Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH Open Access

Privacy-aware and Efficient Student Clustering for Sport Training with Hash in Cloud Environment

Guoyan Diao[1], Fang Liu[2], Zhikai Zuo[1] and Mohammad Kazem Moghimi[3*]

Abstract

With the wide adoption of health and sport concepts in human society, how to effectively analyze the personalized sports preferences of students based on past sports training records has become a crucial and emergent task with positive research significance. However, the past sports training records of students are often accumulated with time and stored in a central cloud platform and therefore, the data volume is too large to be processed with quick response. In addition, the past sports training records of students often contain certain sensitive information, which probably discloses partial user privacy if we cannot protect the data well. Considering these two challenges, a privacy-aware and efficient student clustering approach, named PESC is proposed, which is based on a hash technique and deployed on a central cloud platform connecting multiple local servers. Concretely, in the cloud platform, each student is firstly assigned an index based on the past sports training records stored in a local server, through a uniform hash mapping operation. Then similar students are clustered and registered in the cloud platform based on the students' respective sport indexes. At last, we infer the personalized sport preferences of each student based on their belonged clusters. To prove the feasibility of PESC, we provide a case study and a set of experiments deployed on a time-aware dataset.

Keywords: Cloud platform, User clustering, Privacy, Efficiency, Local server, Hash

Introduction

With the wide spread of COVID-19 all over the world, people are more focusing on health and life than ever before [1–3]. In this situation, health and sport concepts have gained wider adoption in whole human society than ever before [4–6]. To achieve the goal of healthy living and life, sport courses have been playing an increasingly important role in the whole education system. As a precondition of sport course setting and optimization, accurate recognition of personalized sport preference of each student is becoming a crucial and emergent task in front of education systems. Fortunately, past sports training records of students have provided a theoretically feasible evaluation basis to cluster the students and then infer their personalized sports preferences accordingly.

However, the past sports training records of students are often accumulated and stored in a central cloud platform for years and as a result, the data volume is often big enough [7], which probably leads to a long response time in student clustering and subsequent sports preference identification. In addition, past sports training records often contain some sensitive user information of students, which probably discloses student privacy if we cannot protect the data transmitted to the cloud platform well [8–10]. Considering these two challenges, a privacy-aware and efficient student clustering approach, named PESC is proposed for sport preference discovery and mining in cloud environment.

Concretely, in the proposed PESC approach, each student is firstly assigned a sport index (with less private

*Correspondence: k.moghimies@pgs.usb.ac.ir

[3] Department of Communication Engineering, University of Sistan and Baluchestan, Zahedan, Iran
Full list of author information is available at the end of the article

information) by analyzing his or her past sports training records registered in the cloud platform through hash projection operations. Secondly, similar students are clustered into different groups based on students' less-sensitive sport indexes recorded in the cloud platform (since index-based similarity calculation is rather quick, we can guarantee a high clustering efficiency in cloud). Third, we determine the personalized sport preferences of each student based on their belonged groups or clusters. At last, we demonstrate the effectiveness and efficiency of PESC through an intuitive example constructed from real world and a set of simulation experiments.

The major contributions of this article are detailed as below.

(1) Past sports training records as well as their respective time tags are used as a feasible and promising evaluation basis to infer the personalized sport item preferences of students in cloud-enabled smart education.
(2) Hash index mechanism is recruited here to cluster the students into different groups based on the past sports training records registered in the cloud platform. In concrete, firstly, we convert the sensitive user data into a less-sensitive user index through a kind of hash mapping process. Secondly, we use the less-sensitive user indexes to cluster users into different groups without disclosing much user privacy. This way, we can guarantee the user privacy is secure during the accurate user clustering process.
(3) We present a case study extracted from real world applicable scenarios to demonstrate the detailed steps of the proposed PESC solution. In addition, a set of simulated experiments are also provided to show the feasibility of the proposed PESC solution.

The rest of paper is briefly introduced as follows. Literature investigation is conducted in Related literature Section. A motivating example Section clarifies the research background and significance of this paper with a vivid example. Our solution: PESC Section introduces the detailed student clustering process as well as the student sport preference recognition process. A case study is presented in Case study Section and evaluation is presented in Evaluation Section. We summarize the paper and point out the possible research directions in Conclusions Section.

Related literature

In this section, we investigate the current research associated with sport clustering in cloud computing.

In [11], the authors mainly discuss how the socio-economic proximity between organizations or individuals affects the development of sports clusters. In order to solve this problem, the authors mainly investigate the sports groups of surfing and sailing, and puts forward a two-step model of cluster development. The purpose of [12] is to explore whether and how sports clusters correlate with community resilience across regions. To answer this question, the authors apply geographically weighted regression and visualization techniques to the macro data detection of community resilience. The results show that the community resilience of sports industry cluster is significantly correlated with the community resilience, and the two are strongly positively correlated. In [13], the purpose is to use the Sports eFANgelism scale to classify the fans in the process of The Korean league and analyze the differences between groups. Through cluster analysis, three groups were identified according to the level of league fans' behavior. At the same time, through variance analysis, each group had obvious differences in four kinds of preaching behaviors. In recent years, the holding of a large regional event is considered to have a potentially positive impact on the region's economy [14]. Concretely, the authors attempt to delve deeper into this field, focusing on the impact of participatory events on participants themselves. The authors use a cluster analysis procedure to perform a combinatorial analysis of participants, thus confirming and discussing the existence of various effects associated with participation events.

In [15], the authors mainly survey and give feedback to professional athletes. In this study, professional identity, sports identity and self-efficacy are measured, and cluster analysis is used to analyze the survey results. It is proved that identity and self-efficacy are important factors for athletes to choose dual career path. In [16], the aim was to replicate a controlled cluster experiment that demonstrated that behavioral skill training significantly improved motor behavior and self-efficacy in adolescents. Results from repeated trials show heterogeneity in the effectiveness of sports-based interventions, even among apparently similar populations. In [17], physical education has become an important course, and physical education has become an important teaching mode. This paper mainly studies the influence of physical education on college students' physical quality and physical activity level. In this study, the authors used a cluster randomized trial to verify. The results show that under the environment of limited control and self-evolution, appropriate physical education has a positive effect on the development of college students' physical quality. Nowadays, many young athletes are found to be taking performance-enhancing drugs [18]. In order to more accurately understand the causes of adolescent doping, this paper verified the Sports Drug Control model of Adolescent athletes (SDCM-AA), and modified the SDCM-AA model according to the experimental situation. The results show that there is a close relationship between the doping situation of

athletes and the control mode of sports drugs, the cluster effect and the value of athletes' own norms.

