Research Article

Reflections on the Innovation of University Scientific Research Management in the Era of Big Data

Yiming Li (ID)

Adamson University (Philippines), Manila 0900, Philippines

Correspondence should be addressed to Yiming Li; smilelym1119@163.com

In China, universities are important centers for SR (scientific research) and innovation, and the quality of SR management has a significant impact on university innovation. The informatization of SR management is a critical component of university development in the big data environment. As a result, it is crucial to figure out how to improve SR management. As a result, this paper builds a four-tier B/W/D/C (Browser/Web/Database/Client) university SR management innovation information system based on big data technology and thoroughly examines the system's hardware and software configuration. The SVM-WNB (Support Vector Machine-Weighted NB) classification algorithm is proposed, and the improved algorithm runs in parallel on the Hadoop cloud computing platform, allowing the algorithm to process large amounts of data efficiently. The optimization strategy proposed in this paper can effectively optimize the execution of scientific big data applications according to a large number of simulation experiments and real-world multidata center environment experiments.

1. Introduction

With unprecedented power, information technology is promoting the continuous change of thinking mode and behavioral habits recognized in human daily production and life [1]. Big data technology finds out the relevance of data and extracts valuable information through the correlation analysis of data resources such as SR (scientific research) management system, financial system, personnel system, large-scale scientific literature database, and patent database based on the Internet, which can provide an extensive and scientific theoretical basis for traditional expert qualitative decision management [2, 3]. Because most of this knowledge comes directly from inside the database, it is less restricted and influenced by external resources, has relative independence, and has great guiding significance for SR decisions.

Since university informatization is implemented at the end of the last century, many universities have established and operated various database systems. However, the database systems existing among various departments are not related; thus, a number of information islands are formed in the whole school, which not only lead to a large amount of resources and a serious waste of funds in the school but also bring difficulties to the intensive management of teaching and reasonable study. The SR management system also has its own characteristics [4]. The truly idealized university SR management system is a platform that fully realizes the networking of university SR management, and based on it, forms a data center and management communication platform that can be updated instantly, provides comprehensive and accurate SR information for schools, provides a powerful reference for school leaders to make relevant SR decisions, provides convenient, quick, and thoughtful services for all school teachers to carry out SR activities, and provides great convenience for SR managers [5–8]. With the increasing dependence of SR decision-making system on various data, it is very urgent to actively apply big data technology in university SR management informatization; therefore, more attention should be paid

from all aspects, constantly explore and innovate, fully embody the role of big data, and provide assistance for the development of university SR management informatization.

First, A real-time SR business system is established in colleges and universities to improve SR departments' daily business efficiency. Second, an SR decision analysis system is created to ensure that a large amount of SR historical data and current operational data is effectively managed. Then, efficient algorithms are proposed to solve model problems, reduce data communication between data centers for scientific workflow tasks, and thus improve the execution performance of scientific big data applications.

The traditional data mining algorithm and its stand-alone environment have encountered performance bottlenecks, which can no longer satisfy the processing of massive data. Hadoop is a very simple distributed open-source computing platform, which has excellent advantages such as high efficiency, high reliability, high fault tolerance, and high scalability. Therefore, this paper has improved the NB classification algorithm, and based on the Hadoop cloud platform, it provides a parallel running environment for the data mining classification algorithm.

2. Related Work

Although managers can directly obtain clear information by using tools such as query, information such as relationships and trends hidden in a lot of data cannot be obtained from the data surface [9]. In literatures [10, 11], after analyzing the problems of the application method theory of the traditional management information system in the C/S mode, the idea of improvement is put forward. Literature [12] holds that the supervision of university research funds requires that each project team should have at least two supervisors. However, due to the great correlation between supervisors and research project personnel, supervision is difficult to achieve the expected effect. Literature [13] holds that there are some problems such as low utilization rate of fixed assets and insufficient protection of intangible assets.

Literature [14] holds that because there is no scientific budget accounting system at the national level, researchers and financial personnel have no calculation basis, so they can only calculate by experience, which leads to the incomplete and untrue content of budget preparation and cannot well reflect the research cost of the whole process of SR. Literatures [15, 16] hold that there are two problems in the performance of university SR projects: first, the lack of standards for performance evaluation of SR projects makes the overall performance evaluation unable to be effectively reflected; second, the lack of professionalism in SR content in the performance appraisal group leads to the lack of persuasiveness in performance appraisal results. Literature [17] holds that the person in charge of SR should establish the consciousness of budget management and attach importance to the role of budget in the management of SR funds. Literature [18] holds that internal control should be taken as the starting point to improve the management level of SR projects and the use efficiency of funds. Literature [19] holds that strengthening the performance evaluation of SR

funds should start from two aspects: the use benefit evaluation of SR funds and the benefit of SR work. Paying attention not only audits compliance, legality, rationality, and authenticity of the income and expenditure of SR projects but also comprehensively evaluates their economic benefits, technical benefits, and social benefits. In the aspect of performance evaluation of research funding, literature [20] holds that the evaluation method combining individuals and teams can be adopted, which should pay attention to both the cultivation of individual abilities and the collective performance appraisal so as to realize the circular development of individual interests and teams.

