer, the rule layer and the application layer. At the same time, context awareness technology is applied to the data push process to achieve the push of business data. This method has the problem of low accuracy of data push. Zheng, Y et al. [5] reduced the initial data set by using the characteristics of static Sky-line points. Then, according to the characteristics of the searcher moving in the obstacle space, a distance intersection model is constructed, and a pruning strategy is proposed by using the model and the attributes of the data object. According to the pruning strategy, the data objects that have no influence on the query results when the inquirer moves are filtered, so as to reduce redundant data and get the filtered candidate data set. Finally, according to the non-spatial attributes of data objects and the characteristics of mutual dominance, the events that affect the candidate data set are determined, and these events are used to refine the candidate data set, further reducing redundant calculations and obtaining the result set at the current moment. Jelodar, H et al. [6] established an automatic data recommendation system based on user interest classification. The system includes offline module and online module. The offline module processes and cleans the historical storage data, extracts the user's interest characteristics, and forms the user's interest database submodule; The online module uses the recommendation engine to cluster and correlate the information in the user interest database sub-module, and finds the book information similar to the user interest data, forming a preliminary recommendation set sub-module. This method takes a long time to push data. Moreover, the pushed data is too single, which leads to the problems of low push efficiency and low coverage. Panda et al. [7] proposed to realize effective access control of medical data with forward and backward confidentiality, and medical service providers stored patients' electronic medical records in the cloud to provide high-quality medical services. Attribute-based encryption is a promising encryption technology, which can realize fine-grained access control of outsourced encrypted data. In this paper, an attribute-based encryption scheme is proposed to support the update of access policy, and it also provides forward security, backward security and user revocation. Performance analysis shows that the scheme has high communication and computing ability, and is suitable for devices with limited resources, but the push accuracy of this method is low. Nour et al. [8] proposed the access control mechanism in the named data network, and the information center network was recently proposed as an important candidate for the future Internet architecture to solve the problems existing in the current Internet host-centric communication model based on TCP/IP. This paper provides a detailed and comprehensive investigation of access control mechanism in named data network. The access control in named data network is studied by comprehensive method. Firstly, the paradigm of information center network is summarized, the change from channel-based security to content-based security is described, and different encryption algorithms and security protocols in named data network are introduced. Then, we divide the existing access control mechanisms into two categories: access control based on encryption and access control independent of encryption. Each category is classified according to the working principle of access control. Finally, the experience and lessons of the existing access control mechanism are summarized, and the challenges of access control based on named data network are pointed out, and the future research direction is emphasized, but the push recall rate of this method is low. Zhang et al. [9] put forward the monitoring of industrial sewage outlet water pollution based on web crawler and remote sensing interpretation technology, and took Luanhe River basin as the experimental point, combined with web crawler and remote sensing interpretation technology, put forward a feasible and efficient method to obtain sewage outlet data. Grab industrial information and spatial location data on the Internet, and get the location of industrial sewage outlet through remote sensing image interpretation. The distribution of main industrial sewage flowing into the tributaries of Luanhe River basin is simulated. By comparing the results with the actual data collected during the field investigation, the accuracy and reliability of the developed method are verified, but the push coverage rate of this method is low.

In order to solve the problems in the above methods, a design method of university archives business data push system based on big data mining technology is proposed. This method designs the overall architecture and topological structure of the university archives business data push system, and analyzes the functional modules of the system. Using big data mining technology to mine user behavior, modeling according to user behavior sequence, and designing a model to predict user behavior sequence based on hidden Markov model theory. On this basis, the user behavior sequence is analyzed, and the factors such as user collaboration, similarity of user behavior sequence and data timeliness are comprehensively considered to push university archives business data for users. The research shows that the proposed method has high data push accuracy, recall, coverage and push efficiency, and can effectively push the required business data for users with good push effect.

II. System Design

A. Overall System Architecture

The client side of the university archives business data push system adopts the client server mode, namely C/S, and the system adopts the B/S (Browser/Server) mode on the browser side. In general, it follows the classic three-tier architecture, which includes the presentation layer, application layer, and data layer. The system architecture of the university archives business data push system is shown in Fig. 1.

As it can be seen from Fig. 1, that the presentation layer is mainly the university archives business data push mobile terminal and PC browser. The application layer, situated in the middle of the three-tier architecture, consists of the control layer, the business logic layer, and the basic service layer. At the bottom is the data layer, including the data access layer and the data storage layer.

The display layer is mainly Android mobile terminal and PC terminal browser, which respectively displays user login interface and administrator login interface. The user end views the file business by logging into the Android mobile APP. The implementation method mainly uses the Android Activity and Fragment drawing interface, and the implementation architecture mainly relies on the Android operating system. The PC terminal mainly manages the system by logging in to the WEB page, which is realized by JavaScript+HTML. The execution architecture consists of a browser and an operating system.

The application layer is the main part of the data push system, which includes three parts to complete different operations. The control layer includes the control of terminal access, which is mainly composed of parameter analysis and session management. The control layer analyzes the file business type, and hands over specific operations to the business logic layer for processing; The business logic layer includes the specific services of user management, data publisher management, data management, intelligent push management and other functional modules. The basic service layer includes resource caching, resource access control, web crawler and vectorization processing, as well as data processing services and data analysis services. The functional modules of the business logic layer are implemented based on the construction in the basic service layer. The basic services offered provide functional support for the aforementioned business operations. The mobile terminal is realized through MVP mode, and the PC terminal is realized through Spring+Spring MVC+MyBatis. The execution architecture is composed of browser and operating system.

The data layer is composed of two components: the data access layer and the data storage layer. It mainly stores user data, archive business publisher data, business data, comment data, etc. of the system. The operation of structured data in the system is completed by MySQL database. Unstructured data is stored in files by data preprocessing.