With the above literature review, we can simply conclude that existing literatures about population clustering based on sports information often fall short in time efficiency and privacy protection especially in the big data context. Considering this limitation, we propose a highly efficient student clustering approach with privacy protection, named PESC in this paper.

A motivating example

As shown in Fig. 1, three students (i.e., Lucy, Lily and John) as well as their past sport score records are presented and stored in local servers A, B and C, respectively. For example, Lucy took three kinds of sport items (i.e., volleyball, boating and skating) in the past and her volleyball score in 2020 is 80, boating score in 2021 is 90 and skating score in 2020 is 70. Likewise, the scores of the other two students are also presented in Fig. 1. For uniform data analysis and mining, the scores of students stored in different local servers A, B and C need to be transmitted to a remote cloud platform.

As mentioned in Fig. 1, different users' sport-time-score records are stored in different local servers and finally transmitted to a central cloud platform for uniform processing. During such a data transmission process, user privacy contained in the users' sport-time-score records are probably disclosed to other parties. With the known sport-time-score records of students in the cloud platform shown in Fig. 1, we can divide the students into different groups and infer their respective sport preferences for better sport training. However, two challenges often exist in the above student clustering and sport preference recognition process. First, too many students (although only three students are exampled in Fig. 1, the student volume is often big enough in practice) as well as their respective sport score records are involved in uniform data analysis in cloud platform, which often consumes much computational time and leads to a longer waiting time (specifically, the accumulated sport-time-score records of students are continuously growing with time elapsing, which often calls for more processing time). Second, the sport-time-score records of students often contain partial privacy of the involved students and therefore, they are often reluctant to publish their sensitive sport-time-score records to the public.

Motivated by the above two challenges, a privacy-aware and efficient student clustering approach, named PESC is proposed here for achieving uniform data analysis and mining in cloud environment. The concrete steps of PESC will be clarified with more details in the following sections.

Our solution: PESC

Next, we briefly introduce the major steps of the proposed PESC solution: firstly, we convert the sensitive user data into a less-sensitive user index through a kind of hash mapping process; secondly, we use the less-sensitive user indexes to cluster users into different groups without disclosing much user privacy; third, we determine the personalized sport preferences of each student based on their respective belonged groups or clusters. Concrete description of the PESC solution can be found in Fig. 2.

Step 1: Sport index assignment

In this step, we create and assign a sport index to each student based on his or her sport score records in the past registered in the cloud platform. Here, the index is created by a kind of hashing technique [19–21], whose reason is as follows: distributed data integration from mobile clients to cloud platform is often not secure

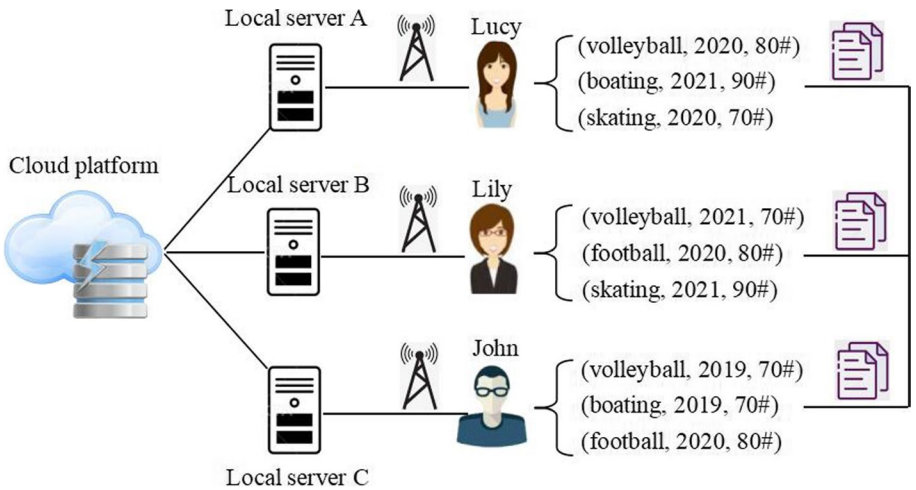

Fig. 1 An integration framework of sport-time-score records in cloud environment

Step 1: Sport index assignment. Through analyzing and mining the past sports training records of students registered in the cloud platform, we create and assign a sport index for each student. The sport index is an indicator of sport performances of a student.

Step 2: Student clustering. According to the sport index of each student recorded in the cloud platform, we cluster the students into different groups. In each group, all the students share the same or similar preferences of sports.

Step 3: Sport preference recognition. According to the clustering results of students, we recognize the personalized sport preferences of students to optimize the sport training course setting and configuration.

Fig. 2 Detailed procedure of PESC

[22–26] while hashing has been proven an effective data protection technology. Please note that the sport score records are time-aware as exampled in Fig. 1. Therefore, to smooth the subsequent sport index creation and assignment procedure, we first model the time-aware sport score records with the sport-time-score matrix M in Eqs. (1)-(2). Here, we assume there are n students, i.e., $Stu\_set = \{s_1, s_2, \ldots, s_n\}$, m kinds of sport items, i.e., $SI\_set = \{si_1, \ldots, si_m\}$ and l time slots, i.e., $T\_set = \{t_1, \ldots, t_l\}$. Thus, each column in matrix M in Eq. (1) denotes a student in Stu_set, each row of s_i in Eq. (2) indicates a time slot and each column of s_i in Eq. (2) represents a sport item. With the above formulation, we can find that each entry $q_{i,j,k}\,(1 \le i \le n, 1 \le j \le m, 1 \le k \le l)$ in Eq. (2) means student s_i's score of sport item si_j at time slot t_k, which has been exampled in Fig. 1.

$$M = (s_1, s_2, \ldots, s_n) \tag{1}$$

$$s_i = \begin{bmatrix} q_{i,1,1} & \cdots & q_{i,m,1} \\ \vdots & \ddots & \vdots \\ q_{i,1,l} & \cdots & q_{i,m,l} \end{bmatrix} \tag{2}$$

Specifically, if student s_i has not taken sport item si_j at time slot t_k, then the corresponding entry $q_{i,j,k}$ in Eq. (2) is equal to zero, i.e., $q_{i,j,k} = 0$ holds. Next, we create and assign a sport index Y_i for each student $s_i (1 \le i \le n)$ in matrix M. Concretely, we convert the matrix s_i in Eq. (2) into a corresponding vector $v(s_i)$ which is formalized in Eq. (3). Please note that as Eq. (3) indicates, $v(s_i)$ is an l*m-dimensional vector. For brief formalization, we assume $l * m = Q$ in the following discussions. Thus, $v(s_i)$ is a Q-dimensional vector.