With the help of data mining technology, the university SR management innovation information system studied in this paper has actually solved the problems of multilevel characteristics and weight distribution of indicators, uncertainty of evaluation results, and deep mining [21] of evaluation data in the process of establishing the evaluation system, which makes the evaluation system more scientific, reasonable, and reliable.

3. Research Method

Construction objectives and system functions of the university SR management information system are put forward, and the overall design scheme of the system is designed. The NB classification algorithms are analyzed, and based on their advantages and disadvantages, improvements are made, and the SVM-WNB classification algorithm, a combination algorithm of SVM and WNB classification algorithm, is proposed. The related models and algorithms of data layout and task scheduling are researched, and a scientific big data application workflow management system is implemented.

3.1. Overall System Design Scheme. It combines B/S with C/S to form B/W/D/C/S (browser/web/database/client/server) architecture [22], and users can access the database locally and remotely at the same time.

A real-time business subsystem of SR adopts the B/S mode. Users of other departments and Internet users access the database server of the network center to obtain the news released to the outside world, and the two servers make a differential copy at a certain time period to achieve the consistency of data.

According to the setting scheme of B/W/D/C architecture, the development platform of the system can adopt the following scheme, as shown in Figure 1.

Windows Server 2003 seamlessly integrates network management with the fundamental operating system, making the network simple to use and manage. Its internal architecture is entirely 32 bit, and there are multiple threads running at the same time, allowing it to support more powerful applications. At the same time, the system's stability is ensured by providing separate memory space for the operating system and application programs to avoid data conflict.

The system's database is SQL Server 2000, a relational database product with many features such as data query
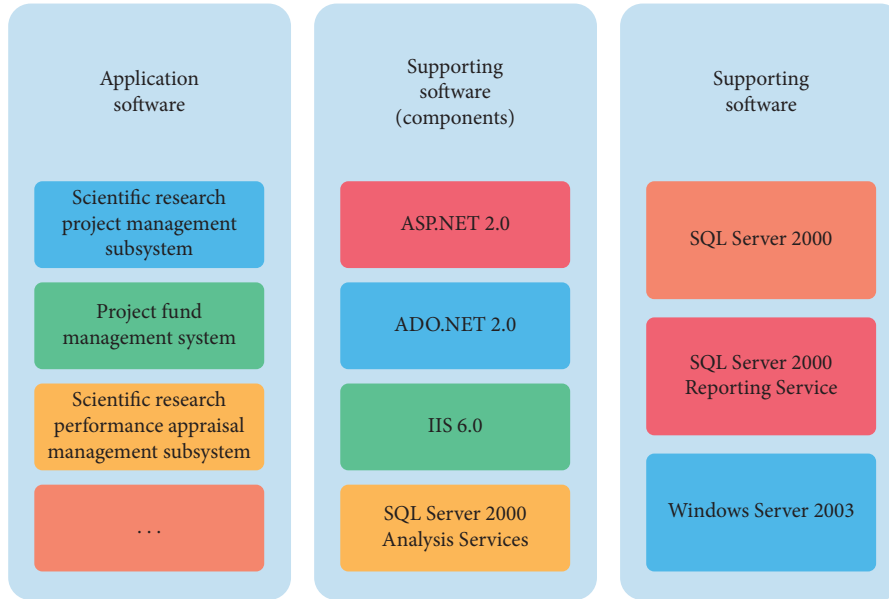
FIGURE 1: System development platform configuration.

diversity (SQL, XML), data integrity, high data access efficiency, concurrency control, transaction processing, data disaster protection, data diversification, security, and ease of use.

Our entire C/S mode SR decision analysis subsystem running platform is based on Visual Basic 6.0, which provides comprehensive component-based programming support, allows us to quickly build professional application systems for customers, reduces development and coding workload, and allows developers to focus more on communicating with users' needs.

3.2. Big Data Classification Algorithm.

The basic idea of the NB (Naive Bayes) classification algorithm is simple and easy to understand, mainly assuming that the attributes between classes are independent of each other and they do not influence each other. Assuming that there are n classes $C_1, C_2, \ldots, C_n$, a sample X to be classified is given, then if and only if

$$P(C_i|X) > P(C_j|X), \quad 1 \le j \le n, \ j \ne i. \tag{1}$$

At this time, the sample X belongs to C_i; that is, the NB classification algorithm assigns the sample X to be classified to the class with the highest posterior probability.

Because the NB classification algorithm assumes that the attributes of classes are independent of each other, the classifier model of the algorithm is very simple, as shown in Figure 2, which is the structure diagram of the NB classifier model.