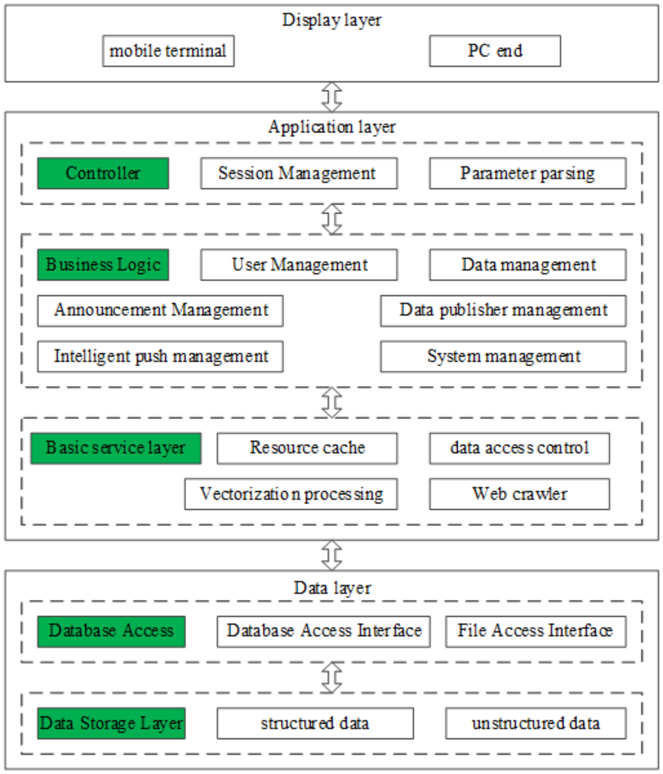

Fig. 1. Overall architecture of university archives business data push system.

In order to maintain the complete extensibility of the system, the system is divided into Controller layer, Business layer, Service layer and DAO layer when implementing the PC side in combination with the specific SSM framework. The business publisher logs in to the system through the browser, accesses the JSP page, and calls the service of the Service layer through the Controller layer.

The request sent by the business publisher to the server first enters the Controller layer. Through the controller layer's control of different businesses, dif-ferent businesses can be distributed to the Business layer to call specific business logic code. The Busi-ness layer encapsulates the database operation ser-vice and the interface provided by the Service layer to the outside. The database operation service can directly call the DAO encapsulated by Mapper. The specific SQL execution statement is implemented in XML. The results are returned to the browser for dis-play through the database operation service.

B. System Topology

Since the university archives business data push system is aimed at users in different geographical locations, the business publisher on the PC side also publishes business in different work areas [10], [11]. The work nodes are distributed in different locations, so the university archives business data push system designed by the proposed method adopts a star to-pology structure, as shown in Fig. 2.

The topological structure of this system is mainly divided into three parts: the network segment where the university archives business data push system is located, the network segment where the third-party organization is located, and the part that communi-cates directly with the Internet. The system is set up inside a communication, with separate WEB server and database server. The WEB server completes the request and response of PC end business publishers and mobile terminal users to the system functions. The database server stores the data information used in the system, including relational data and non-relational file data. At the same time, to ensure concurrent access to the system, the proposed method uses Alibaba Cloud's nginx load balancing to connect to multiple WEB servers. In this network segment, the three roles of system administrator, business administrator and communicator are divided to manage the business information and user information of the system. The roles in the system are connected to the system through the firewall. External users do not need to go through the firewall to access the system, which ensures the speed of external users' access and the security of the internal network.

Through this topology design, different access modes are set for various roles in the system.Whether tourists or registered users, authenticate users [12], [13], use mobile terminals to register and log in to the sys-tem through wireless settings, and complete corresponding queries and other operations. There are three types of file business publishers: correspondent, organization publisher, and individual publisher. The communicator is in the system network segment, the organization publisher is in the third-party organization Ethernet, and the individual publisher is free and can be directly connected to the Internet. The system administrator and news administrator are in the LAN of the system layout to manage user information and file business information.

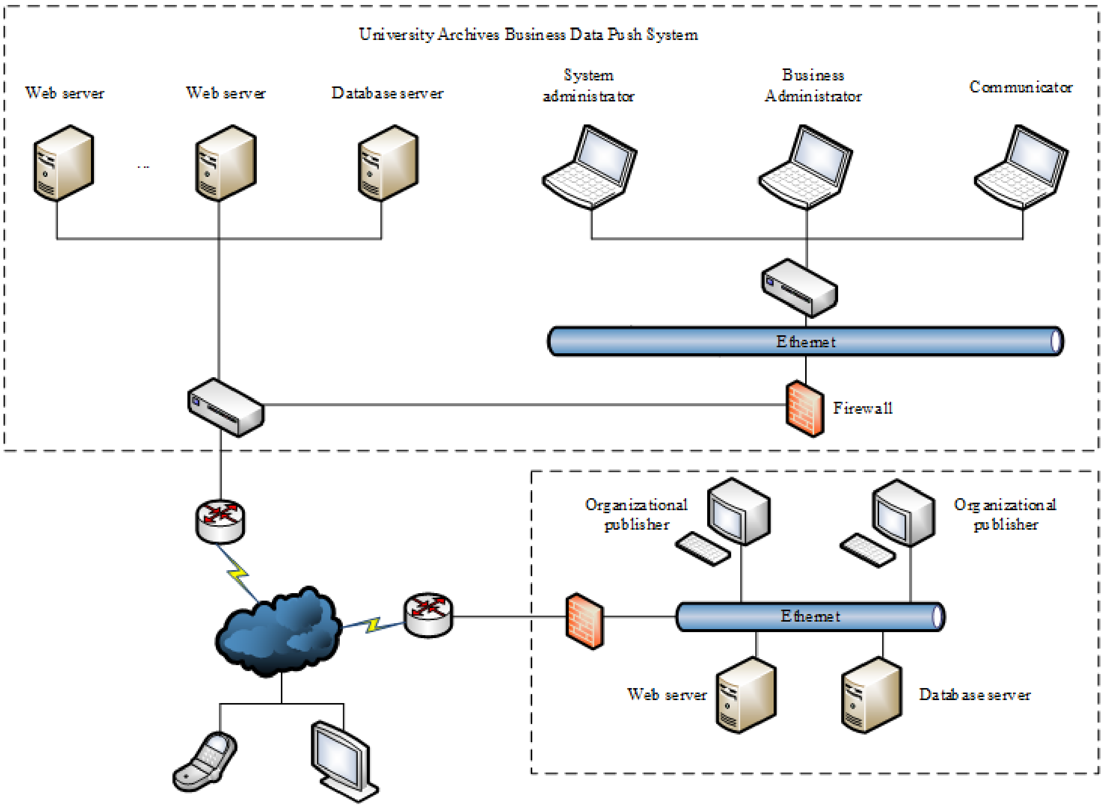

Fig. 2. System network topology.