$$v(s_i) = \left(q_{i,1,1}, \ldots, q_{i,m,1}, q_{i,1,2}, \ldots, q_{i,m,2}, \ldots, q_{i,1,l}, \ldots, q_{i,m,l}\right) \tag{3}$$

Then we randomly produce a new vector which is also Q-dimensional, denoted by $\phi = (\omega_1, \ldots, \omega_Q)$, following the rule in Eq. (4) where each ω_φ is randomly produced between -1 and 1. Next, with two Q-dimensional vectors $v(s_i)$ and ϕ, a product operation is adopted in Eq. (5), through which a real value y_i is obtained. Furthermore, a sign operation is adopted in Eq. (6) which converts the real value y_i into a Boolean one. The reason that we use binary mapping in Eq. (6) is that binary mapping is very time-efficient and effective [27–30], which has been validated and proved by many other literatures.

$$\omega_\varphi = \text{rand}(-1, 1)(1 \le \varphi \le Q) \tag{4}$$

$$y_i = v(s_i) * \Phi = \sum \left(q_\varphi * \omega_\varphi\right)(1 \le \varphi \le Q) \tag{5}$$

$$y_i = \begin{cases} 1 \text{ if } y_i > 0 \\ 0 \text{ if } y_i \le 0 \end{cases} \tag{6}$$

We repeat the above conversion process described in Eqs. (4)-(6) p times and then for each student $s_i\,(1 \le i \le n)$ in matrix M, we obtain p Boolean values: $y_{i(1)}, \ldots, y_{i(p)}$. Thus, a p-dimensional vector Y_i is obtained as formalized in Eq. (7). In this paper, Y_i is the sport index for student s_i and we can use Y_i to represent s_i for further calculation in the subsequent discussions in Step 2 and Step 3. Since Y_i is an index containing little sensitive information of student s_i, the following calculations associated with Y_i can be considered privacy-free. This way, the students' sensitive information contained in historical records can be protected well.

$$Y_i = \left(y_{i(1)}, \ldots, y_{i(p)}\right) \tag{7}$$

Step 2: Student clustering

Through Step 1, we have derived the student-index correspondence relationships, i.e., each student s_i is corresponding to an index Y_i. We summarize the above correspondence relationships, i.e., $s_i \rightarrow Y_i$ with a table shown in Table 1. Here, the hash table is recorded in the cloud platform.

Table 1 A hash table

Student	Index
s_1	$y_{1(1)}, \ldots, y_{1(p)}$
s_2	$y_{2(1)}, \ldots, y_{2(p)}$
$\cdots$	$\cdots$
s_n	$y_{n(1)}, \ldots, y_{n(p)}$

Table 2 R hash tables

Student	$Index_1$	$Index_2$	$\cdots$	$Index_R$
s_1	$y_{1(1)1}, \ldots, y_{1(p)1}$	$y_{1(1)2}, \ldots, y_{1(p)2}$	$\cdots$	$y_{1(1)R}, \ldots, y_{1(p)R}$
s_2	$y_{2(1)1}, \ldots, y_{2(p)1}$	$y_{2(1)2}, \ldots, y_{2(p)2}$	$\cdots$	$y_{2(1)R}, \ldots, y_{2(p)R}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
s_n	$y_{n(1)1}, \ldots, y_{n(p)1}$	$y_{n(1)2}, \ldots, y_{n(p)2}$	$\cdots$	$y_{n(1)R}, \ldots, y_{n(p)R}$

To minimize the "false-negative" and "false-positive" probability in subsequent student clustering results, we create R hash tables instead of one table, which are presented in Table 2. Thus, each student s_i will be corresponding to R indexes: $Index_1, \ldots, Index_R$. Then with Table 2, we can cluster the n students in Stu_set into different groups $G = \{g_1, g_2, \ldots\}$, which is based on the judgment condition in Eqs. (8)-(10). Here, s_x, $s_z \in Stu\_set$. Please note that the time complexity of the judgment condition of Eqs. (8)-(10) is very low and therefore, it is typically suitable for the clustering scenarios in big data context since big data processing often involves high computational costs [31–35] and effective computing offloading capabilities [36–40].

s_x and s_z both belong to group g

$$\text{iff there exists } r(r \in [1, R]) \text{ satisfying } (Y_x)_r = (Y_z)_r \tag{8}$$

$$(Y_x)_r = y_{x(1)r}, \ldots, y_{x(p)r} \tag{9}$$

$$(Y_z)_r = y_{z(1)r}, \ldots, y_{z(p)r} \tag{10}$$

Step 3: Sport preference recognition

For two students s_x and $s_z (\in Stu\_set)$, if they are divided into an identical group $g \in G$, then they are probably with similar sport preferences. In this situation, if s_z likes a sport item $s_i \in SI\_set$, then s_x likes the sport item s_i with high probability, vice versa. This way, we can infer the sport preferences of each student in universities since people belonging to same group often share the same or close preferences with high probability [41–44].

Here, please note that if there are no other students belonging to the group g that contains student s_x, then an exception occurs since s_x has no similar students. In this situation, we loosen the similar student judgment conditions in Eqs. (8)-(10). Concretely, as Eq. (11) shows, if s_z has the minimal Hamming Distance with s_x in any of the R hash tables, then s_z is taken as the similar student of s_x and thus, if s_z likes a sport item $s_i \in SI\_set$, then s_x likes the sport item s_i with high probability, vice versa.

s_x and s_z are similar

$$\text{iff } (Y_x)_r \oplus (Y_z)_r = \text{MIN } (Y_x)_r \oplus (Y_z)_r | r \in [1, R], s_z \in Stu\_set \tag{11}$$

In our proposal, each user is assigned an index with less privacy through a hash mapping process, which can be done in an offline manner since the indexes can be generated beforehand [45–47]. Therefore, the time complexity of this conversion process is approximately zero. Afterwards, user clustering can be performed based on an online user index matching process whose time complexity is approximately O(1). Therefore, the total time cost of our proposal is rather low.