The NB classifier model structure is a tree-like Bayesian network, which includes a root node representing class variables and some leaf nodes representing attributes, among which attributes are independent of each other.

The NB classification algorithm also has its disadvantages. If the training sample set or the sample set to be
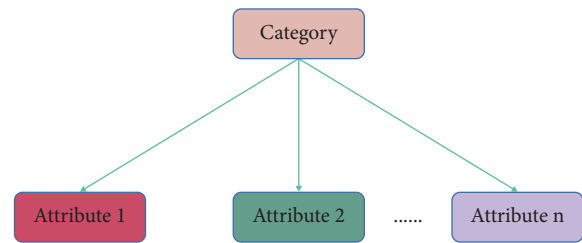

FIGURE 2: Structure diagram of NB classifier model.

classified has a very large scale, the cost of the NB classification algorithm will become very large.

The SVM (Support Vector Machine) algorithm often classifies samples incorrectly near the optimal classification hyperplane. For the NB classification algorithm, theoretically, the NB algorithm has very good efficiency and minimum error rate [23, 24], and the assumption of class condition independence reduces the computational cost of the algorithm. However, in practical application, such a hypothesis is impossible.

Based on the advantages and disadvantages of these two algorithms, the WNB (Weighted NB) classification algorithm can be obtained by improving the NB classification algorithm by weighting. Then, combining the two algorithms is considered to form a new algorithm, which is called the SVM-WNB (Support Vector Machine-Weighted NB) classification algorithm.

Assuming that w_k is the weight coefficient of attribute A_k, the WNB classification algorithm can be expressed by the following formula:

$$\underset{C_i \in C}{\operatorname{argmax}} \ P(C_i) \prod_{k=1}^{n} P(x_k|C_i)^{w_k}. \tag{2}$$

Here, two different methods are selected to calculate the weight coefficient w_k, and the average of w_{k1} and w_{k2} is

obtained, respectively, and then, this average is taken as the final value of w_k, and w_k is put into the above formula for calculation. The following is a description of these two ways.

Assuming that the conditional attribute and decision attribute are X, Y, their mathematical expectations are $E(X), E(Y)$, and covariance, respectively:

$$\text{Cov}(X, Y) = E(X, Y) - E(X)E(Y) = E[(X - E(X))(Y - E(Y))]. \tag{3}$$

The correlation coefficient ρ can be expressed as follows:

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{D(X)D(Y)}}. \tag{4}$$

The smaller the correlation coefficient ρ, the smaller the influence of condition attribute X on decision attribute Y. On the contrary, the larger the correlation coefficient ρ, the greater the influence of condition attribute X on decision attribute Y. Therefore, the correlation coefficient ρ can be used to weight the attributes, and w_{k1} can be expressed as follows:

$$w_{k1} = |\rho| = \left| \frac{\text{Cov}(X, Y)}{\sqrt{D(X)D(Y)}} \right|. \tag{5}$$

The second method is relatively simple; that is, the weight coefficient is obtained by calculating the correlation probability, assuming that there is an attribute A_k, and a_k is the value of the attribute A_k. The number of attributes whose value is a_k can be expressed as $\text{Num}(A_k = a_k)$. The attribute A_k takes the value of a_k, and the number belonging to class C_i can be expressed as $\text{Num}(A_k = a_k \wedge C_i)$. w_{k2} can be expressed as follows:

$$w_{k2} = \frac{\text{Num}(a_k \wedge C_i)}{\text{Num}(A_k = a_k)}. \tag{6}$$

Two different methods are used to select the weight coefficient, and finally, the average value is taken, which makes the selection of the weight coefficient more reasonable. Different attributes and different weight coefficients effectively improve the accuracy of classification.

3.3. Scientific Workflow Scheduling.

Scientific workflow scheduling is built on the foundation of scientific workflow and data center modeling, so the scheduling model must first model scientific workflow and data center and then consider the characteristics of scientific workflow scheduling and model workflow scheduling in a reasonable and accurate manner.

For the application of scientific big data, the DAG (Directed Acyclic Graph) model is usually used to represent the complex execution and data dependency among workflow tasks. In this study, the DAG model of scientific workflow is expressed as follows:

$$G = (V \leftarrow D \cup T, E, \omega, \theta), \tag{7}$$

where V represents the node set of the graph, including the dataset D and the task set T.

As shown in Figure 3, this study assumes that each task has an output dataset that can be used by many subsequent tasks.

At any level of coarsening l, in this paper, the optimized random connection weight matching method is used to aggregate the nodes of graph G_1. The nodes in the graph G_1 are randomly accessed. For each accessed node $v_i \in V^l$, if it has not been aggregated, the node v_j with the largest edge weight among the nodes connected to it and not aggregated is selected for aggregation.

