Wei *et al. EURASIP Journal on Wireless Communications and Networking* (2019) 2019:101
https://doi.org/10.1186/s13638-019-1432-2

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

# Privacy-aware cross-cloud service recommendations based on Boolean historical invocation records

Qiang Wei[1], Wenxue Wang[2], Gongxuan Zhang[1] and Tingting Shao[3*]

## Abstract

In the age of big data, service recommendation has provided an effective manner to filter valuable information from massive data. Generally, by observing the past service invocation records (Boolean values) distributed across different cloud platforms, a recommender system can infer personalized preferences of a user and recommend him/her new services to gain more profits. However, the historical service invocation records are a kind of private information for users. Therefore, how to protect sensitive user data distributed across multiple cloud platforms is becoming a necessity for successful service recommendations. Additionally, the historical service invocation records often update with time, which call for an efficient and scalable service recommendation method. In view of these challenges, we introduce the multi-probe Simhash technique in information retrieval domain into the recommendation process and further put forward a privacy-preserving recommendation method based on historical service invocation records. At last, we design several experiments on the real-world service quality data in set *WS-DREAM*. Experimental results show the feasibility of the proposal in terms of producing accurate recommended results while protecting users' private information contained in historical service invocation records.

**Keywords:** Service recommendation, Historical service invocation records, Simhash, Multi-probe, Privacy-preservation

## 1 Introduction

With the advent of the big data age, the volume and variety of available data both increase quickly, which make it hard for a user to select valuable information that matches his/her preferences [1–4]. Therefore, to decrease the heavy burden on users' service selection decisions, diverse service recommendation techniques are brought forth accordingly [5, 6]. Typically, through analyzing the service lists ever-executed or ever-invoked by historical users, a recommender system, such as the collaborative filtering (CF) recommender system, can infer the possible user preferences and find the users who are similar with a target user (i.e., the "friends" of the target user); afterward, appropriate new services are recommended to the target user according to the service list ever-executed by his/her similar friends. This way, users' decision-making cost on service selection declines significantly, with the help of lightweight recommender systems based on CF technique. Today, intelligent service recommendation techniques have already been successfully applied to various domains [7].

However, traditional service recommendation methods (e.g., CF-based methods) have several obvious shortcomings. First of all, we need to calculate the user similarity or item similarity in CF methods so as to find similar users or similar services for further recommendations. While the abovementioned similarity calculation process requires the historical service quality data (e.g., response time of a service invocation) generated from past service invocations, which may reveal the private information of users contained in the released service quality data, especially in the distributed environment [8–14] where the decision-making data for recommendations are distributed across multiple cloud platforms with conflict of interest. In this situation, how to utilize the less-sensitive Boolean service invocation records (i.e., whether a user

* Correspondence: Shaotingting03@163.com
[3]Medical Information Engineering School, Jining Medical University, Rizhao, China
Full list of author information is available at the end of the article

*u* has invoked a web service *ws*; typically, if *u* invoked *ws* in the past, then (*u*, *ws*) pair is equal to 1; otherwise, 0) instead of real service quality data containing much private information of users becomes a necessary but challenging task.

Second, the historical service invocation records often update frequently, which requires a recommended method to quickly output an optimal recommendation solution for a target user. While traditional CF methods often fail to deliver a quick and accurate recommended list as user similarity or item similarity needs to be calculated repeatedly when the historical service invocation records are updated frequently.

Considering these drawbacks, the Simhash technique that is popular in privacy-aware information retrieval is introduced into the recommendation domain. Afterward, we modify Simhash to be multi-probe Simhash (MPS) and bring forth a privacy-aware recommendation method based on MPS, named $Rec_{MPS}$, to avoid the probably returned too many or too few recommended results. Overall, our major contributions are summarized as below:

(1) We recognize a novel service recommendation scenario where the service recommendation decision-makings are not based on historical service quality values containing much user privacy but based on less-sensitive Boolean service invocation records stored in different cloud platforms.

(2) We modify Simhash technique into multi-probe Simhash, i.e., MPS and bring forth a privacy-preserving recommendation method based on MPS and historical service invocation records, i.e., $\mathrm{Rec}_{\mathrm{MPS}}$.