Table 3 An example of PESC

Random vector	Hash table 1	Hash table 2
ϕ_1	(0.1, 0.4, -0.5, 0.3, -0.8, -0.6, 0.4, 0.3)	(-0.1, 0.5, -0.3, 0.5, -0.7, -0.6, 0.2, 0.2)
ϕ_2	(-0.7, 0.2, 0.4, -0.3, 0.5, 0.2, -0.9, -0.5)	(-0.8, 0.3, -0.5, 0.2, -0.9, -0.6, 0.4, 0.1)
ϕ_3	(-0.2, 0.8, -0.6, 0.7, -0.3, -0.9, 0.5, 0.2)	(0.6, 0.3, -0.7, 0.8, -0.1, -0.3, 0.7, 0.3)
$y_1 = v(s_1) * \phi$	(-2.4, -1.5, -0.9)	(-2.4, -8.4, 3.4)
$y_2 = v(s_2) * \phi$	(1.6, -7, 3.5)	(1.4, -4.1, 8.6)
$y_3 = v(s_3) * \phi$	(-2.4, -2.7, 1.5)	(-1.3, -6.1, 6.6)
y_1	(0, 0, 0)	**(0, 0, 1)**
y_2	(1, 0, 1)	(1, 0, 1)
y_3	(0, 0, 1)	**(0, 0, 1)**

Formally, our proposed PESC solution is described with Algorithm 1.

Require: (1) $Stu\_set = \{s_1, s_2, \ldots, s_n\}$
(2) $SI\_set = \{si_1, ..., si_m\}$
(3) $T\_set = \{t_1, ..., t_l\}$
(4) $q_{i,j,k}(1 \le i \le n, 1 \le j \le m, 1 \le k \le l)$
(5) s_x: a target student
Ensure: (1) si // sport items preferred by s_x
1: $Q = l * m$
2: **for** $\varphi = 1$ to Q **do**
3: $\omega_\varphi =$ rand (-1, 1)
4: **end for**
5: $\phi = (\omega_1, \ldots, \omega_Q)$
6: **for** $i = 1$ to n **do**
7: $v(s_i) = (q_{i,1,1}, \ldots, q_{i,m,1}, q_{i,1,2}, \ldots, q_{i,m,2}, \ldots, q_{i,1,l}, \ldots, q_{i,m,l})$
8: $y_i = v(s_i) * \phi$
9: **if** $y_i > 0$ **then**
10: $y_i = 1$
11: **else**
12: $y_i = 0$
13: **end if**
14: **end for**
15: Repeat Lines 2 $\sim$ 13 p times
16: **for** $i = 1$ to n **do**
17: $Y_i = (y_{i(1)}, \ldots, y_{i(p)})$
18: **end for**
19: generate a hash table // see Table 1
20: Repeat Lines 14 $\sim$ 18 R times
21: generate R hash tables // see Table 2
22: **for** $r = 1$ to R **do**
23: **for** $z = 1$ to n **do**
24: $(Y_x)_r = y_{x(1)r}, \ldots, y_{x(p)r}$
25: $(Y_z)_r = y_{z(1)r}, \ldots, y_{z(p)r}$
26: **if** $(Y_x)_r = (Y_z)_r$ **then**
27: s_x and s_z are similar
28: put s_z into set g
29: **end if**
30: **end for**
31: **end for**
32: **for** each $s_z \in g$ **do**
33: **for** $j = 1$ to m **do**
34: **if** s_z prefers si_j **then**
35: **return** si_j
36: **end if**
37: **end for**
38: **end for**

Algorithm 1 PESCCase study

A case study is constructed to show the concrete steps of the proposed PESC solution when we need to cluster students according to their historical sport records in cloud platform and infer their respective sport preferences. Here, we assume there are three students: s_1, s_2, s_3 and each student's historical sport preferences (totally four sport items) with time (totally two time slots) are presented as follows.

$$M = (s_1, s_2, s_3)$$

$$s_1 = \begin{bmatrix} 5 & 4 & 3 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

$$s_2 = \begin{bmatrix} 4 & 5 & 3 & 4 \\ 2 & 4 & 5 & 5 \end{bmatrix}$$

$$s_3 = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 5 & 3 & 4 & 3 \end{bmatrix}$$

Then we convert the three 2*4 matrices corresponding to s_1, s_2, s_3 into three 8-dimensional vectors, i.e., $v(s_1)$, $v(s_2)$, $v(s_3)$. Next, we randomly generate two hash tables, each of which is with three vectors ϕ_1, ϕ_2 and ϕ_3 (as shown in Table 3). The according to Eqs. (4)-(7), we can derive the indexes of the three students s_1, s_2, s_3, i.e., y_1, y_2, y_3 in the two hash tables, respectively. Since $y_1 = y_3$ holds in the second hash table (i.e., $y_1 = y_3 = (0, 0, 1)$), student s_3 is similar with s_1 with high probability. Therefore, the sport items preferred by s_3 is also preferred by s_1 with high probability. This way, we can achieve privacy-aware and efficient student clustering and sport preference inference for sport training in universities.

$$v(s_1) = (5, 4, 3, 2, 4, 3, 2, 1)$$
$$v(s_2) = (4, 5, 3, 4, 2, 4, 5, 5)$$
$$v(s_3) = (2, 3, 4, 5, 5, 3, 4, 3)$$

Evaluation

Through a time-aware quality performance dataset WS-DREAM, we design a set of experiments to show the effectiveness and efficiency of the PESC solution. We compare the performances (i.e., MAE for measuring clustering accuracy and time cost for measuring clustering efficiency) of PESC with another two existing solutions: UCF (user-based collaborative filtering) and ICF (item-based collaborative filtering). Concretely, we measure the performances of three solutions under different parameter settings. Here, two parameters are recruited: number of students varied from 100 to 300 and matrix density varied from 5% to 40%. Experiments are run in a laptop with 3.20 GHz CPU and 4.0 GB RAM.

Concrete experiment results are reported in the following four profiles.

