

Received August 19, 2020, accepted September 3, 2020, date of publication September 7, 2020,
date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022404

Secure User Privacy in Population Physique Clustering and Prediction Based on Sport Questionnaires

CHUNLAN TIAN¹, CHONGMIN ZHANG¹, WANLI HUANG²,
AND HAO WANG³, (Member, IEEE)

¹College of Sport Science, Qufu Normal University, Rizhao 276826, China

²School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China

³Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway

Corresponding authors: Wanli Huang (wanlih1983@126.com) and Hao Wang (hawa@ntnu.no)

ABSTRACT Population physique is one of the key aspects for measuring and evaluating the healthy degree or living level of the population of a nation. Many population physique evaluation models or tools have been developed in the past decades, among which questionnaire is an essential and promising way to help achieve the above population physique measurement goal. Typically, through designing, distributing and collecting a variety of sport questionnaires associated with people's living conditions or sport preferences, we can quantify, observe and cluster the healthy degree of the whole nation's population objectively and scientifically. However, the above sport questionnaire-based population physique analysis methods are often time-consuming as a considerable amount of sport questionnaires needs to be compared. Moreover, the sport questionnaires filled out by people often contains some sensitive information that is not supposed to be disclosed to other people, which also call for appropriate privacy protection measures. Inspired by the above two challenges, we introduce hash techniques into population physique evaluation process and afterwards, we propose a hash-based population physique clustering and prediction method that is efficient in time cost and effective in terms of privacy protection. At last, we designed experiments based on a dataset to prove the effectiveness of our proposal in this research work.


INDEX TERMS Population physique, clustering, hashing, privacy protection, time efficiency, sport questionnaires.

I. INTRODUCTION

With the economic growth and social progress in the last decades, people's living conditions have been improved considerably in terms of both physical and mental aspects. As a result, more and more people are gradually changing their attentions from living to life, through focusing on various health-guaranteed measures such as scientific and healthy diet, sufficient and high-quality sleeping and moderate sport activities, and so on [1], [2]. Therefore, in recent years, healthcare industry is gaining a rapid development and has constituted an important part of the whole nation's or society's economical system. Correspondingly, healthcare-related research (e.g., intelligent medical care, scientific exercise) is also becoming a hot and significant

research topic that draws wide attentions from both academy and industries [3], [4].

As one of the popular healthcare-related research topics, population physique monitoring has always been a hot research direction as it can effectively reflect the current health and development conditions of all the individuals of a nation or a country. To achieve this goal, a classic way is through statistics-related techniques. Typically, to realize population physique monitoring, we can design, distribute and collect a variety of sport questionnaires associated with people's living conditions or sport preferences from individuals. Thus, through the collected sport questionnaires, we can quantify, observe and evaluate the healthy degree of the whole nation's population objectively and scientifically. Furthermore, according to the sport questionnaires, a variety of related applications can also be developed, such as population physique prediction with time, population physique

The associate editor coordinating the review of this manuscript and approving it for publication was Gautam Srivastava .

trend analyses, population clustering based on physique, and so on.

However, there are often several critical challenges existing in the above sport questionnaires-based population physique monitoring and utilization process. First of all, the traditional population physique analysis solutions based on sport questionnaires often fall short in quick response as there are often a huge amount of collected sport questionnaires that are necessary to be integrated and compared. In addition, sport questionnaires collected from individuals are usually sensitive enough as they often include the personalized information of individuals. In most cases, individuals are often reluctant to release their private information to the third party.

Therefore, it is becoming a necessity to develop novel techniques or solutions to cope with the abovementioned two challenges existing in sport questionnaires-based population physique monitoring and utilization process. Inspired by this observation, hash techniques with privacy-preservation effects are introduced into population physique analyses and prediction, to further improve the performances of present population physique monitoring and utilization.

In summary, the major contributions of this research work are three-fold.

(1) We recognize the significant importance of hash techniques in securing the sensitive information hidden in sport questionnaires from individuals.

(2) We utilize hash techniques to create the healthcare status indices for individuals and use the indices to achieve time-efficient and privacy-free population physique clustering and prediction.

(3) A variety of simulation experiments are enacted and developed based on a real-world dataset. Reported experimental results show the feasibility of our algorithm in coping with sport questionnaires-based population physique monitoring and utilization.

The rest of this research paper is organized as follows. We investigate the up-to-date research work in the same field in Section II to better outline research significance of our paper. Section III presents a real-world example to demonstrate the paper motivation more clearly. We clarify the concrete steps of our suggested sport questionnaires-based population physique clustering and prediction method in Section IV. Evaluations are made in Section V. Finally, we summarize the research significance as well as its advantages and disadvantages in Section VI.

II. RELATED WORK

In this section, we investigate the current research status of the field of privacy-free data analyses and prediction, and compare the related work in existing literatures with a discussion about their respective advantages and disadvantages, so as to further outline the research significance of this work.