If the edge connected with multiple nodes has the largest weight, the node with the best weight balance after aggregation is selected as follows:

$$f_{v_u} = (\tau(v_u) - \tau(k))^2, \tag{8}$$

where v_u is the vertex after aggregation, $\tau(v_u) = \omega_{cp}(v_u)/\sum_{v \in V}\omega_{cp}(v)/\omega_{st}(v_u)/\sum_{v \in V}\omega_{st}(v)$, $\tau(k) = cp_k/\sum_{k \in K}cp_k/st_k/\sum_{k \in K}st_k$. The smaller the q f_{v_u} value, the better the balance.

In order to make the hybrid GA (genetic algorithm) search only the feasible solution space, the fitness function set by this algorithm is as follows:

$$\text{fit}(I) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij} * \sum_{k=1}^{K}Z_{ijk}. \tag{9}$$

The general idea of the algorithm is based on the concept of revenue, which is the reduction of the total division cost when the nodes are moved between the two divided parts. The benefits can be positive or negative, and the positive benefits indicate that the cost of mobile node division is reduced; it is consistent with the purpose of the study, which is the hope of this study; and negative returns need to be avoided.

$(G_t, G_{t'})$ denotes the two disjoint partitions, namely, $G_t \cap G_{t_{,}} = \varnothing$. The symbol $\text{gain}(v_i, G_{t'})$ indicates the reduction of the division cost when the node $v_i \in G_{t'}$ moves to $G_{t'}$ gains; $\text{gain}(v_i, G_{t'})$ can be calculated by the following formula:

$$\text{gain}(v_i, G_{t'}) = g(v_i, G_{t'}) = g(v_i, G_t), \tag{10}$$

where $g(v_i, G_t)$ represents the sum of the weights of the edges connected with v_i in the partition G_t.

$v_i \in V_t$ iteration is moved to the appropriate data center according to the descending order of maximum profit until the partition meets the constraints of the corresponding data center.

4. Analysis and Discussion

4.1. Parallel Processing of Algorithm and Experimental Analysis.

It is feasible to parallelize the SVM algorithm on the Hadoop cloud computing platform. In the NB classification algorithm, when the dataset or attributes is too large, the cost of storing samples and calculating probability is very huge, and the classification performance will be significantly reduced. However, because the data processed by the NB algorithm are mutually independent [25], and the process of calculating the probability of data is also mutually independent, the data
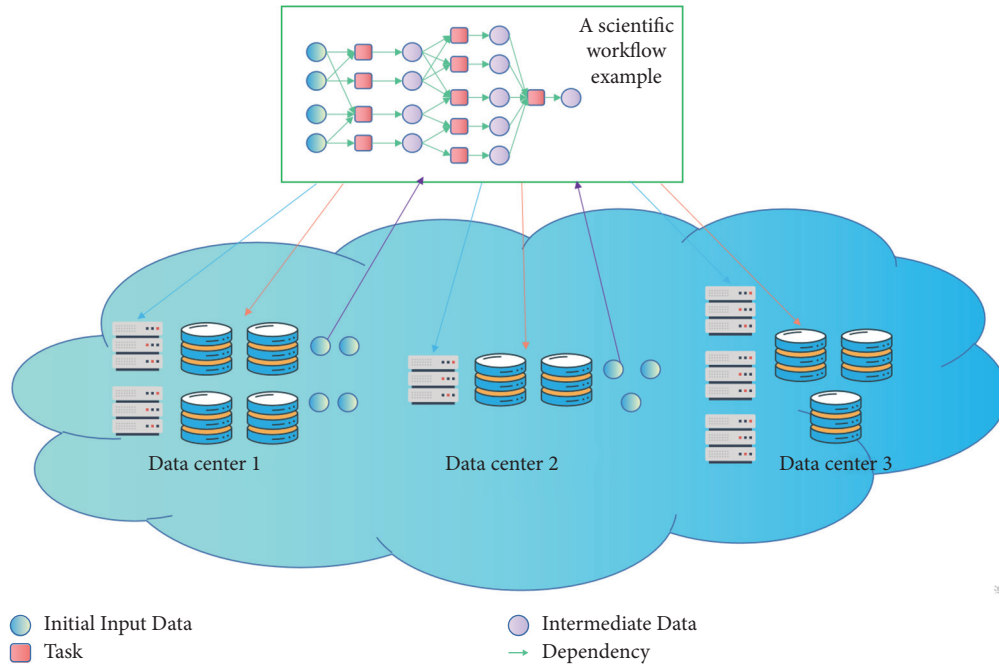
FIGURE 3: Schematic diagram of scientific workflow scheduling.

processed by the NB algorithm can be easily divided into balanced blocks. Through the above analysis, it is clear that the SVM-WNB classification algorithm is feasible.