(1)MAE comparison

We measure the clustering accuracy of three solutions via MAE performances [48–51] under different parameter settings. In concrete, the MAE of three solutions with respect to the number of students is presented in Fig. 3a. As Fig. 3a indicates, the MAE of ICF solution is the highest, which means that the clustering accuracy of ICF is often poor. On the contrary, PESC and UCF solutions both perform well in MAE, which indicates that their clustering accuracy is often high. Furthermore, the MAE value of PESC is close to that of UCF. This means that PESC can achieve approximate clustering accuracy with the baseline UCF solution, because the hash index technique adopted in PESC can guarantee to output the most

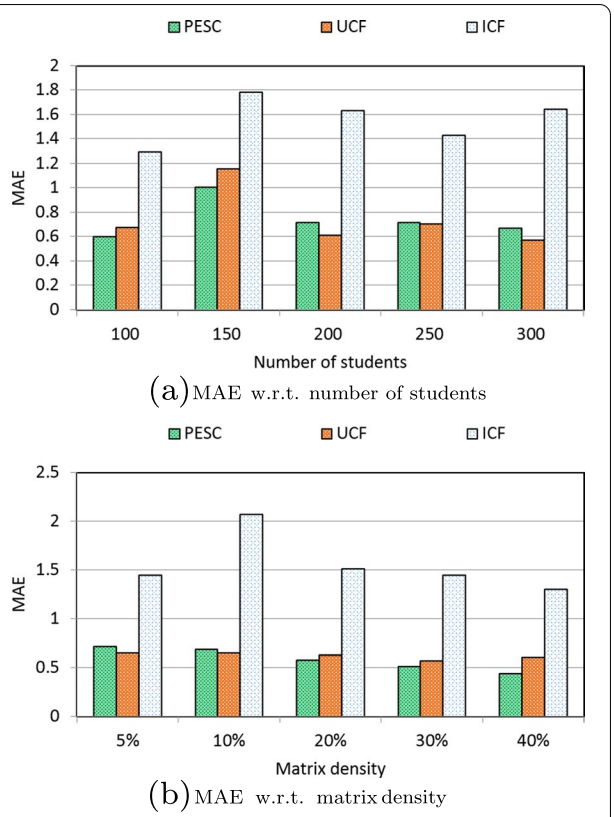

Fig. 3 MAE comparisons

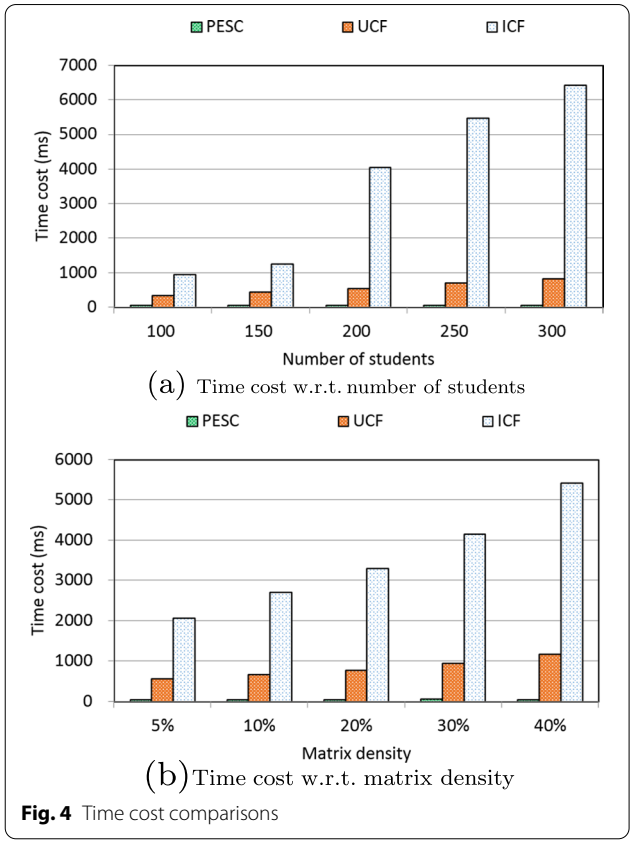

Fig. 4 Time cost comparisons

similar students since the has index technique formalized in Eqs. (4)-(11) is with a good property of similarity keeping. Therefore, our PESC performs well in terms of clustering accuracy.

Moreover, the MAE of three solutions with respect to matrix density is presented in Fig. 3b. As Fig. 3b shows, the MAE of ICF solution is also the highest, which indicates a poor clustering accuracy. On the contrary, PESC and UCF solutions both perform well in MAE, which indicates that their clustering accuracy is often high. Furthermore, like Fig. 3a, the MAE value of PESC is close to that of UCF, which indicates an approximate clustering accuracy between our PESC and baseline UCF. The reason is the same as that analyzed in Fig. 3a, which will not be repeated here.

(2) Time cost comparison

Time cost is a key metric to indicate the algorithm performance especially in the big data environment [52–56]. Here, we measure the clustering efficiency of three solutions via time cost metric under different parameter settings. In concrete, the clustering efficiency of three solutions with respect to the number of students is presented in Fig. 4a. As Fig. 4a indicates, the time cost of ICF solution is the highest since more computational cost is needed for ICF when the number of students is increasing. In addition, the time cost of UCF solution is smaller than ICF, which is still high since all the students need to take part in the similarity calculation. Compared to UCF and ICF, our proposed PESC performs the best in terms of time cost since the hash index technique recruited in PESC is quite efficient due to its low complexity of O(1). Another observation from Fig. 4a is that the time cost of UCF and ICF both increases with the growth of the number of students; while the time cost of our PESC stays approximately stable with the increment of student volume, which indicates a good scalability of PESC in coping with big data.

In addition, the clustering efficiency of three solutions with respect to matrix density is presented in Fig. 4b. Figure 4b show a close result with that of Fig. 4a: the time cost of PESC performs better than UCF and ICF; the time cost of UCF and ICF both increases with the rising of matrix density, while the time cost of PESC stays approximately stable with the growth of matrix density. The reason is the same as that analyzed in Fig. 4a, which will not be repeated here.

(3) MAE of PESC w.r.t. number of hash functions

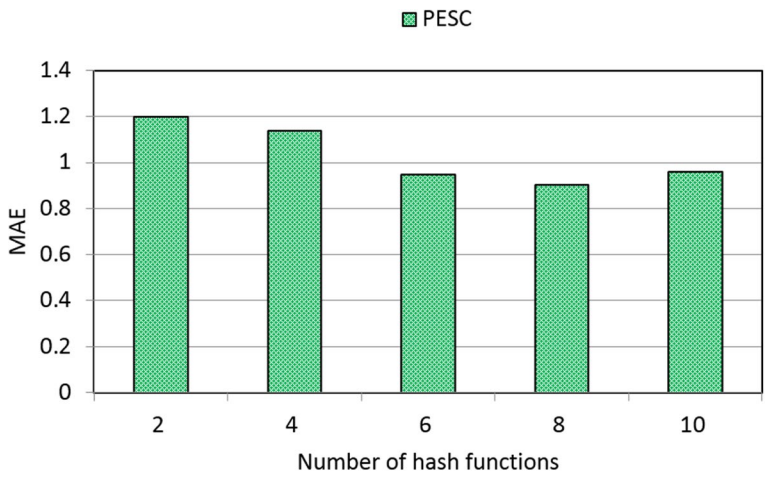

Fig. 5 MAE of PESC

In this profile, we measure the clustering accuracy of PESC via MAE metric under different parameter settings. Here, the parameter is the number of hash functions. Measurement results are presented in Fig. 5. As can be seen from Fig. 5, the MAE of PESC approximately decreases with the growth of the number of hash functions. This result can be explained by the inherent property of the hash index technique adopted in PESC. In concrete, when more hash functions are used to generate the hash indexes of students according to Eqs. (4)-(7), the clustering conditions are very strict and therefore, only a smaller number of really similar students are clustered into a same group. In this situation, the clustering accuracy is improved and MAE is decreased accordingly.