C. System Function Module

According to the requirements of archives business in colleges and universities, the functional modules of archives business data push system in colleges and universities are divided, and the specific steps are as follows:

Step 1: Demand analysis: It deeply understand the specific needs of university archives business, including interviews with university archives managers, questionnaires or research on existing business processes, with the goal of clarifying the specific requirements and objectives that the system needs to meet.

Step 2: Function identification: Based on the demand analysis, identify the core functions required by the data push system of university archives business to meet the daily operation and management needs of university archives business.

Step 3: Module division: It group and modularize the identified functions. The division of modules should follow the principle of high cohesion and low coupling, so as to ensure that the functions of each module are relatively independent but can work together. The function of the data push system of university archives business is shown in Fig. 3.

1) Push management: You can query the push task records of push users and the records of accessing push pages; Users can also manually add push accounts to the policy server; At the same time, the user can also query the push task record of the active push customer and the record of accessing the push page [14].

2) White list management: This module is mainly responsible for adding users to the white list if they do not want to receive page push reminders, and then adding and deleting the white list if they do not want to be pushed in the future.

3) Customized push management: This module is responsible for adding and modifying the customized push task of the recommendation file business, and is responsible for performing association rule function analysis [15], [16] in the background, and applying the results to the foreground to batch add or delete users of customized recommendation products, and formulating push rules.

4) Push statistics: This module is mainly used to make statistics and analysis on the business data records of each push file in the push system, and present the results to the administrator in the form of a summary table.

5) System management: The system management function module primarily encompasses various functions such as system department management, user management, authority management, password modification, role management, and log query.

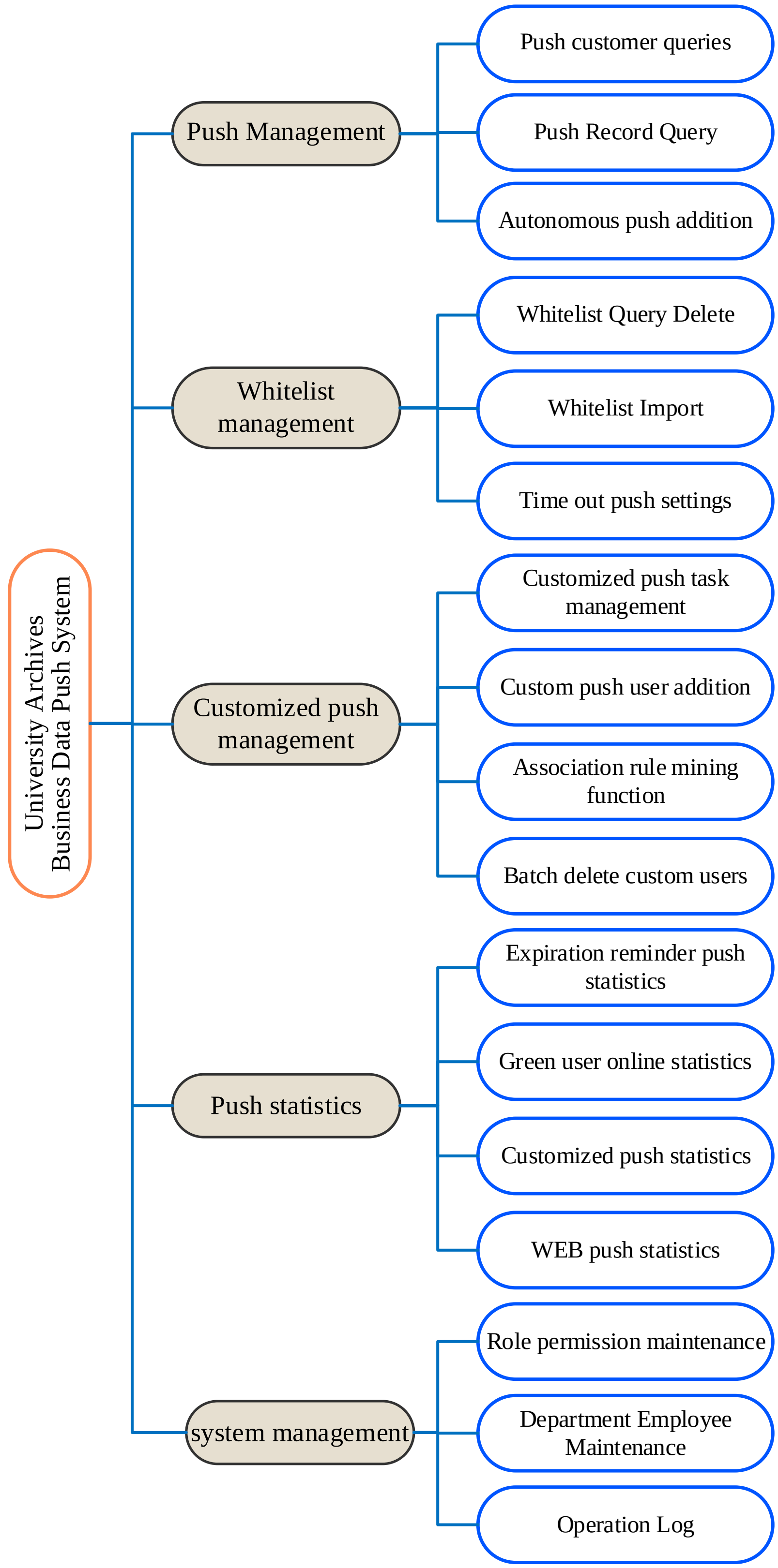

Fig. 3. Function division of college archives business data push system.

III. The Method and Implementation of College Archives Business Data Push

A. User Behavior Analysis

Considering that each user behavior sequence may represent a certain "interest" of the user, and the "entity" of the "interest" is the university file business page pointed to by the behavior sequence. Therefore, user behavior templates are used to record various "interests" of users and their corresponding behavior sequences. The generation rule is to take the user identifier as the name of the file. Each line in the file records a behavior sequence and the identifier of the file business that the behavior sequence points to. This file can be generated synchronously in the process of user behavior sequence extraction.

B. Similarity Calculation

The university archives business data push system designed by the proposed method mainly adopts the idea of user collaborative filtering [17], [18]. The recommendation system based on collaborative filtering needs to find the collection of items or users' nearest neighbor points through similarity calculation, and then carry out the next recommendation work according to the relevant information of these nearest neighbor points. The search of recommended file business data is divided into the following two steps:

- Find the nearest neighbor user of the current user through user feature similarity calculation;
- From the behavior sequences of these neighboring users, we can calculate the similarity of user behavior sequences to find the file business data that the current user may be "interested" in.