A. DATA ENCRYPTION FOR PRIVACY PROTECTION

As a classical data protection manner, encryption has been studied for thousands of years. The authors have focused on

multiple keywords-driven information search with privacy-preservation [5]. This method uses symmetric public key mechanism to achieve data encryption when searching for the pre-designated multiple keywords. Although secure data transmission and search are achieved, this method often falls short in time efficiency. The authors use oval curve encryption mechanism to pursue privacy-aware information reuse and show the competitive advantages of the proposal compared to existing solutions [6]. Although this solution can achieve high privacy protection capability, the applicable domain is a bit narrow as it is especially designed for Boolean type data-driven information retrieval. The authors adopt homomorphic manner for secure data transmission [7]. Although high privacy-protection performances of this solution are available, additional data transmissions among different parties are inevitable. As a consequence, computational cost is increased accordingly. Similar work can be found in [8] where homomorphic manner is adopted to secure sensitive keywords that are ready to be retrieved. Although this solution can protect data privacy better, fuzzy keywords search from users' undetermined requirements are not considered, which decrease the proposal's comprehensiveness.

B. DIFFERENTIAL PRIVACY FOR DATA PROTECTION

Differential technology has recently been put forward for measurable and computational privacy protection in various applications associated with sensitive information. The authors combine Differential technology and collaborative techniques for trusted data protection when performing multiple-party information fusion and integration [9]. Although this solution can alleviate the noise issues raised by Differential technology, the computational cost is still high. Similar job is done in [10] where a tradeoff between privacy protection and data utilization is achieved. In [11], the authors combine Differential technology and Matrix Factorization techniques for missing value prediction that considers both prediction accuracy and privacy disclosure probability. This solution can secure user privacy to some extent; however, the data availability and prediction precision are reduced with the growth of privacy-protection performances. Differential technology and trust information are combined in [12] to balance the performances of privacy protection and data use. Other related work includes [13] where Differential technology and Huffman Coding are fused for location privacy protection, and [14] where the authors combine Differential technology and information entropy for sensitive information protection.

However, the abovementioned Differential technology-based data protection solutions often fall short in noise disturbance and high computational cost raised by Differential technology itself.

C. ANONYMIZED DATA FOR PRIVACY PROTECTION

Eliminating or generalizing the sensitive user identity information from the data to be released is called data anonymization or generalization, and has been widely used in various

domains [15]. Typical user identity information includes user name, user identity ID, and so on. After eliminating this sensitive information, the rest data without much privacy can be opened to the third party for reuse [16]. Motivated by this advantage, the authors introduce K-anonymity strategy for expert system decisions that involve a considerable amount of private information [17]. Similar job is done in [18] to secure the sensitive locations of users. In summary, the above-mentioned research work can all secure private information to some extent; however, it is inevitable to drop some valuable information after data are anonymized.

The above analyses indicate that although many research works have paid attentions to the value extraction from massive data [19]–[24], data privacy and data utilization cannot be balanced well due to the natural tradeoffs between them. Inspired by this observation, we introduce hash technology into data analyses and prediction when sport questionnaires-based population physique monitoring and utilization are performed.

To perform population physique clustering and prediction, we need to compare the sport questionnaires from *Jack* and *Carolina* so as to analyze their physique similarity or physique differences. While as seen in the scenario, the sport questionnaires often contain certain private information of *Jack* and *Carolina*. As a result, analyzing these distributed sport questionnaires from different people often inevitably disclose the people’s privacy. In addition, due to the existence of so many people as well as their respective sport questionnaires, the comparison and analyses of the existing sport questionnaires are often a time-consuming computational task that cannot be done in a limited time period.

In view of the above limitation analyses, we employ hash techniques that can make privacy-free similarity calculation to perform the above population physique clustering and prediction task. Concrete procedure will be described in detail in the following section. Besides, for formal description and solving of the problem focused in this article, here, we specify the symbols and their meanings with Table 1.

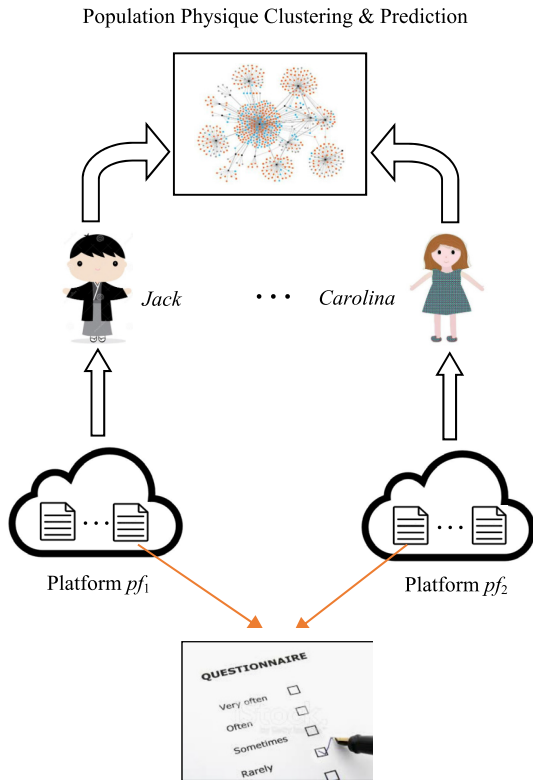


FIGURE 1. Sport physique clustering and prediction based on sport questionnaires.

III. MOTIVATION

We use the example shown in Fig.1 to motivate our research in this article. In the scenario, *Jack* and *Carolina* both have their respective answered sport questionnaires (e.g., problem “sport frequency” with four answer choices: {*Very often*, *Often*, *Sometimes*, *Rarely*, *Never*}), recorded in platforms p_1 and p_2 , respectively.

TABLE 1. Symbols & Meanings.