(3) Through experiments on WS-DREAM dataset, we show the major advantages of our proposal compared to other competitive work.

The organizational structure of our paper is summarized as below. In Section 2, the related work is briefly introduced. We formulate the privacy-preserving service recommendation problems based on historical service invocation records in Section 3. Section 4 presents the proposed recommendation method $\mathrm{Rec}_{\mathrm{MPS}}$ that is based on the multi-probe Simhash technique. Extensive experiments executed in Section 5 demonstrates the feasibility of $\mathrm{Rec}_{\mathrm{MPS}}$. At last, in Section 6, we conclude the whole paper and point out the prospective improvement directions in future research.

## 2 Related work

Next, we introduce the related work associated with this paper from the following two aspects: accuracy-oriented web service recommendations and privacy-aware web service recommendations.

### 2.1 Accuracy-oriented web service recommendations

Collaborative filtering is a promising recommendation technique widely adopted in existing recommender systems, e.g., item-based collaborative filtering [15]. Generally, the CF recommendation methods can first look for the users who are similar with a target user or look for the services which are similar to a target service, based on the historical service quality data; afterward, the appropriate services that may be preferred by the target user are filtered out and put into the final recommended list. As service execution environment is often not fixed but varied, the service quality data are also dynamic and dependent on the service running context, such as service running time and location information; considering this, several context-aware recommendation methods are proposed to enact more accurate recommendation decisions, e.g., location-based CF [16] and time-based CF [17]. While the abovementioned CF variants have an obvious shortcoming, i.e., they only recruit objective decision data (e.g., historical service quality) for recommendations, while neglecting other key factors that may affect a target user's recommendation decision-makings, e.g., users' personalized preferences. In view of this shortcoming, CF recommendation methods are improved in [18] by considering the preferences of users, to support personalized and preference-aware service selection decisions of different users.

However, in the above CF-based recommendation variants, the data for recommendation decisions (e.g., historical service quality) are often supposed to be centralized. Namely, existing methods seldom consider the multi-source recommendation problems in which the data are distributed. Furthermore, existing CF-based recommendation methods seldom consider the privacy leakage risks when the multi-source decision-making data are needed to be integrated or fused together for more comprehensive and rational service recommendations.

### 2.2 Privacy-preserving web service recommendations

Privacy-preservation is a crucial goal in a distributed environment. To achieve the goal of privacy protection, in [19], a service user is advised to release only partial service quality data observed by him/her; thus, most private information of users are protected and secure. There is often a trade-off between the amount of published service quality data and the accuracy of final recommended results, which is observed and studied in [20]. Furthermore, the authors regard the size of the released data as a variable parameter and convert the sensitive recommendation problem into one about the parameter successfully. However, in the above method, the private information contained in the published data is still at risk and may be exposed to the outside world.

Data encryption is a popular way to achieve the privacy-preservation goal in recommendations. However, as a heavyweight data protection mechanism, encryption may bring heavy computational cost and transmission delay, and hence is often not applicable to the recommendation scenarios with little computational capabilities [21, 22]. K-anonymity strategies are employed in [23] to hide the real data and make sure that user privacy is secure. However, the anonymous data are often of low availability and therefore, the accuracy of recommended results is decreased accordingly.

In order to solve the above challenge, in [24], the data are firstly obfuscated and then the obfuscated data are recruited to recommend appropriate new service items to the target users; this way, the sensitive information of the target users are secure. However, the accuracy of recommended results is not as high as expected; this is because the obfuscated data, not real data, are employed to make recommendation decisions. In [25], the sensitive QoS data are firstly split into multiple pieces (i.e., QoS pieces); and then the QoS pieces are sent to different users for storage; finally, the less-sensitive QoS pieces are utilized as the service recommendation bases. This method can achieve a partial privacy-preservation goal in service recommendation; however, it still fails to protect other key user privacy information, e.g., the set of web services that were executed by different users in the past.