1) User feature representation and user feature file: The main data input that the system relies on is the various "filtering and provocation" selected by the user when searching, so that users and data, users and users, and data and data in the system can be associated through user behavior sequences, which can reflect the characteristics of users or data. Therefore, the proposed method takes the user behavior sequence as the original data, uses the vector space model to represent the user, and converts the feature vector βexpressed as:

$$\beta = \{(g_1, w_1), (g_2, w_2), \cdots, (g_n, w_n)\} \tag{1}$$

Among them, w_i is the i feature items weight of g_i, indicating that the current user's behavior sequence or behavior template the appears times of g_i ; Characteristic item $g_i$indicates a behavior included in the user behavior sequence; n indicates the type of user behavior in the system.

In order to improve the efficiency of user collaborative search, the proposed method records the feature vectors of all historical users in the form of feature files. This file is a text file, each line of which is composed of the historical user feature vector exported from the user behavior template and the number of this behavior template.

2) User similarity calculation: The common vector-based similarity calculation methods in recommendation systems are as follows:

a) Cosine similarity: Cosine similarity is a commonly used method in information retrieval for measuring the similarity between two documents [19], [20]. It involves treating the feature vectors of the documents as vectors. The cosine similarity is determined by the angle between these two vectors, with a smaller angle indicating a higher degree of similarity, the more parallel they tend to be, and parallelism means they are completely similar. The calculation formula is as follows:

$$s_{im}(u_a, u_b) = \frac{\beta \times \sum_{i=1}^{n} w_{ai} \times w_{bi}}{\sqrt{\sum_{i=1}^{n} w_{ai}^2 \times \sum_{i=1}^{n} w_{bi}^2}} \tag{2}$$

Among them, u_a, u_b represent users eigenvector of a, b; w_{ai},w_{bi} respectively represent eigenvectors u_a,u_b of elements i.

The advantage of cosine similarity is that it is not affected by the size of vector module and the calculation is simple. However, when encountering high-dimensional sparse vectors, the accuracy of similarity calculation will decline.

b) Pearson correlation coefficient similarity: In cosine similarity calculation, two user feature vectors are regarded as two independent variables, but in reality, there may be some connection between them. Especially in the scoring vector based on "user item", there is a certain relationship between different users' ratings of items. Therefore, Pearson correlation coefficient is proposed to measure the approximation of two users [21]. The calculation formula is as follows:

$$s_{im}(u_a, u_b)' = \frac{\beta \times \sum_{s \in S_{ab}} (w_{as} - m_a) \times (w_{bs} - m_b)}{\sqrt{\sum_{s \in S_{ab}} (w_{as} - m_a)^2 \times \sum_{s \in S_{ab}} (w_{bs} - m_b)^2}} \tag{3}$$

Among them, S_{ab} indicates the user a, b subscript set of common features; s express an element in S_{ab} ; w_{as} , w_{bs} respectively represent eigenvectors u_a, u_b of elements s; m_a, m_b respectively represent the average weight of u_a, u_b.

Considering that in the process of college archives business data push, the feature vector of online users is extracted from a behavior sequence, and a behavior sequence contains fewer repeated behaviors, the obtained feature vector of online users is basically a binary vector. Therefore, the similarity calculated by Pearson correlation coefficient is basically equivalent to cosine similarity in effect, but its calculation process is obviously more complex than cosine similarity, so the proposed method uses the cosine value of the vector as the similarity of user characteristics.

3) Similarity calculation of behavior sequence: In general, the number of items found through the current user's nearest neighbor set is still quite considerable, and through previous analysis, it can be seen that each user's behavior sequence is associated with at least one university file, and different behavior sequences of the same user may reflect the user's more subtle interests and preferences. Therefore, the current user's interest preferences can be more accurately predicted under the direction of the adjacent user's behavior sequence.

An action sequence mainly contains two aspects of information: the relative order in which the user actions in the sequence occur and the value of the actions. Therefore, the similarity of two behavior sequences can be measured from two perspectives, namely, behavior similarity and behavior value similarity.

Define S a behavior sequence the behavior state string of B, which only contains the name of the behavior in B, without the value of the corresponding behavior. Meanwhile in S, the string representing a behavior name is regarded as a state, which is an atomic weight and is indivisible. The calculation formula of the behavioral sequence B_1, B_2 of the behavior similarity of $X_{simseq}(s_1, s_2)$ is as follows:

$$X_{simseq}(B_1, B_2) = \frac{max[L_{len}(B_1), len(B_2)]}{max[L_{len}(B_1, B_2)]} \times sim(u_a, u_b)' \quad (4)$$

Among them, B_1, B_2 represents a common subsequence; $L_{len}(\cdot)$indicates the number of states contained in the behavior state string; $max[\,L_{len}(B_1, B_2)]$ express the number of states s_1, s_2 contained in the maximum common subsequence of.

Values of the same behavior are comparable, and the relative order factors between states should also be included in the scope of value similarity. So, the behavior sequence B_1, B_2 value similarity of $S_{simvalue}(B_1, B_2)$ is based on the maximum common subsequence of s_1, s_2, the calculation formula is as follows:

$$S_{simvalue}(B_1, B_2) = \frac{max[(X_{simseq}(B_1, B_2)]}{max[S_{subseqcount}(B_1, B_2)]} \quad (5)$$

Among them, $S_{subseqcount}(B_1, B_2)$ express the number of the same public status values of B_1, B_2.

Finally, the behavior sequence B_1, B_2 the similarity of $S_{simb}(B_1, B_2)$ given by the linear combination of behavior similarity and value similarity, set parameters $z = \frac{\sigma \times L_{len}(s_1)}{max[L_{len}(s_1), L_{len}(s_2)]}$, the calculation formula of $S_{simb}(B_1, B_2)$ is as follows:

$$S_{simb}(B_1, B_2) = (1 - z)S_{simseq}(s_1, s_2) + S_{simvalue}(B_1, B_2) \quad (6)$$

Among them, $\sigma$is a constant between 0 and 1, used to adjust z. Because according to practical experience, the importance of value similarity and behavior similarity is related to the ratio of the number of two behavior states. Generally, the closer the number of behavior states is, the greater the component of value similarity.