Symbols	Meanings
P_1, \dots, P_n	A set of people whose physiques are evaluated
q_1, \dots, q_m	Questions in sport questionnaires
f_1, \dots, f_M	Hash functions
T_1, \dots, T_N	Hash tables
pf_1, \dots, pf_h	Platforms recording sport questionnaires
v_1, \dots, v_m	m dimensions of each hash function
$h_1(P_x), \dots, h_M(P_x)$	Hash values of P_x based on M hash functions
$H_1(P_x), \dots, H_N(P_x)$	Indices of P_x in N hash tables

IV. PPCP METHOD

The proposed population physique clustering and prediction method is named PPCP. The basic rationale of PPCP is: we first convert each sensitive sport questionnaire into a privacy-free index value through hash. Then we perform privacy-free similar individuals finding based on hash theory. At last, we cluster or predict the population physiques with the obtained similar individuals. Concrete procedure of PPCP is briefly introduced in Fig. 2.

(1) Step-1: From sport questionnaires to individual indices.

As presented in the example of Fig.1, an individual’s sport questionnaire is often constituted by a set of pre-designed multiple-choice questions, represented by q_1, \dots, q_m as formalized in Table 1. Each question in the sport questionnaire is often accompanied by a several choices such as the five choices in Fig.1, i.e., {*Very often*, *Often*, *Sometimes*, *Rarely*, *Never*}. Such a kind of discrete textual description cannot be

Step-1: From sport questionnaires to individual indices. We convert the questionnaire answers to scale 1~Z. Then we convert each individual's sport questionnaire whose elements are scaled from 1 to Z to an individual index based on hash.

Step-2: Similar individuals finding based on hash. With the individuals' index values derived in Step-1, we find the similar individuals who share the similar physiques based on hash.

Step-3: Population physique clustering and prediction based on similar individuals. With the similar individuals obtained in Step-2, we cluster the individuals by their population physiques and predict the missing physique of a new comer.

FIGURE 2. Procedure of our PPCP.

directly used for population physique clustering and prediction that are focused in this research work.

Therefore, we convert the textual descriptions of choices of the multiple-choice questions into numeric expressions, so as to ease the subsequent calculation. Typically, we can use scale 1~Z to represent the multiple choices of the questions in a sport questionnaire. Taking the example in Fig.1 for illustration: the five choices in answer set $\{Very\ often, Often, Sometimes, Rarely, Never\}$ can be converted into the five elements in set $\{5, 4, 3, 2, 1\}$, respectively (assume a larger number means more frequent sport exercises, vice versa). In other words, parameter Z is equal to 5 here.

However, for different multiple-choice questions in an identical sport questionnaire, their respective Z values are not always the same or equal. Let's consider the following multiple-choice question with three answer choices. In the example, Z is equal to 3 and the three choices should be converted into the three elements in set $\{3, 2, 1\}$, respectively.

Question: sport strength ratings.

Answer choices: {High, Median, Low}.

With the above two examples with different Z values, we can observe that the different Z values are not beneficial to the uniform calculation. Considering this drawback, we normalize the different Z values corresponding to different multiple-choice questions in an identical sport questionnaire, through the scaling formulas in equations (1)-(2). Here, $Norm(q_i)$ means the normalized value of original value of dimension q_i ; $min(q_i)$ and $max(q_i)$ represent the minimal and maximal values of dimension q_i , respectively. For example, in the example in Fig.1, $min(q) = 1$ and $max(q) = 5$. Besides, $value(q_i)$ denotes a certain concrete value of dimension q_i ; for example, in the above example, $value(q) = 1$ or 2 or 3 or 4 or 5. This way, we can get a set of normalized values for the m dimensions $\{q_1, \dots, q_m\}$, i.e., $\{Norm(q_1), \dots, Norm(q_m)\}$ falling into range $[0, 1]$.

For q_i ($1 \leq i \leq m$) of "larger is better" property

$$Norm(q_i) = \frac{value(q_i) - \min(q_i)}{\max(q_i) - \min(q_i)} \quad (1)$$

For q_i ($1 \leq i \leq m$) of "smaller is better" property

$$Norm(q_i) = \frac{\max(q_i) - value(q_i)}{\max(q_i) - \min(q_i)} \quad (2)$$

Thus, through (1)-(2), we can convert each individual P_x ($1 \leq x \leq n$)'s sport questionnaire into a normalized vector $Vec(P_x) = (Norm(q_1), \dots, Norm(q_m))$ in which each element belongs to range $[0, 1]$. $Vec(P_x)$ is typically sensitive as its elements $Norm(q_1), \dots, Norm(q_m)$ often contain certain privacy of the individual. Next, we try to minimize the privacy amount contained in $Vec(P_x)$ through the well-known hash technology.

Concretely, we generate an m-dimensional vector $X = (v_1, \dots, v_m)$ where each dimension's value falls into range $[-1, 1]$. Thus, we use the hash function f in equations (3)-(4) to achieve the projection from $Vec(P_x)$ with privacy to Boolean value of 0 or 1 with little privacy. In concrete, in (3), " \odot " represents the dot product; in other words, considering two vectors $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$, $A \odot B = a_1 * b_1 + a_2 * b_2 + \dots + a_n * b_n$. Thus, through " \odot " operation, the value of f ($Vec(P_x)$) is either positive or negative or 0. Then according to the projection function in (4), we can further convert f ($Vec(P_x)$) with partial individual privacy into a Boolean value $h(P_x)$ with no privacy. This way, through the two-step hash projection process in (3)-(4), the privacy protection goal is achieved partially. Concrete procedure of this step is available in Algorithm1.

$$f(Vec(P_x)) = Vec(P_x) \odot X \quad (3)$$

$$h(P_x) = \begin{cases} 1 & \text{if } f(Vec(P_x)) > 0 \\ 0 & \text{else} \end{cases} \quad (4)$$

(2) Step-2: Similar individuals finding based on hash.