Differential privacy (DP) is regarded as an effective way for privacy-aware service recommendation in [26]. Typically, through DP, the real service quality data are first confused by the injected noise data and then sent to recommender systems for decision-makings. Thus, the sensitive information contained in the real service quality data can be protected very well. While the time costs of DP-based recommendation solutions are generally high; therefore, the time delay is often high enough. Besides, the accuracy of the returned recommended list would be reduced to some extent when the accumulated noise data after DP are high.

Locality-sensitive hashing (LSH) technique is employed in work [8, 9, 17, 21, 22] to protect the sensitive QoS values generated from historical invocations. However, these work focus more on protecting the historical QoS values (typically, continuous values) instead of the historical service invocation records (Boolean values) that we focus on in this paper.

Considering the drawbacks of existing methods, a privacy-preserving recommendation method named $\text{Rec}_{\text{MPS}}$ is suggested, which will be specified in detail in the rest of this paper.

## 3 Formulation

For a better understanding of our proposal, we introduce the symbols to be used in the subsequent paragraphs. Concretely, we assume that there are $m$ users $\{u_1, \ldots, u_m\}$ and $n$ web services $\{ws_1, \ldots, ws_n\}$; $u^*$ is a target user waiting for a recommended list from the recommender system; historical service invocation records are depicted by $r_{i,j}$ $(1 \le i \le m,\ 1 \le j \le n)$ that indicates whether $u_i$ has invoked $ws_j$ in the past: $r_{i,j} = 1$ if the answer is yes, otherwise $r_{i,j} = 0$.

Compared to the real service quality of services invoked by historical users, the service invocation records (i.e., the Boolean values $r_{i,j}$) are less sensitive. However, we argue that the values of $r_{i,j}$ $(1 \le i \le m,\ 1 \le j \le n)$ are still a kind of private information for users. Therefore, the problem that we need to solve in this paper can be clarified as below: a recommender system needs to analyze the preferences of $u^*$ and make appropriate recommendations to $u^*$ based on the $r_{i,j}$ values produced from past service invocations without revealing the real $r_{i,j}$ values. In the next section, we will introduce our resolution to this problem.

## 4 Service recommendation based on multi-probe Simhash

Next, we will detail the concrete algorithm of $\text{Rec}_{\text{MPS}}$. The general idea of $\text{Rec}_{\text{MPS}}$ is first, we create less sensitive user indices based on historical service invocation records and Simhash (as Simhash is more suitable for protecting the Boolean data compared to other hash variants), and find neighbors of target user $u^*$ based on user indices; second, to avoid too few (even null) or too many returned neighbors of $u^*$, multi-probe Simhash strategy is adopted; finally, optimal recommendations are made to $u^*$ based on the returned neighbors of $u^*$. Next, we will elaborate on the concrete process of $\text{Rec}_{\text{MPS}}$.

### 4.1 Step 1: create less sensitive user indices and find neighbors of target user $u^*$ based on user indices

Historical service invocation records of $n$ service $\{ws_1, \ldots, ws_n\}$ by $m$ users $\{u_1, \ldots, u_m\}$ can be represented by the matrix in (1), where $r_{i,j}$ is a Boolean value indicating whether $u_i$ has invoked $ws_j$ in the past. Thus, each row vector $(r_{i,1}, \ldots, r_{i,n})$ denotes the historical service invocation records of user $u_i$. As a service community often contains a large number of web services, i.e., $n$ is large, vector $(r_{i,1}, \ldots, r_{i,n})$ for user $u_i$ is often high-dimensional and hence requires much computational time when $(r_{i,1}, \ldots, r_{i,n})$ takes part in the subsequent service recommendation process. Therefore, to reduce the time cost, Simhash technique is employed to convert the high-dimensional vector $(r_{i,1}, \ldots, r_{i,n})$ for $u_i$ into a low-dimensional vector for $u_i$, i.e., $(R_{i,1}, \ldots, R_{i,p})$ where $p = \lceil \log_2^n \rceil$ holds.

$$\begin{bmatrix} r_{1,1} & \cdots & r_{1,n} \\ \vdots & \ddots & \vdots \\ r_{m,1} & \cdots & r_{m,n} \end{bmatrix} \tag{1}$$