In this experiment, the penalty parameter $C = 5$, the decision threshold $\xi = 0.6$, the radial basis kernel function parameter $g = 0.1$, and $K = 5$ in the nearest neighbor algorithm are selected, and mapred.map.tasks to 3 and mapred.reduce.tasks to 4 in Hadoop are set; that is, the number of maps is set to 3 and the number of reduce is set to 4.

For the multiclassification problem of the SVM algorithm, a one-to-one classification method is adopted (69), and $n(n-1)/2$ SVM classifiers are constructed for n categories.

The test sample set will be divided here in order to ensure the comprehensiveness and accuracy of the experimental results, and some samples will be randomly selected as experimental test data, and the final mean value will be taken as the result of many experiments.

The hSVM-WNB classification algorithm, NB algorithm, and SVM-WNB classification algorithm are compared after randomly selecting 10,000, 50,000, 100,000, 250,000, 500,000, 750,000, and 1 million test cases from the test sample set. The accuracy and processing time of these three algorithms are compared in Figures 4 and 5, respectively.

Figure 4 shows that the SVM-WNB classification algorithm is more accurate than the NB algorithm. On the one hand, it improves classification accuracy by optimizing the algorithm by weighting attributes in various ways; on the other hand, it selects and uses the optimized WNB algorithm to process near the classification hyperplane of the SVM algorithm. The hSVM-WNB classification algorithm maintains good accuracy even when the number of test cases is large due to parallel processing.

Figure 5 shows that the SVM-WNB classification algorithm requires two classifiers to be trained, whereas the NB algorithm is very fast, so the NB algorithm is faster than the SVM-WNB classification algorithm in terms of processing time.

4.2. Analysis of the University SR Management System.

The workflow used for testing in this experiment was generated by the Pegasus workflow generator. The ratio between the amount of resources provided and the amount of resources required in this group of experiments is 1 : 2.

For each workflow task, the required computing resources are uniformly distributed among (1, 10), and the unit is the number of CPU cores. The size of each data obeys the uniform distribution of (1, 100) in GB.

It can be seen from Figure 6 that among the four candidate algorithms for Montage workflow, the hybrid GA algorithm (hybrid GA) has the best scheduling effect, reducing the data transmission by 40.1% and 30.4% compared with the RRLocality and KCut algorithm. Because the KCut algorithm is designed based on the maximum flow minimum cut theory, all parts of the workflow after division are extremely unbalanced, which shows that some divided parts contain most workflow tasks and data nodes.

However, due to the limitation of computing and storage resources in the data center, it is necessary to adjust the partition that does not meet the constraint conditions, which leads to the deterioration of the results.

From Figure 7, it can be seen that with the increase of data centers, the amount of data transmission across data centers also shows an increasing trend. If the workflow is dispatched to more data centers, more dependent edges will
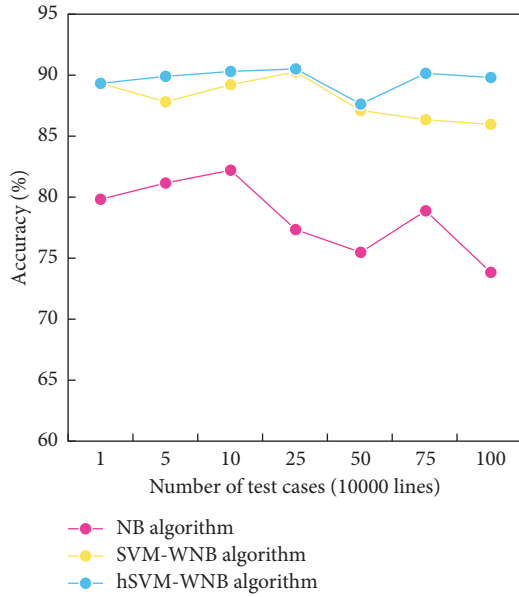
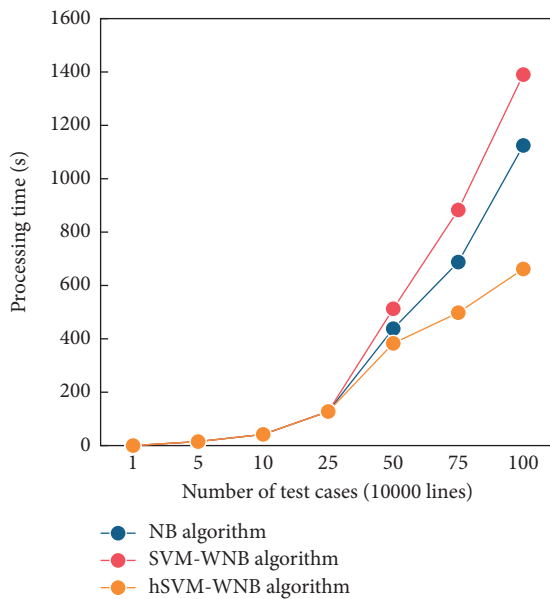FIGURE 4: Comparison chart of classification algorithm accuracy.