We repeat the above process in Step-1 by introducing more hash functions: f_1, \dots, f_M . Thus, for each individual P_x ($1 \leq x \leq n$), we can obtain n Boolean values: $h_1(P_x), \dots, h_M(P_x)$ according to equations (3)-(4). Afterwards, we get an m-dimensional Boolean vector $H(P_x)$ as in equation (5). Thus, $H(P_x)$ (e.g., $(1, 0, 1, 1, 0)$ if $M = 5$) is considered as the index value of the individual P_x .

$$H(P_x) = (h_1(P_x), \dots, h_M(P_x)) \quad (5)$$

According to similarity-unchanged nature of hash theories, two individuals with similar index values are also similar. Inspired by this nature, we can use index values $H(P_1), \dots, H(P_n)$ containing less privacy to seek for the similar individuals who own similar physiques, instead of using sensitive $answer_x(q_1), \dots, answer_x(q_m)$ ($1 \leq x \leq n$).

In concrete, we put the $(P_x, H(P_x))$ ($1 \leq x \leq n$) pairs in a hash table T. Due to the probabilistic nature of hash techniques, multiple hash tables are generated by repeating the above process N times, i.e., T_1, \dots, T_N . Detailed procedure is available in Algorithm 2.

Algorithm 1**Inputs:**

- (1) P_1, \dots, P_n : individuals whose physiques are to be evaluated
- (2) q_1, \dots, q_m : Questions in sport questionnaires.
- (3) $\text{answer}_x(q_1), \dots, \text{answer}_x(q_m)$: P_x 's answer set for questions.

Output:

- (1) $h(P_x)$: Boolean value of individual P_x .

```

1: for x = 1, ..., n do
2:   for i = 1, ..., m do
3:     convert  $\text{answer}_x(q_i)$  into  $\text{value}(q_i)$  by the example in Step-1
4:     if  $q_i$  is positive
5:       then calculate  $\text{Norm}(q_i)$  by (1)
6:     else calculate  $\text{Norm}(q_i)$  by (2)
7:     end if
8:   end for
9: end for
10: for i = 1, ..., m do
11:    $v_j = \text{random}[-1, 1]$ 
12: end for
13: for x = 1, ..., n do
14:    $\text{Vec}(P_x) = (\text{Norm}(q_1), \dots, \text{Norm}(q_m))$ 
15:    $f(\text{Vec}(P_x)) = 0$ 
16:   for i = 1, ..., m do
17:      $\text{Vec}(P_x) = \text{Vec}(P_x) + \text{Norm}(q_i) * v_j$ 
18:   end for
19:   if  $\text{Vec}(P_x) > 0$ 
20:     then  $h(P_x) = 1$ 
21:   else  $h(P_x) = 0$ 
22:   end if
23: return  $h(P_x)$ 
24: end for

```

For two individuals P_{x1} and P_{x2} , if their corresponding index values are the same in any of the N tables, then it is probably concluded that P_{x1} and P_{x2} are similar individuals, vice versa. More formally, the similar individual evaluation process is based on the equation (6), where $y = 1$ or 2 or ... or N .

$$P_{x1} \text{ is similar with } P_{x2} \text{ iff } H_y(P_{x1}) = H_y(P_{x2}) \quad (6)$$

(3) Step-3: Population physique clustering and prediction based on similar individuals.

Through Algorithm 2, similar individuals who share the same or similar sport physiques are discovered. Next, according to the individual similarity, we can cluster the populations into different groups. In each group, all the individuals hold approximately same sport preferences or habits. With the derived population groups, various group-based population analysis operations can be performed further.

Another alternative application based on population groups is individual physique prediction. Generally, the feedback rate of sport questionnaires is not high enough.

Algorithm 2**Inputs:**

- (1) $h(P_1), \dots, h(P_n)$: Boolean values of individuals after projection;
- (2) f_1, \dots, f_M : projection functions.

Output:

- (1) T_1, \dots, T_N : N hash tables.

```

1: for y = 1, ..., N do
2:    $T_y = \text{Null}$ 
3:   for j = 1, ..., M do
4:     execute Algorithm 1 by  $f_j$ 
5:   end for
6:   for x = 1, ..., n do
7:      $H(P_x) = (H_1(P_x), \dots, H_M(P_x))$ 
8:     put  $(P_x, H(P_x))$  pair in  $T_y$ 
9:   end for
10: return  $T_y$ 
11: end for

```

As a consequence, the collected sport physique data of population are often sparse, which probably decrease the availability and dependability of the collected sport questionnaires.