Next, we introduce the concrete conversion process. Each of the $n$ services $\{ws_1, \ldots, ws_n\}$ is recoded according to binary code (the number of 0/1 bits is equal to $p$). For example, $ws_1 = (0, 0, \ldots, 0, 0, 1)$, $ws_2 = (0, 0, \ldots, 0, 1, 0)$, $ws_3 = (0, 0, \ldots, 0, 1, 1)$, and so on. Assume that $u_i$ has invoked $n_1$ services ($n_1 \leq n$), then we pick these $n_1$ services as well as their binary codes to form an $n_1{*}p$ matrix constituted by 0 and 1. For example, if there are totally 30 candidate web services (here, $p = \lceil \log_2^{30} \rceil = 5$) and $u_i$ has invoked $ws_1$ and $ws_3$, then we can derive a 2*5 0/1 matrix in (2). Next, we substitute "− 1" for the element "0" in (2). Thus, we can obtain another 2*5 matrix in (3) where each entry is either − 1 or 1.

$$u_i\text{:} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} (2).$$

$$u_i\text{:} \begin{bmatrix} -1 & -1 & -1 & -1 & 1 \\ -1 & -1 & -1 & 1 & 1 \end{bmatrix} (3).$$

For the − 1/1 matrix in (3), we calculate the sum of each column and then obtain a 5-dimensional vector $H(u_i) = (-2, -2, -2, 0, 2)$. Afterward, in vector $H(u_i)$, we substitute "0" for the negative entries and substitute "1" for the positive entries. Then, we obtain a new 5-dimensional 0/1 vector (0, 0, 0, 0, 1), which can be considered as the index for user $u_i$, denoted by $h(u_i)$. Here, index $h(u_i)$ has two advantages: first, $h(u_i)$ is less sensitive as it contains little even no private information of user $u_i$; second, $h(u_i)$ is a low-dimensional vector $(R_{i,1}, \ldots, R_{i,p})$ compared to the original high-dimensional vector $(r_{i,1}, \ldots, r_{i,n})$ for user $u_i$.

Next, with the user indices $h(u_i)$ $(1 \leq i \leq m)$, we can look for the similar users (i.e., neighbors) of target user $u^*$. Concretely, if index values $h(u_i) = h(u^*)$ holds, then $u_i$ is deemed as a qualified neighbor of $u^*$ with high probability according to the Simhash theory.

### 4.2 Step 2: improved neighbor search for target user *u** based on multi-probe Simhash

In Step 1, neighbors of target user $u^*$ can be discovered and returned for recommendation decision-makings based on Simhash technique. However, the neighbor search condition in Step 1, i.e., $h(u_i) = h(u^*)$ cannot always work well as it is probably too loose or too tight in certain situations. Concretely, if the condition $h(u_i) = h(u^*)$ is too loose, then too many neighbors of target user $u^*$ can be returned, which may reduce the recommendation accuracy to some extent; otherwise, if the condition $h(u_i) = h(u^*)$ is too tight, then few (even null) neighbors of target user $u^*$ will be returned, which may decrease the recommendation feasibility. In other words, the traditional Simhash technique needs to be improved or modified to avoid the probably returned too many or too few (even null) neighbors of $u^*$.

Next, we improve the traditional Simhash technique to be multi-probe Simhash. Concretely, if the neighbor search condition $h(u_i) = h(u^*)$ is too loose, then we will tighten it; otherwise, if the neighbor search condition $h(u_i) = h(u^*)$ is too tight, then we will loosen it to some extent.

#### *4.2.1 Case 1: search condition relaxation*

The neighbor search condition $h(u_i) = h(u^*)$ introduced in Step 1 is probably too rigid or tight in certain situations and thereby finds too few (even null) neighbors of the target user $u^*$. In this situation, we need to relax the too tight neighbor search condition $h(u_i) = h(u^*)$ so that the number of returned neighbors of $u^*$ can exceed the pre-defined threshold $P$.