FIGURE 6: Influence of the number of workflow tasks on data transmission.

FIGURE 5: Comparison chart of classification processing time.
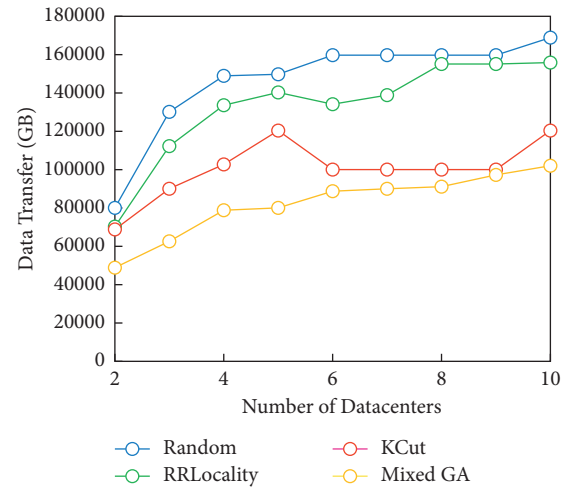
FIGURE 7: Influence of the number of data centers on data transmission.

be divided, which will lead to the increase of data transmission across data centers.

Figure 8 shows a Montage workflow with 500 tasks that was sent to four different data centers to investigate the impact of changing data center capacity on data transmission across data centers. The ratio between the amount of resources provided by the data center and the amount of resources required in this experiment is set to gradually increase from 1.2 to 1.5, with a 0.05 increment.

As shown in Figure 8, for Montage workflow, the hybrid GA algorithm is the best candidate algorithm. It is not difficult to find that with the increase of data center capacity, the data transmission volume of workflow across data centers shows a downward trend. This is due to the increase
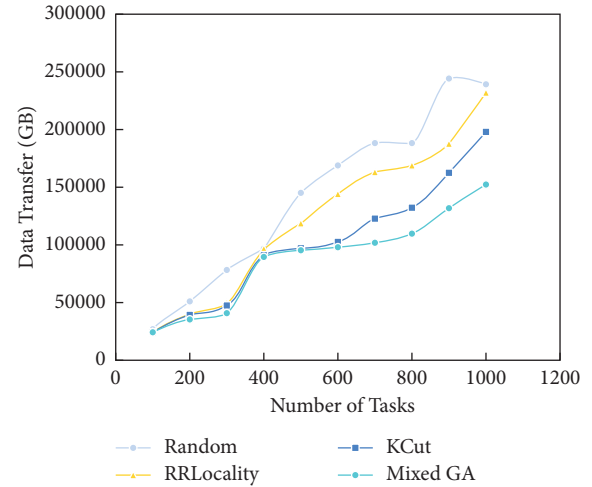
of data center capacity, which greatly reduces the tasks or data that are forced to be placed in nonoptimal data centers due to data center capacity limitations, thus reducing the amount of data transmitted across data centers.

This system middleware is based on Globus Toolkit-6.0, which consists primarily of the following components: grid security of multidata centers is handled by GSI and MyProx; data management is handled by 2GridFTP; task management is handled by GRAM5, and grid infrastructure is built by a group of C language common libraries.

After the system middleware has been installed, the workflow management software must be installed on the submission node in order to complete the workflow definition, mapping, and submission for execution. This section of the function is based on the Pegasus system, an open-source workflow management software that implements the data layout and workflow scheduling strategy proposed in this paper. The Pegasus workflow management system includes an implementation interface for defining and
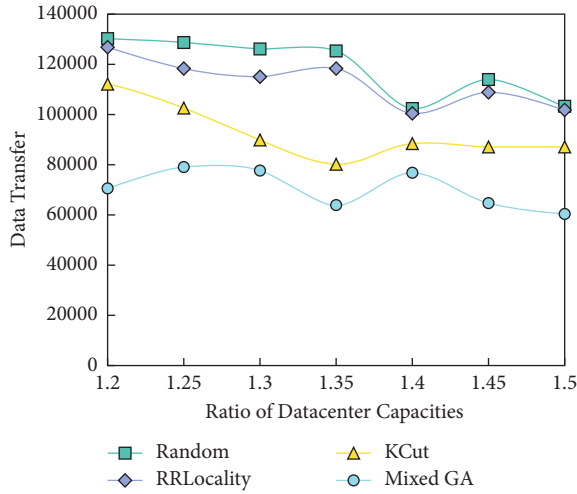
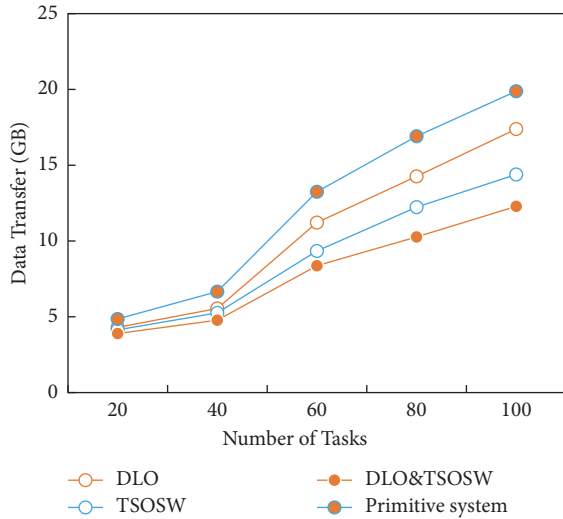Figure 8: Influence of data center capacity on data transmission.