In this situation, to improve the quality of collected sport questionnaires and enhance the availability of questionnaire data, a promising way is to preprocess the collected data and predict the missing values existing in the questionnaires. Specifically, for an individual P_x whose answer choice for question q is absent from the collected questionnaires, his/her missing data, denoted by $\text{value}_x(q)$, can be predicted by equation (7). Here, $\text{Sim}(P_x)$ records all the similar individuals of P_x . Formal procedure is available in Algorithm 3.

$$\text{value}_x(q) = \frac{1}{|\text{Sim}(P_x)|} * \sum_{P_k \in \text{Sim}(P_x)} \text{value}_k(q) \quad (7)$$

V. EVALUATION

In this section, we simulate a set of experiments to test the feasibility of our PPCP method in dealing with population physique clustering and prediction.

A. EXPERIMENTAL SETTING

As the answer choices of the questions in the sport questionnaires are often discrete values, we employ the classic MovieLens dataset [25] for experimental evaluation purpose. Similar to sport questionnaires, the ratings in MovieLens dataset are also discrete values scaled from 1-star to 5-star.

In the simulation, the population volume (n) is up to 6040, the question volume in the sport questionnaires (m) is up to 3900, the hash function volume (M) is up to 15 and the hash table volume (N) is up to 15.

For performance comparisons, we introduce two baseline methods: UPCC and IPCC. These two compared methods are both classic and effective in performing similar objects search. Moreover, we mainly measure the RMSE and time

Algorithm 3

Inputs:

- (1) T_1, \dots, T_N : hash tables;
- (2) P_x : an individual whose questionnaire data are missing.
- (3) q : a question of the sport questionnaires

Output:

$value_x(q)$: predicted value of q of P_x

```

1: Sim ( $P_x$ ) = Null
2: for  $y = 1$  to  $N$  do // hash tables
3:   for  $i = 1, \dots, n$  do // individuals
4:     if  $H_y(P_i) = H_y(P_x)$ 
5:       then include  $P_i$  in Sim ( $P_x$ )
6:     end if
7:   end for
8: end for
9: TOTAL = 0
10: for  $k = 1$  to | Sim ( $P_x$ )| do
11:   TOTAL = TOTAL +  $value_k(q)$ 
12: end for
13:  $value_x(q) = TOTAL / | \text{Sim} (P_x) |$ 
14: return  $value_x(q)$ 
    
```

costs of different solutions. The experiment running environment is: 3.20 GHz processor, 4.0 GB memory, Windows 7 OS and JAVA 8.

B. RESULTS

1) RMSE OF THREE METHODS

RMSE is regarded as a key indicator that influences the prediction accuracy when missing values are predicted. In this test, we measure the RMSE of three methods with the variation of parameters m and n , respectively. Other parameters: $M = 15, N = 15$. Compared results are reported in Fig. 3.

From Fig.3, there is no obvious variation tendency of RMSE for each of the three methods. While an obvious observation is available in Fig.3, i.e., PPCP’s RMSE value is often superior to the other two methods. We analyze the reasons as below: in PPCP, three tuning the parameter settings, only the most similar individuals are returned for missing data prediction. As a result, the prediction RMSE is decreased accordingly.

2) TIME COST OF THREE METHODS

In big data environment, response time is a key factor that influences the final satisfactions of people. Therefore, we measure the time cost of three methods to quantify their respective response speed. Other parameters: $M = 15, N = 15, m = 3900, n = 6040$. Compared results are reported in Fig. 4.

A clear variation tendency is found in Fig.4 for both UPCC and IPCC methods, as these two solutions often involve heavy-weight similarity computational costs. Especially when the population or question volume increases, the computational time grows approximately linearly. On the

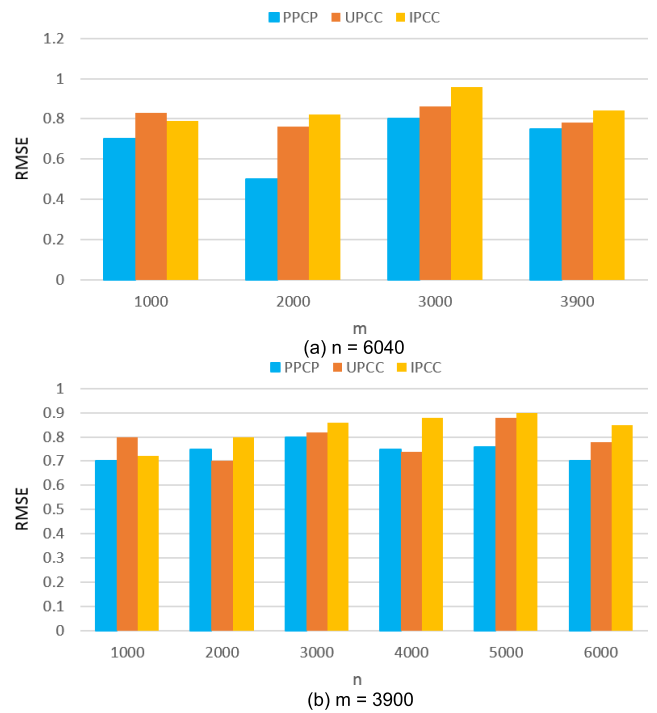


FIGURE 3. RMSE comparisons.