C. User Behavior Prediction

The user behavior prediction of the proposed method is mainly based on a set of hidden Markov models [22], [23]. First, relevant machine learning methods are used offline to train the prediction model parameters of various behaviors from historical data; Then, according to the partial behaviors of current online users, a finite step Markov process prediction is performed to obtain a relatively complete sequence of user behaviors; Finally, this behavior sequence is used to guide the push of university archives business data.

1) Prediction model of user behavior sequence: The user behavior sequence prediction model of the proposed method is a model systemcomposed of multiple "single two-layer" mixed state hidden Markov models designed on the basis of hidden Markov model theory [24], in which each sub model is independent of each other and deals with the prediction task of different types of user behavior sequences.

The difference between the "single two-layer" mixed state hidden Markov model and the traditional hidden Markov model is that its state set$S$the contained state elements can be divided into two categories: ordinary state elements and double-layer state elements. The double-layer state is mainly introduced to simulate the double-layer condition in the "filter condition". The detailed description of the model is as follows:

a) Status and observations: In the model, the name of a behavior in the user behavior sequence is regarded as a state, and the behavior value is regarded as the output value of the state. Corresponding to the front search function page of the college file business data push system, the type of the "filter condition" tag is considered as the status, and the "filter condition" represented by the "filter condition" tag is the output value of the corresponding status.

b) Transition probability and initial state vector: In order to reduce the complexity of model calculation, when designing the transition probability matrix, the two-layer state is reduced to a common state, and the root state is used as its representative [25], [26]. From any state$i$transfer to a two-layer state$j$transition probability ofa_{ij}. It can be calculated by the following formula:

$$a_{ij} = \sum_{s \in S_j} p_{is} \times S_{simb}(B_1, B_2) \quad (7)$$

Among them, s is j substatus collection for an element in S_j ; p_{is} is status $i$transfer to status$j$sub state of probability of s. Similar to the transition probability matrix, in the initial state vector, the two-level state is reduced to a common state for processing.

c) Probability vector of observation value: For the double-layer state, the observed values are also divided into two layers. For the root state, all its sub states are regarded as its observed values [27], [28]. Therefore, the double-layer state i of k output probability of observation values $b_i(k)$ is:

$$b_i(k) = \sum_{i=1,k=1}^{m} v_{ik}(j) \times a_{ij} \quad (8)$$

Among them, $v_{ik}(j)$ indicates a sub state k of j the probability of output values.

2) Prediction of user behavior sequence: The prediction of user behavior sequence is to start from a given state, predict the subsequent states and the output values of the states in a limited step, and generate a more complete sequence of user behavior [29]. According to the finite stage optimal decision theory of Markov process, the prediction process is to find the starting point from the specified state in the transfer matrix n step state transfer to obtain the maximum utility of the transfer path. Define the utility that can be obtained in each step of decision-making, and the predicted result of user behavior sequence is:

$$\gamma(i,j) = a_{ij} \times max[b_i(k)] \quad (9)$$

D. Historical Project Recommendation

Historical project recommendation is to push the university file business data related to historical users to current users. The recommended method is top-N. First, the recommended item set is found by combining user collaborative filtering and user behavior sequence search[30]; Then use the relevant sorting strategy to sort the data in the file business data set according to its sorting weight; Finally select N business data is pushed to users.

1) Historical item search: The search of historical items is to use the similarity between the characteristics of the current user and the historical user, and the similarity between the behavior sequence of the historical user and the behavior sequence of the current user as a clue to filter out the item set that may meet the interests of the current user from the historical items. The specific process is divided into the following steps:

a) User behavior prediction: Use the prediction model to predict backwards n step according to the behavior sequence entered by the current user, to get a more complete sequence of user behaviors s. The size of n is related to the number of behavior states entered by the current user, the more states there are, the small the n is. The maximum number of states contained in a user behavior sequence in the system is ϑ, then the calculation formula of n is as follows:

$$n = \begin{cases} \vartheta - \gamma(i,j) & \vartheta > \gamma(i,j) \\ 0 & \vartheta \le \gamma(i,j) \end{cases} \quad (10)$$

b) Finding user behavior template based on user feature similarity: According to the current user behavior sequence s, generate the user's feature vector, calculate the cosine similarity between the historical user and the current user's feature vector, and take out the similarity greater than the threshold $\sigma_1$user behavior template file.$\sigma_1$can be adjusted by the experimental feedback data.

2) Ranking of recommended items: The ranking of recommended items comprehensively considers factors such as feature similarity between users, behavior sequence similarity, popularity of business data and timeliness of business data, calculates the ranking weight of business data, sorts it, and finally generates N list of recommended items of file business data. The specific process is divided into the following steps:

a) Calculate the total feature similarity of users associated with data: One file may be related to multiple users, so the overall feature similarity of file business data $simusers(c,S)$ is the user behavior sequence s similarity between the generated characteristics of the current user and those of all neighboring users containing the data $A_{sim}(c,u_i)$ which is calculated as follows:

$$S_{simusers}(c,S) = n \times S_{simusers}(c,S) \times A_{sim}(c,u_i) \quad (11)$$

b) Calculate the total similarity of the behavior sequence associated with the data: One file business may be related to multiple user behavior sequences, so the similarity of one file business behavior sequence $S_{simstrings}(s,Q)$ is the behavior sequence and prediction sequence associated with s the similarity $S_{simb}(s,q_i)$ which is calculated as follows:

$$S_{simstrings}(s,Q) = \sum_{q_i \in S} S_{simb}(s,q_i) \quad (12)$$

c) Calculate the popularity of data: Count the number of different users associated with each file business, and calculate the popularity of each data. The calculation method is as follows:

$$Z(i) = exp\{\, m / \sum_{j=1}^{m} \delta(i,j)\}^{-1} \quad (13)$$

where, $Z(i)$ indicates file business data i popularity, m is the total number of related users.