Figure 9: Influence of different policies on data transmission amount of AMS workflow execution.

Figure 10: The effect of different policies on AMS workflow execution time.

analyzing workflows, as well as submitting them for execution in a multidata center environment.

In order to test the optimization efficiency of the scientific big data application implementation of the optimization strategy proposed in this article, this paper considers three optimization strategies: DLO (data layout optimization), TSOSW (task scheduling optimization of scientific workflow), and DLO and TSOSW (comprehensive optimization of two strategies), and the original system is compared.

The system execution environment consists of three data centers, namely, the cloud data center of a university, the dawning data center, and the data center built by several servers. The experimental data are AMS scientific data, which are distributed and stored in a multidata center environment. Due to the long running time of a large-scale scientifi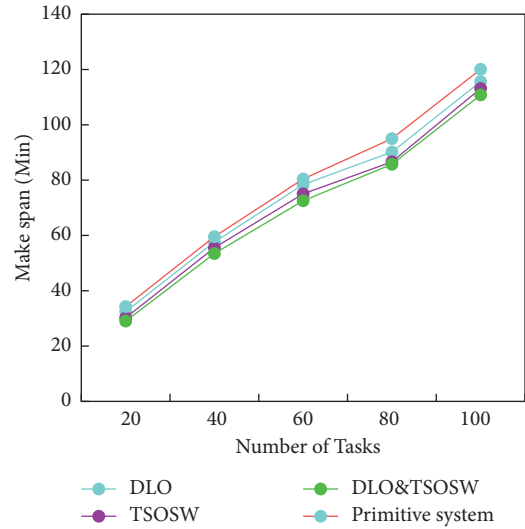c workflow and the huge data scale required, this paper uses small- and medium-scale scientific workflow to perform the test, and the number of workflow tasks ranges from 20 to 100.

Figures 9 and 10 are statistical charts of data transmission volume and corresponding execution time of AMS scientific workflows of different scales across data centers under different strategies.

Compared with the original system, DLO and TSOSW strategy reduces the data transmission by 34.1% on average and the execution time of workflow by 26.2%. Compared with TSOSW, the optimization effect of DLO is slightly worse. The reason may be that the execution position of the task has a greater influence on the data transmission. Only when the task is dispatched to the data center where the input data are located, the data transmission can be better optimized, and the execution efficiency can be improved.

5. Conclusion

SR is a source of national innovation and development, as well as a catalyst for scientific and technological progress and social progress. University SR management is a powerful guarantee for university SR's rapid development. In light of the rapid advancement of educational informatization, university SR management departments should introduce and utilize big data technology to provide an impetus for university SR's healthy and rapid development. This paper builds a comprehensive university SR management innovation information system using data warehouse and big data technology. The SVM-WNB algorithm is proposed, and attributes are weighted and improved. The hSVM-WNB algorithm was then transplanted to the Hadoop cloud computing platform for distributed processing. A series of experiments have shown that the parallelization of the new algorithm is possible, which could be very useful in practice. Using a special hybrid GA algorithm, a novel heuristic

method is proposed that can effectively reduce data transmission across data centers during workflow execution.

The front end of the system does not provide good data mining functions for users because the data mining model is not included in the background design and development of the system. One of the next major tasks is to add a data mining function to this system so that it can function as a comprehensive decision support system.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] C. Xin and C. He, "Research on university scientific research patent management information system based on BS mode," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 3, pp. 1–10, 2019.

[2] K. Benmoussa, M. Laaziri, S. Khoulji, M. L. Kerkeb, and A. E. Yamami, "Evaluating the usability of a Moroccan university research management web platform," *Procedia Manufacturing*, vol. 32, pp. 1008–1016, 2019.

[3] Y. Zhao, "Research on the construction of Chinese universities from intelligent library to smart library," *Management Science and Research: Chinese and English Version*, vol. 9, no. 1, p. 7, 2020.

[4] Q. Huang, "Design of information management system for scientific research projects based on time management," *Chinese Journal of Medical Research Management*, vol. 032, no. 3, pp. 220–225, 2019.

[5] S. Mao, W. Xu, Z. Ge, C. Jun, and X. Zhijie, "Design and implementation of an integrated information platform for emergency interconnection," *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue*, vol. 31, no. 7, pp. 884–889, 2019.