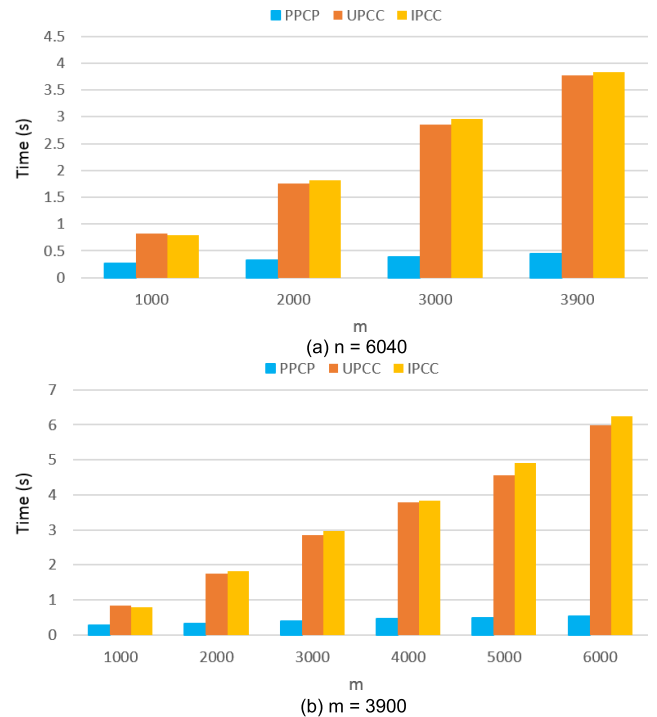


FIGURE 4. Time consumption of three methods.

contrary, the time efficiency of PPCP is relatively high compared to UPCC and IPCC. We analyze the reasons as follows: in PPCP, the first two steps are often done offline, whose time cost is $O(1)$; in the third step, we only need to query the individual index tables produced offline to make population physique clustering and prediction, whose time is $O(1)$. As a result, the response speed of PPCP is high enough.

3) RMSE OF PPCP

As observed from Section IV, the proposed three algorithms are related to several parameters. In this test, we observe the relationship between the RMSE of PPCP and the involved parameters such as M and N. Other performance parameters: $m = 3900$, $n = 6040$. Compared results are reported in Fig. 5.

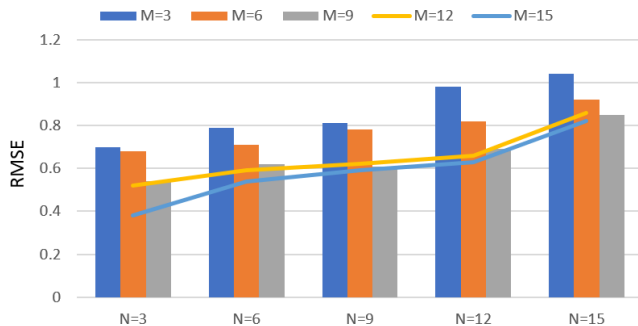


FIGURE 5. RMSE of PPCP with M and N.

Fig.5 shows the obvious change tendency of RMSE of PPCP with the growth of M and N. In general, when M value increases, the RMSE of PPCP decreases as more hash functions often indicate a more rigid similar individual discovery condition; as a result, RMSE drops accordingly. On the contrary, when N value increases, the RMSE of PPCP rises as more hash tables often indicate a looser similar individual discovery condition; as a result, RMSE grows accordingly.

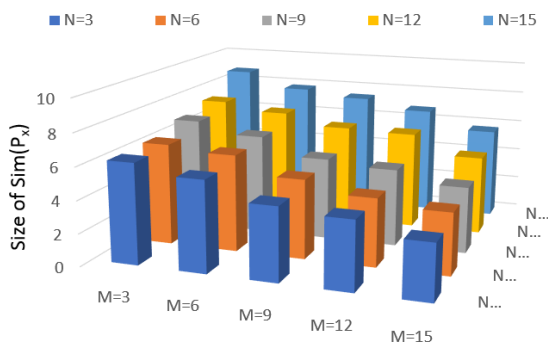


FIGURE 6. Similar individual volume of PPCP with M and N.

4) SIZE OF $Sim(P_x)$ IN PPCP

As we discussed in the last paragraph, parameters M and N determine the similar individual discovery condition to some extent. Therefore, M and N can also influence the number of similar individuals of a pre-designated individual. To evaluate such a relationship between the size of $Sim(P_x)$ and the parameters, we conduct an experiment whose results are reported in Fig.6.

The figure data show that there is a relatively clear variation tendency of the size of $Sim(P_x)$ when parameters M or N varies. In concrete, when M value increases, the size of $Sim(P_x)$ decreases as more hash functions often indicate a more rigid similar individual discovery condition; as a result,

size of $Sim(P_x)$ drops accordingly. On the contrary, when N value increases, the size of $Sim(P_x)$ rises as more hash tables often indicate a looser similar individual discovery condition; as a result, the size of $Sim(P_x)$ grows accordingly.

C. FURTHER DISCUSSIONS

In the experiments, we did not measure the performance of privacy guaranteeing due to the characteristics of hash technology [26]–[28]. In addition, only the prediction accuracy and efficiency are compared in the evaluation section. In our upcoming research, more prediction metrics (e.g., diversity [29]–[30], robustness [31], [32], and so on) should be added for better evaluating the performances of the suggested PPCP method.

VI. CONCLUSION

The existence of various medical data or healthy data has enabled the intelligent Internet of Health (IoH) services. As an example IoH application, we use collected sport questionnaires to investigate and monitor the population physiques. Meanwhile, we use hash techniques to secure the private data included in the sport questionnaires. Finally, we simulate a range of experiments to show that the proposed PPCP method is superior to other methods, especially for the RMSE and time cost.