The values of function $\delta(i,j)$ is as follows:

$$\delta(i,j) = \begin{cases} 1 & m_i \in I_j \\ 0 & m_i \notin I_j \end{cases} \quad (14)$$

where, I_j represents a collection of items contained in a user behavior template.

d) Calculation data timeliness: The timeliness of data is determined by the online time and current time of archive business data. The greater the timeliness, the more innovative the data. The calculation method is as follows:

$$T(i) = exp[\,(t_c - t_i)/\phi] \times Z(i) \times \delta(i,j) \quad (15)$$

Among them, $T(i)$ represent data i timeliness; t_c represents the current time; t_i represent data i online time; ϕ represents a constant greater than 0.

e) Calculate the sorting weight of data: Based on the above factors, the ranking weight of a university's archive business data is:

$$W(i) = \xi_1 \times S_{simusers} + \xi_2 \times S_{simstrings} + \xi_3 \times Z(i) + \xi_4 \times T(i) \quad (16)$$

Among them, ξ_i is the influence factor of each factor, which is set through experimental effect feedback or artificial experience.

f)Generate recommendation list: According to the sorting weight of file business data, the data is sorted from the largest to the smallest, and is selected the first N to generate recommendation lists to return to the user.

IV. Experiment and Discussion

In order to verify the overall effectiveness of the design method of university archives business data push system based on big data mining technology, it is necessary to test it. Before the test, prepare three virtual machines and modify the corresponding host names to mini0l, min02, and mini03. Next, set their network mode to NAT and modify their IP addresses [31]. Then modify the hosts file of each machine to configure the mapping relationship between the host name and IP address. Close the firewall of each machine and restart the machine. Finally, configure ssh password free login between virtual machines. Now the Linux environment is ready.

Set five datasets, labeled as Dataset 1, Dataset 2, Dataset 3, Dataset 4, and Dataset 5, specifically, it includes the following contents:

Data set 1: Student file data: This data set contains students' personal information, such as name, gender, date of birth, home address, contact information, etc. It also contains academic information, such as enrollment date, major, course results, rewards and punishments, etc.

Data set 2: Staff file data: This data set contains personal information of staff, such as name, gender, date of birth, contact information, etc. It also contains professional information, such as position, employment date, education, work experience, etc.

Data set 3: School business data: This data set contains various business data of the school, such as school curriculum arrangement, examination arrangement, activity arrangement, etc.

Data set 4: User behavior data: This data set contains the behavior data of users (students and faculty) in the system, such as login times, pages visited, time spent on pages, links clicked, etc.

Data set 5: System log data: This data set contains the running logs of the system, such as error logs and operation logs.

Use the proposed method, reference [4] method, reference [5] method, and reference [6] method to conduct business data push test in the Linux environment, and use the accuracy rate $Precision$, recall rate Z_{Recall} , coverage rate C_{ov} as an evaluation indicator of the method. Accuracy can measure the accuracy of pushed data, recall can measure the integrity of pushed data, and coverage can measure the satisfaction of pushed data to users' needs.

Accuracy P_{re} indicates the ratio of the number of test sets contained in the recommended data list to the number of all recommended items. The calculation formula is as follows:

$$P_{re} = \frac{A}{B} \times 100\% \tag{17}$$

where, A indicates the number of correctly predicted samples, B indicates the number of recommended samples for all users obtained from the test set.

It can be seen from the analysis of Fig. 4 that when the data push test is carried out under the same test environment, the data push accuracy of the proposed method is the highest and relatively stable. The data push accuracy of the reference [4] method and reference [5] method fluctuates greatly, and the data push accuracy of reference [6] method is stable but relatively low overall.

Recall rate Z_{Recall} indicates the ratio between the number of user recommendations and the number of items that users have acted in the test set:

$$Z_{Recall} = \frac{A}{C} \times 100\% \tag{18}$$

where, C indicates the number of samples that all users in the test set have had historical behaviors.

According to the test in Fig. 5, compared with the test results of the method in reference [4], the method in reference [5] and the method in reference [6], the recall rate of the proposed method is higher, indicating that the proposed method has good recommendation ability in the field of large-scale business data push.

Coverage rate C_{ov} represents the ratio of the number of recommended items to the total number of items, and its expression is as follows:

$$C_{ov}=B/D\times 100\% \tag{19}$$

Where, Drepresents the number of samples in the entire dataset.

It can be seen from Fig. 6 that under different data sets, the data push coverage of the proposed method is more than 90%, which indicates that the proposed method includes many types of data pushed by users with full coverage, and the coverage of the reference [4] method, reference [5] method and reference [6] method is low, which indicates that when the above methods are used to launch data push, the pushed data is relatively simple and cannot meet the needs of users.

The time required for the proposed method, reference [4] method, reference [5] method and reference [6] method in the process of data push test is shown in Table I.

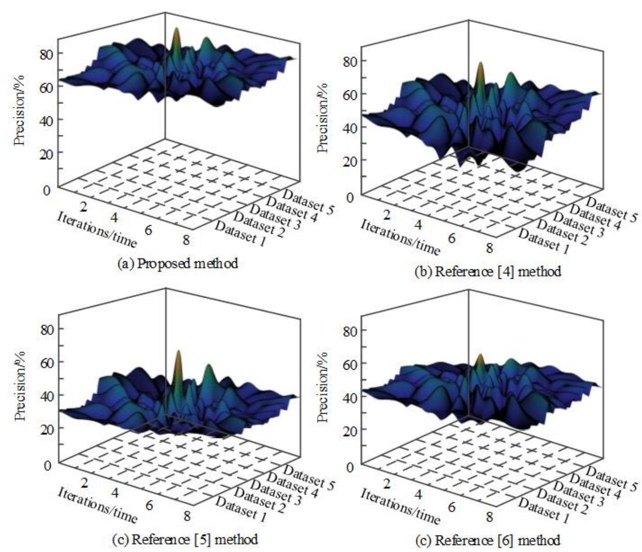

Fig. 4. Accuracy test results.

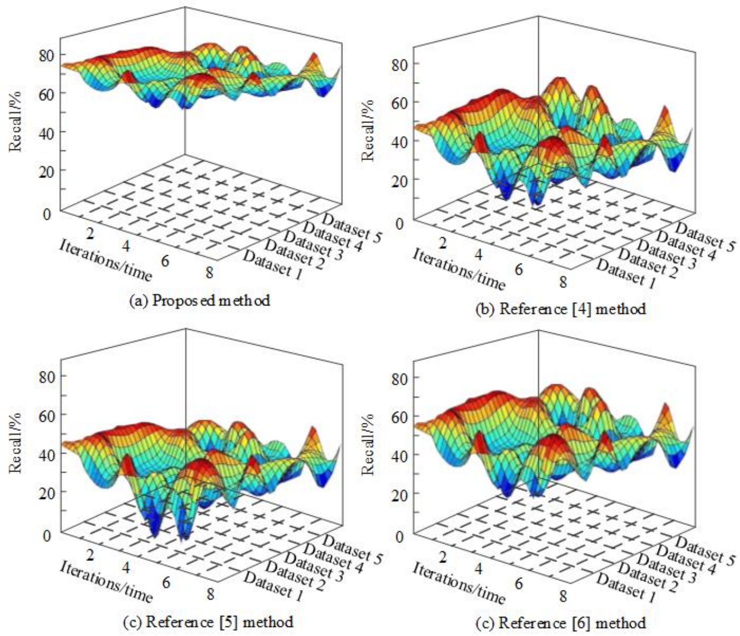

Fig. 5. Recall rate test results.