[6] G. Wei, "The influence of university teachers' scientific research pressure on scientific research performance, job satisfaction, mediation and adjustment of emotional intelligence," *International l Journal of Management Innovation Systems*, vol. 4, no. 1, p. 20, 2019.

[7] D. Dudziak-Gajowiak, G. Kolaczek, and K. Juszczyszyn, "Solving problems relating to ICT security management systems including various characteristics of the environment and system," *Scientific Journal of the Military University of Land Forces*, vol. 192, no. 2, pp. 321–334, 2019.

[8] N. S. Alekseeva, "Developing conceptual model of intellectual capital management system for innovation and industrial cluster in terms of digitalization," *Vestnik of Astrakhan State Technical University. Series: Economics*, vol. 2021, no. 1, pp. 26–31, 2021.

[9] O. Кочнева and O. Kochneva, "The knowledge management system in organizations as a tool of labor increase efficiency: stages of construction," *Bulletin of Kemerovo State University Series Political Sociological and Economic sciences*, vol. 2019, no. 1, pp. 14–23, 2019.

[10] M. A. Mumen, F. P. Oganda, N. Lutfiani, and I. Handayani, "Implementation of OJS based iJC media E-journal system at university of pramita Indonesia," *Aptisi Transactions on Management (ATM)*, vol. 4, no. 2, pp. 169–178, 2020.

[11] X. Dong, "Physical training information system of college sports based on big data mobile terminal," *Mobile Information Systems*, vol. 2021, no. 1, 7 pages, Article ID 4109794, 2021.

[12] A. N. Popsuyko, E. A. Batsina, G. V. Artamonova, and E. A. Morozova, "Labour productivity in medical institutions as an object of scientific analysis," *Vestnik of the Plekhanov Russian University of Economics*, vol. 18, no. 2, pp. 77–89, 2021.

[13] J. Morel, R. Crouzevialle, and A. Massoni, "Construction and management of a geo-historical information system for an interdisciplinary and contributory atlas: the historical atlas of limousin," *History and Computing*, vol. 14, no. 1-2, pp. 27–45, 2020.

[14] A. Rolgayzer, K. Demidenko, and A. Kadnikova, "Possibilities of applying information and communication technologies in teaching foreign languages at universities," *Scientific Research and Development. Socio-Humanitarian Research and Technology*, vol. 9, no. 3, pp. 49–57, 2020.

[15] J. Ma, D. Xu, and J. Xu, "Structuring China business credit management system and relevant characteristic variables based on the tree model," *International Journal of Information Systems and Supply Chain Management*, vol. 12, no. 2, pp. 22–35, 2019.

[16] M. H. Qasem, N. Obeid, A. Hudaib, A. M. Alamaiah, A. AI-Zahrani, and A. Al-Khasawneh, "Multi-agent system combined with distributed data mining for mutual collaboration classification," *IEEE Access*, vol. 9, p. 99, 2021.

[17] S. Kilicarslan, M. Celik, and Ş. Sahin, "Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification," *Biomedical Signal Processing and Control*, vol. 63, no. 1, Article ID 102231, 2021.

[18] J. Di and Y. Xu, "Decision tree improvement algorithm and its application," *International Core Journal of Engineering*, vol. 5, no. 9, pp. 151–158, 2019.

[19] D. Xiao, H. Li, and X. Sun, "Coal classification method based on improved local receptive field-based extreme learning machine algorithm and visible-infrared spectroscopy," *ACS Omega*, vol. 5, no. 40, pp. 25772–25784, 2020.

[20] M. Han, J. Ding, and J. Li, "PatHT: an efficient method of classification over evolving data streams," *The International arab Journal of Information Technology*, vol. 16, no. 6, pp. 1098–1105, 2019.

[21] Z. Huang, Y. Liu, C. Zhan, C. Lin, W. W. Cai, and Y. Chen, "A novel group recommendation model with two-stage deep learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, In press.

[22] Y. Cui, "Intelligent recommendation system based on mathematical modeling in personalized data mining," *Mathematical Problems in Engineering*, vol. 2021, no. 3, 11 pages, Article ID 6672036, 2021.

[23] N. Phuong, "An expectation maximization algorithm for LiDAR point cloud classification," *Environmental Science and Engineering:B*, vol. 009, no. 2, pp. 71–81, 2020.

[24] S. Arumugam, A. Kovalan, and A. E. Narayanan, "A novel educational data mining model using classification algorithm for evaluating students E-learning performance," *INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING*, vol. 7, no. 5, pp. 616–624, 2019.

[25] R. Sharma, S. N. Singh, and S. Khatri, "Data mining classification techniques-comparison for better accuracy in prediction of cardiovascular disease," *International Journal of Data Analysis Techniques and Strategies*, vol. 11, no. 4, p. 356, 2019.