To make the PPCP method more general and comprehensive, we will further refine PPCP to adapt multiple data categories and structures [33]. Besides, we will study the possibility of combining different privacy protection ways for advanced data security.

REFERENCES

- [1] S. Din and A. Paul, "Smart health monitoring and management system: Toward autonomous wearable sensing for Internet of Things using big data analytics," *Future Gener. Comput. Syst.*, vol. 111, p. 939, 2020.
- [2] N. C. Benda, T. C. Veinot, C. J. Sieck, and J. S. Ancker, "Broadband Internet access is a social determinant of health!," *Amer. J. Public Health*, vol. 110, no. 8, pp. 1123–1125, Aug. 2020.
- [3] E. Sillence, J. M. Blythe, P. Briggs, and M. Moss, "A revised model of trust in Internet-based health information and advice: cross-sectional questionnaire study," *J. Med. Internet Res.*, vol. 21, no. 11, Nov. 2019, Art. no. e11125.
- [4] K. Szulc and M. Duplaga, "The impact of Internet use on mental wellbeing and health behaviours among persons with disability," *Eur. J. Public Health*, vol. 29, no. 4, p. 425, Nov. 2019.
- [5] T. Peng, Y. Lin, X. Yao, and W. Zhang, "An efficient ranked multi-keyword search for multiple data owners over encrypted cloud data," *IEEE Access*, vol. 6, pp. 21924–21933, 2018.
- [6] H. Dai, Y. Ji, G. Yang, H. Huang, and X. Yi, "A privacy-preserving multi-keyword ranked search over encrypted data in hybrid clouds," *IEEE Access*, vol. 8, pp. 4895–4907, 2020.
- [7] T. V. Xuan Phuong, G. Yang, W. Susilo, F. Guo, and Q. Huang, "Sequence aware functional encryption and its application in searchable encryption," *J. Inf. Secur. Appl.*, vol. 35, pp. 106–118, Aug. 2017.
- [8] Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 340–352, Feb. 2016.
- [9] H. Ming, C. Mengmeng, and W. Xiaofei, "A collaborative filtering recommendation method based on differential privacy," *J. Comput. Res. Develop.*, vol. 54, no. 7, pp. 1439–1451, 2017.
- [10] W. Tong and H. Shubin, "An improved collaborative filtering recommendation algorithm with differentially privacy," *Inf. Secur. Technol.*, vol. 7, no. 4, pp. 26–28, 2016.