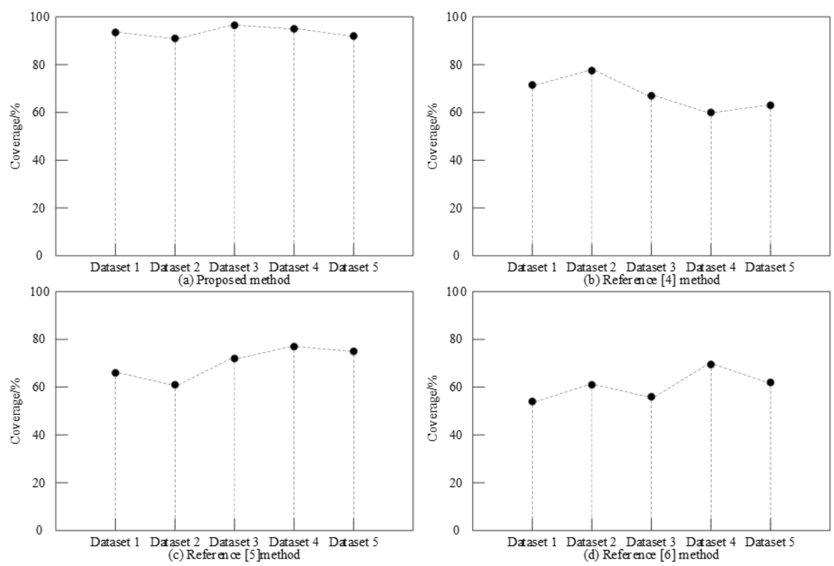

Fig. 6. Coverage test results.

TABLE I. DATA PUSH TIME OF DIFFERENT METHODS

Data samples/piece	Data push time/s			
	Proposed method	***Reference [4] Method***	***Reference [5] Method***	***Reference [6] Method***
100	5.2	9.8	7.6	8.8
200	5.8	10.5	8.3	9.6
300	6.3	11.3	8.9	10.5
400	6.9	12.2	9.7	11.2
500	7.4	13.0	10.4	12.7
600	7.7	14.9	11.7	13.5
700	8.2	15.4	12.6	14.4
800	8.6	16.6	13.7	15.6
900	9.1	17.3	14.9	16.8
1000	9.5	18.1	15.6	17.3

Analysis of the data in Table I shows that the da-ta push time of the proposed method, the reference [4] method, the reference [5] method and the refer-ence [6] method increases with the increase of the number of data samples, but under the same data samples, the push time of the proposed method is far lower than the other three methods, indicating that the proposed method has high data push efficiency.

To sum up, when the data push test is carried out in the same test environment, compared with the methods in [4], [5] and [6], the data push accuracy of the proposed method is the highest and relatively stable; The recall rate is high, the data push coverage rate is above 90%, and the push time required is much lower than the other three methods, which shows that the proposed method has high data push efficiency, and the data pushed for users contains more types and covers all the fields, and it has good push ability in the field of large-scale business data push.

V. CONCLUSION

On the one hand, it takes a lot of time to screen the data, and it is difficult to find the data you need from a large number of data; On the other hand, it will make a lot of redundant information become "hidden information" in the network, which cannot be obtained by ordinary users. In this context, a data push system is proposed. At present, there are some problems in the design method of data push system, such as low push accuracy, low recall, low coverage and low push efficiency. Therefore, a design method of university archives business data push system based on big data mining technology [31] is proposed, which is of great significance and provides decision-making and management support for university management departments. Through the data push function, timely and accurate data transmission and information sharing can be realized, and the efficiency and quality of university archives business can be improved. In this paper, the overall architecture and functional modules of the system are designed, and according to the characteristics of users' behavior, the push of university archives business is realized. It is verified that the data push accuracy of the proposed method is the highest and relatively stable. The recall rate is high, the data push coverage rate is above 90%, and the push time required is much shorter than the other three methods, which can effectively push the required business data for users in a short time. Provide more intelligent and scientific decision-making and management support for university management departments, and improve the efficiency and quality of university archives business.

The prospect of future research work can be carried out from the following aspects:

1) Dig deep into the archives business data of colleges and universities: In the future, we can further expand the data sources, including students, faculty, courses, scientific research projects and other dimensions, in order to obtain more comprehensive data information. Mining hidden rules and trends in data can help university management departments make better decisions and management.

2) Data security and privacy protection: In future research, we need to pay more attention to data security and privacy protection. Explore how to use emerging technologies such as blockchain to ensure the security and credibility of data, and study the methods and technologies of privacy protection to protect sensitive information of individuals and institutions.

COMPETING OF INTERESTS

The authors declare no competing of interests.

AUTHORSHIP CONTRIBUTION STATEMENT

Jun Li: Writing-Original draft preparation

Conceptualization,Supervision, Project administra-tion.

Zhongke Wang: Language review, Methodology, Software.

REFERENCES

[1] J. Yang, H. Chen, and X. Li, "Intelligent products' recommendation system based on machine learning algorithm combined with visual features extraction," Int J Biom, vol. 14, no. 2, pp. 125–137, 2022.

[2] B. Peng, "Research and Implementation of Electronic Commerce Intelligent Recommendation System Based on the Fuzzy Rough Set and Improved Cellular Algorithm," Math Probl Eng, vol. 2021, pp. 1–8, 2021.

[3] C. Liang, R. Fan, W. Lu, and S. Zhao, "Personalized recommendation based on CNN-LFM model," Computer simulation, vol. 37, no. 03, pp. 399–404, 2020.

[4] H. Pan and Z. Zhang, "Research on context-awareness mobile tourism e-commerce personalized recommendation model," J Signal Process Syst, vol. 93, pp. 147–154, 2021.