(4) Time cost of PESC w.r.t. number of hash functions

In this profile, we measure the clustering efficiency of PESC under different parameter settings. Here, the parameter is the number of hash functions. Concrete results are presented in Fig. 6. As Fig. 6 indicates, the time cost of PESC first decreases with the growth of the number of hash functions and then stays approximately stable with the growth of the number of hash functions. This result can also be explained by the property of the hash index technique adopted in PESC. Concretely, when more hash functions (e.g., from 2 functions to 4 functions) are recruited in the generation of hash indexes of students according to Eqs. (4)-(7), the clustering conditions become more strict and therefore, only a smaller number of really similar students are clustered into a same group. In this situation, the clustering efficiency is improved significantly. Furthermore, when the number of hash functions continue to become larger (e.g., from 4 functions to 10 functions), the students belonging to same clusters stay stable even the filtering conditions are becoming stricter. In this situation, the time cost also stays approximately stable.

Conclusions

With the wide adoption of health and sport concepts in human society, how to effectively analyze the personalized sports preferences of each student has become a crucial and emergent task in education. In this situation, past sports training records of students recorded in a central cloud platform have provided a theoretically feasible evaluation basis to cluster the students and then infer their respective sports preferences accordingly. However, the past sports training records of students stored in the cloud platform are often accumulated for years and therefore, the data volume is often large, which often leads to a long processing time of cloud platform for student clustering and sports preference identification. In addition, the past sports training records of students registered in the cloud platform often contain some sensitive user information, which probably discloses user privacy if we cannot protect the data well. Considering these two challenges, a privacy-aware and efficient student clustering approach, i.e., PESC is proposed for sport training cloud-assisted smart education. In concrete, we use hash mapping operations to secure the sensitive user privacy. Firstly, we convert the sensitive user data into a less-sensitive user index through a kind of hash mapping process. Secondly, we use the less-sensitive user indexes to cluster users into different groups without disclosing much user privacy. This way, we can

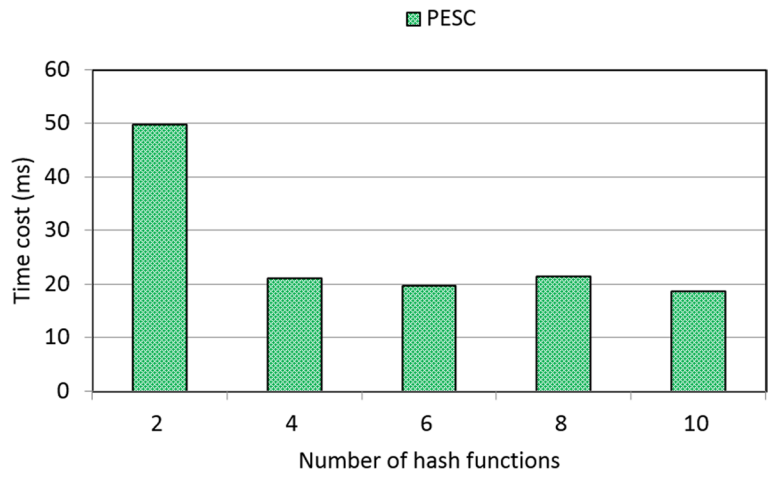

Fig. 6 Time cost of PESC

guarantee the user privacy is secure during the accurate user clustering process. We demonstrate the effectiveness of PESC through an intuitive example and a set of experiments.

However, there are some classic privacy protection solutions besides hash adopted in this paper, such as encryption, anonymization, differential privacy, etc [57–59]. In the future work, we will further compare PESC with other privacy-preserving techniques through experiment comparison. In addition, for practical sport-related applicable scenarios, multiple dimensions as well as their weights are meaningful and crucial. Therefore, we will refine PESC by considering weight information of different sport dimensions involved in student clustering scenarios. At last, we only consider one kind of user data (i.e., score) for simplicity in user clustering, while neglecting the diversity of user data [60–63]. In the future work, we will take the data type diversity into consideration to make our research more robust and applicable.

Abbreviations

PESC: Privacy-aware and efficient student clustering approach;; SDCM-AA: Sports Drug Control model of Adolescent athletes;; UCF: User-based Collaborative Filtering;; ICF: Item-based Collaborative Filtering..

Acknowledgements

We would like to thank the provider of the WS-DREAM dataset.

Authors' contributions

Guoyan Diao: Idea and writing. Fang Liu: Formulation and motivation. Zhikai Zuo: Algorithm and experiment design. Mohammad Kazem Moghimi: Literature investigation and English writing. All authors read and approved the final manuscript.

Availability of data and materials

The WS-DREAM dataset: https://wsdream.github.io/

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Physical Education, Southwest Forestry University, Kunming, China. [2]Faculty of Foreign Languages, Southwest Forestry University, Kunming, China. [3]Department of Communication Engineering, University of Sistan and Baluchestan, Zahedan, Iran.