- [11] X. Zheng-zheng, L. Qi-liang, H. Xiao-yu, L. Ji-yuan, and L. Lei, "Differential privacy protection for collaborative filtering algorithms with explicit and implicit trust," *Acta Electronica Sinica*, vol. 46, no. 12, pp. 3050–3059, 2018.
- [12] C. Yin, L. Shi, R. Sun, and I. Jin Wang, "Improved collaborative filtering recommendation algorithm based on differential privacy protection," *J. Supercomput.*, vol. 76, pp. 5161–5174, Jan. 2020.
- [13] Y. Xiao, L. Xiong, S. Zhang, and Y. Cao, "LocLok: Location cloaking with differential privacy via hidden Markov model," *Proc. VLDB Endowment*, vol. 10, no. 12, pp. 1901–1904, Aug. 2017.
- [14] S. Yang, J. Xu, X. Yang, and X. Ren, "Bayesian network-based high-dimensional crowdsourced data publication with local differential privacy," *Scientia Sinica Informationis*, vol. 49, no. 12, pp. 1586–1605, Dec. 2019.
- [15] J. Wang, Z. Cai, Y. Li, D. Yang, J. Li, and H. Gao, "Protecting query privacy with differentially private k-anonymity in location-based services," *Pers. Ubiquitous Comput.*, vol. 22, no. 3, pp. 453–469, Jun. 2018.
- [16] S. Zhang, X. Li, Z. Tan, T. Peng, and G. Wang, "A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services," *Future Gener. Comput. Syst.*, vol. 94, pp. 40–50, May 2019.
- [17] F. Casino, J. Domingo-Ferrer, C. Patsakis, D. Puig, and A. Solanas, "A k-anonymous approach to privacy preserving collaborative filtering," *J. Comput. Syst. Sci.*, vol. 81, no. 6, pp. 1000–1011, Sep. 2015.
- [18] P. Zhao, J. Li, and F. Zeng, "ILLIA: Enabling k-anonymity-based privacy preserving against location injection attacks in continuous LBS queries," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1033–1042, Apr. 2018.
- [19] L. Qi, Q. He, F. Chen, X. Zhang, W. Dou, and Q. Ni, "Data-driven Web APIs recommendation for building Web applications," *IEEE Trans. Big Data*, early access, Feb. 24, 2020, doi: [10.1109/TBDDATA.2020.2975587](https://doi.org/10.1109/TBDDATA.2020.2975587).
- [20] C. Zhou, A. Li, A. Hou, Z. Zhang, Z. Zhang, P. Dai, and F. Wang, "Modeling methodology for early warning of chronic heart failure based on real medical big data," *Expert Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113361, doi: [10.1016/j.eswa.2020.113361](https://doi.org/10.1016/j.eswa.2020.113361).
- [21] J. Li, T. Cai, K. Deng, X. Wang, T. Sellis, and F. Xia, "Community-diversified influence maximization in social networks," *Inf. Syst.*, vol. 92, Sep. 2020, Art. no. 101522.
- [22] L. Wang, X. Zhang, R. Wang, C. Yan, H. Kou, and L. Qi, "Diversified service recommendation with high accuracy and efficiency," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106196, doi: [10.1016/j.knosys.2020.106196](https://doi.org/10.1016/j.knosys.2020.106196).
- [23] Y. Xu, C. Zhang, G. Wang, Z. Qin, and Q. Zeng, "A blockchain-enabled deduplicatable data auditing mechanism for network storage services," *IEEE Trans. Emerg. Topics Comput.*, early access, Jun. 29, 2020, doi: [10.1109/TETC.2020.3005610](https://doi.org/10.1109/TETC.2020.3005610).
- [24] W. Zhong, X. Yin, X. Zhang, S. Li, W. Dou, R. Wang, and L. Qi, "Multi-dimensional quality-driven service recommendation with privacy-preservation in mobile edge environment," *Comput. Commun.*, vol. 157, pp. 116–123, May 2020.
- [25] *Multi-Dimensional Quality-Driven Service Recommendation With Privacy-Preservation in Mobile Edge Environment*. Accessed: Apr. 14, 2020. [Online]. Available: <https://grouplens.org/datasets/movielens/>
- [26] L. Qi, X. Wang, X. Xu, W. Dou, and S. Li, "Privacy-aware cross-platform service recommendation based on enhanced locality-sensitive hashing," *IEEE Trans. Netw. Sci. Eng.*, early access, Jan. 27, 2020, doi: [10.1109/TNSE.2020.2969489](https://doi.org/10.1109/TNSE.2020.2969489).
- [27] L. Qi, C. Hu, X. Zhang, M. R. Khosravi, S. Sharma, S. Pang, and T. Wang, "Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment," *IEEE Trans. Ind. Informat.*, early access, Jul. 28, 2020, doi: [10.1109/TII.2020.3012157](https://doi.org/10.1109/TII.2020.3012157).
- [28] Q. Liu, P. Hou, G. Wang, T. Peng, and S. Zhang, "Intelligent route planning on large road networks with efficiency and privacy," *J. Parallel Distrib. Comput.*, vol. 133, pp. 93–106, Nov. 2019.
- [29] Y. Xu, J. Ren, Y. Zhang, C. Zhang, B. Shen, and Y. Zhang, "Blockchain empowered arbitrable data auditing scheme for network storage as a service," *IEEE Trans. Services Comput.*, vol. 13, no. 2, pp. 289–300, Mar./Apr. 2020.
- [30] L. Wang, X. Zhang, T. Wang, S. Wan, G. Srivastava, S. Pang, and L. Qi, "Diversified and scalable service recommendation with accuracy guarantee," *IEEE Trans. Comput. Social Syst.*, early access, Jul. 21, 2020, doi: [10.1109/TCSS.2020.3007812](https://doi.org/10.1109/TCSS.2020.3007812).
- [31] Q. Liu, G. Wang, F. Li, S. Yang, and J. Wu, "Preserving privacy with probabilistic indistinguishability in weighted social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 5, pp. 1417–1429, May 2017.
- [32] X. Chi, C. Yan, H. Wang, W. Rafique, and L. Qi, "Amplified LSH-based recommender systems with privacy protection," *Concurrency Comput., Pract. Exper.*, to be published, doi: [10.1002/CPE.5681](https://doi.org/10.1002/CPE.5681).
- [33] X. Xu, R. Mo, F. Dai, W. Lin, S. Wan, and W. Dou, "Dynamic resource provisioning with fault tolerance for data-intensive meteorological workflows in cloud," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6172–6181, Sep. 2020, doi: [10.1109/TII.2019.2959258](https://doi.org/10.1109/TII.2019.2959258).



CHUNLAN TIAN received the bachelor's degree from the College of Sport Science, Qufu Normal University, China, in 1987. She is currently an Associate Professor and the master's Supervisor of the College of Sport Science, Qufu Normal University. Her research interest includes sport education.



CHONGMIN ZHANG received the bachelor's degree from the College of Sport Science, Qufu Normal University, China, in 1985. He is currently a Full Professor and the master's Supervisor of the College of Sport Science, Qufu Normal University. His research interest includes sport education training theory and methods.



WANLI HUANG received the bachelor's and master's degrees from Shandong University, China, in 2004 and 2007, respectively. She is currently a Lecturer with the School of Information Science and Engineering, Qufu Normal University, China. Her research interests include big data processing and data privacy protection.



HAO WANG (Member, IEEE) received the B.Eng. and Ph.D. degrees in computer science and engineering from the South China University of Technology, Guangzhou, China, in 2006. He is currently an Associate Professor with the Norwegian University of Science and Technology, Gjøvik, Norway. He has authored or coauthored more than 130 articles in reputable international journals and conference papers. His current research interests include big data analytics, the Industrial Internet of Things, high-performance computing, and safety-critical systems. He is a member of the IEEE IES Technical Committee on Industrial Informatics. He has served as the TPC Co-Chair of the IEEE DataCom 2015, the IEEE CIT 2017, and ES 2017. He is a Reviewer of many journals, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and the IEEE TRANSACTIONS ON BIG DATA.

...