[5] Y. Zheng and D. X. Wang, "A survey of recommender systems with multi-objective optimization," Neurocomputing, vol. 474, pp. 141–153, 2022.

[6] H. Jelodar et al., "Recommendation system based on semantic scholar mining and topic modeling on conference publications," Soft comput, vol. 25, pp. 3675–3696, 2021.

[7] S. Panda, S. Mondal, R. Dewri, and A. K. Das, "Towards achieving efficient access control of medical data with both forward and backward secrecy," Comput Commun, vol. 189, pp. 36–52, 2022.

[8] B. Nour, H. Khelifi, R. Hussain, S. Mastorakis, and H. Moungla, "Access control mechanisms in named data networks: A comprehensive survey," Acm computing Surveys (cSuR), vol. 54, no. 3, pp. 1–35, 2021.

[9] J. Zhang, T. Zou, and Y. Lai, "Novel method for industrial sewage outfall detection: Water pollution monitoring based on web crawler and remote sensing interpretation techniques," J Clean Prod, vol. 312, p. 127640, 2021.

[10] S. D. Veeramachaneni, A. K. Pujari, V. Padmanabhan, and V. Kumar, "A hinge-loss based codebook transfer for cross-domain recommendation with non-overlapping data," Inf Syst, vol. 107, p. 102002, 2022.

[11] S. Soundrya, B. Kumaran, and V. Harini, "An Efficient Two-Layer Framework for Tour Sense Recommendation," ECS Trans, vol. 107, no. 1, p. 4913, 2022.

[12] A. Jabbari and J. B. Mohasefi, "A secure and LoRaWAN compatible user authentication protocol for critical applications in the IoT environment," IEEE Trans Industr Inform, vol. 18, no. 1, pp. 56–65, 2021.

[13] C. Hsu, L. Harn, and Z. Xia, "An HSS - based robust and lightweight multiple group authentication for ITS towards 5G," IET Intelligent Transport Systems, vol. 15, no. 11, pp. 1454-1460, 2021.

[14] U. Chaisoong, S. Tirakoat, and C. Jareanpon, "Tourist information-seeking behaviours using association rule mining," ICIC Express Letters, vol. 15, no. 9, pp. 915–923, 2021.

[15] Z. Zhao, Z. Jian, G. S. Gaba, R. Alroobaea, M. Masud, and S. Rubaiee, "An improved association rule mining algorithm for large data," Journal of Intelligent Systems, vol. 30, no. 1, pp. 750–762, 2021.

[16] Z. Chen, Y. Wang, S. Zhang, H. Zhong, and L. Chen, “Differentially private user-based collaborative filtering recommendation based on k-means clustering,” Expert Syst Appl, vol. 168, p. 114366, 2021.

[17] F. Wang, H. Zhu, G. Srivastava, S. Li, M. R. Khosravi, and L. Qi, “Robust collaborative filtering recommendation with user-item-trust records,” IEEE Trans Comput Soc Syst, vol. 9, no. 4, pp. 986–996, 2021.

[18] M. Verma and A. Rawal, “An enhanced item-based collaborative filtering approach for book recommender system design,” ECS Trans, vol. 107, no. 1, p. 15439, 2022.

[19] E. Türkarslan, J. Ye, M. Ünver, and M. Olgun, “Consistency fuzzy sets and a cosine similarity measure in fuzzy multiset setting and application to medical diagnosis,” Math Probl Eng, vol. 2021, pp. 1–9, 2021.

[20] B. Il Kwak, M. L. Han, and H. K. Kim, “Cosine similarity based anomaly detection methodology for the CAN bus,” Expert Syst Appl, vol. 166, p. 114066, 2021.

[21] I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua, “Prediction of solar energy guided by pearson correlation using machine learning,” Energy, vol. 224, p. 120109, 2021.

[22] T. Tamaoka et al., “Denoising electron holograms using the wavelet hidden Markov model for phase retrieval—Applications to the phase-shifting method,” AIP Adv, vol. 11, no. 2, 2021.

[23] Y. Aoudni et al., “Cloud security based attack detection using transductive learning integrated with Hidden Markov Model,” Pattern Recognit Lett, vol. 157, pp. 16–26, 2022.

[24] C. Raskar and S. Nema, “Metaheuristic enabled modified hidden Markov model for traffic flow prediction,” Computer Networks, vol. 206, p. 108780, 2022.

[25] S. Rahimipour, M. Ghatee, S. M. Hashemi, and A. Nickabadi, “A hybrid of neuro-fuzzy inference system and hidden Markov Model for activity-based mobility modeling of cellphone users,” Comput Commun, vol. 173, pp. 79–94, 2021.

[26] S. Sefati and N. J. Navimipour, “A qos-aware service composition mechanism in the internet of things using a hidden-markov-model-based optimization algorithm,” IEEE Internet Things J, vol. 8, no. 20, pp. 15620–15627, 2021.

[27] Rashid, H. K.,Farkhund, I.,Benjamin, C. M. Fung,J. B. Enabling Secure Trustworthiness Assessment and Privacy Protection in Integrating Data for Trading Person-Specific Information.IEEE Transactions on Engineering Management, vol. 68, no. 1, pp. 149-169, 2021.

[28] Anshu, J.,Vijay, A. Design of Novel Key Generation Technique Based RSA Algorithm for Efficient Data Encryption and Decryption.ECS transactions, vol. 107, no. 1, pp. 2585-2592, 2022.

[29] Vidya, R.,Prema, K.V. DEC-LADE: Dual elliptic curve-based lightweight authentication and data encryption scheme for resource constrained smart devices.IET wireless sensor systems, vol. 11, no. 2, pp. 91-109, 2021.

[30] Abdelhak, E.,Abderrahim, M.,Mohamed, O. Migrating Data Semantics From Relational Database Systems To NoSQL Systems To Improve Data Quality For Big Data Analytics Systems.ECS transactions, vol. 107, no. 1, pp. 19495-19503, 2022.

[31] Gunasekaran, P., Mohamed,S. B.,Ching-Hsien, S. N. K.,Revathi, S.,Priyan, M. K., Bala, A. M. FDM: Fuzzy-Optimized Data Management Technique for Improving Big Data Analytics.IEEE Transactions on Fuzzy Systems: A Publication of the IEEE Neural Networks Council, vol. 29, no. 1, pp. 177-185, 2021.