Received: 8 July 2022 Accepted: 9 September 2022
Published online: 28 September 2022

References

1. Li K, Zhao J, Hu J, et al. Dynamic Energy Efficient Task Offloading and Resource Allocation for NOMA-enabled IoT in Smart Buildings and Environment. Building and Environment; 2022. https://doi.org/10.1016/j.buildenv.2022.109513.
2. Kumari R, Kumar S, Poonia RC, Singh V, Raja L, Bhatnagar V et al (2021) Analysis and predictions of spread, recovery, and death caused by COVID-19 in India. Big Data Min Analytics 4(2):65–75
3. Yang Y (2015) Attribute-based data retrieval with semantic keyword search for e-health cloud. J Cloud Comput 4(1):1–6
4. Kong L, Wang L, Gong W, Yan C, Duan Y, Qi L (2021) LSH-aware multitype health data prediction with privacy preservation in edge environment. World Wide Web. https://doi.org/10.1007/s11280-021-00941-z:1-16
5. Xu X, Jiang Q, Zhang P, Cao X, Khosravi MR, Alex LT, Qi L, Dou W (2022) Game Theory for Distributed IoV Task Offloading with Fuzzy Neural Network in Edge Computing. IEEE Trans Fuzzy Syst. https://doi.org/10.1109/TFUZZ20223158000:1-1
6. Xu X, Tian H, Zhang X, Qi L, He Q, Dou W (2022) DisCOV: Distributed COVID-19 Detection on X-Ray Images With Edge-Cloud Collaboration. IEEE Trans Serv Comput 15(3):1206–19
7. Nitu P, Coelho J, Madiraju P (2021) Improvising personalized travel recommendation system with recency effects. Big Data Min Analytics 4(3):139–154
8. Cai Z, Xiong Z, Xu H, Wang P, Li W, Pan Y (2021) Generative adversarial networks: A survey toward private and secure applications. ACM Comput Surv (CSUR). 54(6):1–38
9. Kou H, Liu H, Duan Y, Gong W, Xu Y, Xu X et al (2021) Building trust/distrust relationships on signed social service network through privacy-aware link prediction process. Appl Soft Comput 100:106942

10. Zheng X, Cai Z (2020) Privacy-Preserved Data Sharing Towards Multiple Parties in Industrial IoTs. IEEE J Sel Areas Commun 38(5):968–979
11. Gerke A, Dalla Pria Y (2018) Cluster concept: Lessons for the sport sector? Toward a two-step model of sport cluster development based on socioeconomic proximity. J Sport Manag 32(3):211–226
12. Kim C, Kim J, Jang S (2021) Sport clusters and community resilience in the United States. J Sport Manag 35(6):566–580
13. Park S, Kim S, Chiu W (2021) Segmenting sport fans by eFANgelism: a cluster analysis of South Korean soccer fans. Manag Sport Leis 1-15
14. Hautbois C, Djaballah M, Desbordes M (2020) The social impact of participative sporting events: a cluster analysis of marathon participants based on perceived benefits. Sport Soc 23(2):335–353
15. Cartigny E, Fletcher D, Coupland C, Bandelow S (2021) Typologies of dual career in sport: A cluster analysis of identity and self-efficacy. J Sports Sci 39(5):583–590
16. Schnider L, Schilling R, Cody R, Kreppke JN, Gerber M (2022) Effects of behavioural skill training on cognitive antecedents and exercise and sport behaviour in high school students: a cluster-randomised controlled trial. Int J Sport Exerc Psychol 20(2):451–473
17. Choi SM, Sum KWR, Leung FLE, Wallhead T, Morgan K, Milton D et al (2021) Effect of sport education on students' perceived physical literacy, motivation, and physical activity levels in university required physical education: a cluster-randomized trial. High Educ 81(6):1137–1155
18. Nicholls AR, Levy AR, Meir R, Sanctuary C, Jones L, Baghurst T et al (2020) The susceptibles, chancers, pragmatists, and fair players: an examination of the sport drug control model for adolescent athletes, cluster effects, and norm values among adolescent athletes. Front Psychol 11:1564
19. Qi L, Yang Y, Zhou X, Rafique W, Ma J (2021) Fast Anomaly Identification Based on Multi-Aspect Data Streams for Intelligent Intrusion Detection Toward Secure Industry 4.0. IEEE Trans Ind Inf https://doi.org/10.1109/TII20213139363
20. Wang F, Li G, Wang Y, Rafique W, Khosravi MR, Liu G et al (2022) Privacy-aware traffic flow prediction based on multi-party sensor data with zero trust in smart city. ACM Trans Internet Technol (TOIT). https://doi.org/10.1145/3511904
21. Qi L, Hu C, Zhang X, Khosravi MR, Sharma S, Pang S et al (2021) Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment. IEEE Trans Ind Inf 17(6):4159–4167
22. Huang J, Tong Z, Feng Z () Geographical POI recommendation for Internet of Things: A federated learning approach using matrix factorization. Int J Commun Syst e5161. https://doi.org/10.1002/dac.5161
23. Yuan L, He Q, Tan S, Li B, Yu J, Chen F et al (2021) Coopedge: A decentralized blockchain-based platform for cooperative edge computing. Proceedings of the Web Conference 2021:2245–2257
24. Chen Y, Zhao F, Lu Y, Chen X () Dynamic task offloading for mobile edge computing with hybrid energy supply. Tsinghua Sci Technol 10. https://doi.org/10.26599/TST.2021.9010050
25. Zhou X, Li Y, Liang W (2020) CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. IEEE/ACM Trans Comput Biol Bioinforma 18(3):912–921
26. Shao Q, Yu R, Zhao H, Liu C, Zhang M, Song H et al (2021) Toward intelligent financial advisors for identifying potential clients: a multitask perspective. Big Data Min Analytics 5(1):64–78
27. Catlett C, Beckman P, Ferrier N, Nusbaum H, Papka ME, Berman MG et al (2020) Measuring Cities with Software-Defined Sensors. J Soc Comput 1(1):14–27
28. Sandhu AK (2022) Big data with cloud computing: Discussions and challenges. Big Data Min Analytics 5(1):32–40
29. Bouras MA, Farha F, Ning H (2020) Convergence of computing, communication, and caching in Internet of Things. Intell Converged Netw 1(1):18–36
30. Xu X, Fang Z, Zhang J, He Q, Yu D, Qi L, et al (2021) Edge Content Caching with Deep Spatiotemporal Residual Network for IoV in Smart City. ACM Trans Sen Netw 17(3). https://doi.org/10.1145/3447032
31. Chen Y, Gu W, Li K () Dynamic task offloading for Internet of Things in mobile edge computing via deep reinforcement learning. Int J Commun Syst e5154. https://doi.org/10.1002/dac.5154
32. Chen H, Yang C, Zhang X, Liu Z, Sun M, Jin J (2021) From Symbols to Embeddings: A Tale of Two Representations in Computational Social Science. J Social Comput 2(2):103–156
33. Qi L, Lin W, Zhang X, Dou W, Xu X, Chen J (2022) A Correlation Graph based Approach for Personalized and Compatible Web APIs Recommendation in Mobile APP Development. IEEE Trans Knowl Data Eng 1-1
34. Chen Y, Liu Z, Zhang Y et al (2021) Deep reinforcement learning-based dynamic resource management for mobile edge computing in industrial internet of things. IEEE Trans Ind Inf 17(7):4925–4934
35. Dai H, Xu Y, Chen G, Dou W, Tian C, Wu X et al (2022) (2022) ROSE: Robustly Safe Charging for Wireless Power Transfer. IEEE Trans Mob Comput 21(6):2180–2197
36. Chen Y, Zhao F, Chen X, Wu Y (2022) Efficient Multi-Vehicle Task Offloading for Mobile Edge Computing in 6G Networks. IEEE Trans Veh Technol 71(5):4584–4595
37. Zhou X, Liang W, Li W, Yan K, Shimizu S (2022) Wang KIK (2022) Hierarchical Adversarial Attacks Against Graph-Neural-Network-Based IoT Network Intrusion Detection System. IEEE Int Things J 9(12):9310–9319
38. Zhu K, Zhang T (2021) Deep reinforcement learning based mobile robot navigation: A review. Tsinghua Sci Technol 26(5):674–691
39. Ying C, Hua X, Zhuo M, et al. (2022) Cost-Efficient Edge Caching for NOMA-enabled IoT Services. China Commun
40. Zhou J, Li L, Vajdi A, Zhou X, Wu Z (2021) Temperature-Constrained Reliability Optimization of Industrial Cyber-Physical Systems Using Machine Learning and Feedback Control. IEEE Trans Autom Sci Eng 1-12. https://doi.org/10.1109/TASE.2021.3062408.
41. Wernke SA (2022) Explosive Expansion, Sociotechnical Diversity, and Fragile Sovereignty in the Domain of the Inka. J Soc Comput 3(1):57–74
42. Xu Y, Liu Z, Zhang C, Ren J, Zhang Y, Shen X (2022) Blockchain-Based Trustworthy Energy Dispatching Approach for High Renewable Energy Penetrated Power Systems. IEEE Int Things J 9(12):10036–10047
43. Li T, Li C, Luo J, Song L (2020) Wireless recommendations for Internet of vehicles: Recent advances, challenges, and opportunities. Intell Converged Netw 1(1):1–17
44. Zhou X, Yang X, Ma J, Wang KIK (2021) Energy Efficient Smart Routing Based on Link Correlation Mining for Wireless Edge Computing in IoT. IEEE Int Things J 1-1. https://doi.org/10.1109/JIOT.2021.3077937.
45. Gu R, Zhang K, Xu Z, Che Y, Fan B, Hou H, et al. (2022) Fluid: Dataset Abstraction and Elastic Acceleration for Cloud-native Deep Learning Training Jobs. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE). p. 2182-2195
46. Cai Z, He Z (2019) Trading Private Range Counting over Big IoT Data. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). p. 144-153
47. Xu Y, Zhang C, Wang G, Qin Z, Zeng Q (2021) A Blockchain-Enabled Deduplicatable Data Auditing Mechanism for Network Storage Services. IEEE Trans Emerg Top Comput 9(3):1421–1432
48. Cheung J (2021) Real Estate Politik: Democracy and the Financialization of Social Networks. J Soc Comput 2(4):323–336
49. Zhang C, Xu Y, Hu Y, Wu J, Ren J, Zhang Y (2021) A Blockchain-Based Multi-Cloud Storage Data Auditing Scheme to Locate Faults. IEEE Trans Cloud Comput 1-1. https://doi.org/10.1109/TCC.2021.3057771.
50. Cai P, Zhang Y (2020) Intelligent cognitive spectrum collaboration: Convergence of spectrum sensing, spectrum access, and coding technology. Intell Converged Netw 1(1):79–98
51. Xu Y, Ren J, Zhang Y, Zhang C, Shen B, Zhang Y (2020) Blockchain Empowered Arbitrable Data Auditing Scheme for Network Storage as a Service. IEEE Trans Serv Comput 13(2):289–300
52. Huang J, Lv B, Wu Y et al (2022) Dynamic Admission Control and Resource Allocation for Mobile Edge Computing Enabled Small Cell Network. IEEE Trans Veh Technol 71(2):1964–1973
53. Zhang K, Tian Z, Cai Z, Seo D (2021) Link-privacy preserving graph embedding data publication with adversarial learning. Tsinghua Sci Technol 27(2):244–256
54. Zeng Q, Zhou Q, He X, Sun Y, Li X, Chen H (2021) Polar Codes: Encoding/Decoding and Rate-Compatible Jointly Design for HARQ System. Intelligent and Converged Networks 2(4):334–346
55. Gu R, Chen Y, Liu S, Dai H, Chen G, Zhang K et al (2022) Liquid: Intelligent Resource Estimation and Network-Efficient Scheduling for Deep Learning Jobs on Distributed GPU Clusters. IEEE Transactions on Parallel and Distributed Systems 33(11):2808–2820
56. Xu J, Li D, Gu W et al (2022) UAV-assisted Task Offloading for IoT in Smart Buildings and Environment via Deep Reinforcement Learning. Building and Environment. https://doi.org/10.1016/j.buildenv.2022.109218

57. Zhou J, Zhang M, Sun J, Wang T, Zhou X, Hu S (2022) DRHEFT: Deadline-Constrained Reliability-Aware HEFT Algorithm for Real-Time Heterogeneous MPSoC Systems. IEEE Transactions on Reliability. 71(1):178–189
58. Dai H, Wang X, Lin X, Gu R, Shi S, Liu Y, et al. (2021) Placing Wireless Chargers with Limited Mobility. IEEE Trans Mob Comput 1-1. https://doi.org/10.1109/TMC.2021.3136967.
59. Cai Z, Zheng X (2020) A Private and Efficient Mechanism for Data Uploading in Smart Cyber-Physical Systems. IEEE Trans Netw Sci Eng 7(2):766–775
60. Zhou X, Liang W, Kevin I, Wang K, Huang R (2018) Jin Q (2018) Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data. IEEE Trans Emerg Top Comput 9(1):246–257
61. Zhang W, Li Z, Chen X (2021) Quality-aware user recruitment based on federated learning in mobile crowd sensing. Tsinghua Sci Technol 26(6):869–877
62. Zhou X, Xu X, Liang W, Zeng Z, Yan Z (2021) Deep-Learning-Enhanced Multitarget Detection for End–Edge–Cloud Surveillance in Smart IoT. IEEE Int Things J 8(16):12588–12596
63. Zhou J, Cao K, Zhou X, Chen M, Wei T, Hu S (2022) Throughput-Conscious Energy Allocation and Reliability-Aware Task Assignment for Renewable Powered In-Situ Server Systems. IEEE Trans Comput Aided Des Integr Circ Syst 41(3):516–529

Publisher's Note