Next, we elaborate on the concrete condition relaxation process. Suppose $h(u_i) = (R_{i,1}, \ldots, R_{i,p})$ and $h(u^*) = (R_{*,1}, \ldots, R_{*,p})$, then $h(u_i) \oplus h(u^*)$ can be defined as in (4). Thus the original neighbor search condition that is too tight, i.e., $h(u_i) = h(u^*)$ can be converted into another condition $h(u_i) \oplus h(u^*) = 0$. Therefore, we can relax the neighbor search condition to be $h(u_i) \oplus h(u^*) = 1$ or 2 or 3 or … or $p$, depending on the number of returned neighbors of $u^*$ according to the neighbor search condition. At last, the returned neighbors of $u^*$ are put into set *Neig_Set*.

$$h(u_i) \oplus h(u^*).$$
$$= (R_{i,1} \oplus R_{*,1}) + (R_{i,2} \oplus R_{*,2}) + \ldots + (R_{i,p} \oplus R_{*,p}) \text{ (4).}$$

#### *4.2.2 Case 2: search condition tightness*

In Step 1, user index $h(u_i)$ is a super simplification (i.e., coarse-grained expression) of the historical service invocation records of user $u_i$, e.g., $h(u_i) = (0, 0, 0, 0, 1)$ holds in the example of Step 1. While coarse-grained $h(u_i)$ may lead to too loose search condition (i.e., $h(u_i) = h(u^*)$) for the neighbors of target user $u^*$. Considering this drawback, we use relatively fine-grained index for $u_i$, i.e., $H(u_i)$ (in the example of Step 1, $H(u_i) = (-2, -2, -2, 0, 2)$ holds) to replace coarse-grained $h(u_i)$ so as to tighten the search condition and produce fewer neighbors of target user $u^*$.

Concretely, if $H(u_i) = H(u^*)$ holds, we can reach a conclusion that $u_i$ and $u^*$ are similar users because $H(u_i) = H(u^*)$ is a tighter neighbor search condition compared to the original condition $h(u_i) = h(u^*)$. Therefore, through $H(u_i) = H(u^*)$, we can expect to obtain fewer but more similar neighbors of $u^*$. However, if condition $H(u_i) = H(u^*)$ is too tight, then an appropriate relaxation is necessary. Concretely, we do not expect $H(u_i) = H(u^*)$ (i.e., $H(u_i) \oplus H(u^*) = 0$) but expect the result of xor operation $H(u_i) \oplus H(u^*)$ is close to 0. This way, we can relax the neighbor search condition if $H(u_i) = H(u^*)$ is too tight. Concrete condition relaxation degree denoted by the value of $H(u_i) \oplus H(u^*)$ depends on the pre-defined threshold $P$ of the number of $u^*$'s neighbors. At last, the returned neighbors of $u^*$ are put into set *Neig_Set*.

### 4.3 (3) Step 3: recommend new services to target user *u** through returned neighbors in Neig_Set

For each user $u_i$ in Neig_Set, if he or she has invoked candidate service $ws_j$ $(1 \le j \le n)$ in the past, i.e., $r_{i,j} = 1$, then $u_i$ is put into a new set Neig_Set*; furthermore, $ws_j$'s historical quality value by $u_i$ (denoted by $q_{i,j}$) can be used to predict the missing quality value of $ws_j$ by the target user $u$* (denoted by $q_{*,j}$), based on the prediction equation in (5), where | Neig_Set* | is the size of set Neig_Set*.

$$q_{*,j} = \frac{1}{|\text{Neig\_Set}^*|} * \sum_{u_i \in \text{Neig\_Set}*} q_{i,j} \ (5).$$

Thus, for each candidate service $ws_j$ $(1 \le j \le n)$ that has never been executed by the target user $u$*, its missing quality value invoked by $u$*, i.e., $q_{*,j}$ can be predicted by (5). Finally, we select one candidate service with the optimal predicted value $q_{*,j}$ and recommend it to $u$*. This is the end of our suggested recommendation method $\text{Rec}_{\text{MPS}}$.

## 5 Experiments

### 5.1 Experiment configurations

To demonstrate the feasibility of $\text{Rec}_{\text{MPS}}$ method, we deploy extensive experiments with popular WS-DREAM dataset [12]. This dataset contains the historical QoS data of 4532 services collected by 142 users. We extract the Boolean (user, service) pairs from these QoS data for experiment purpose. To show the advantages of our solution, we compare $\text{Rec}_{\text{MPS}}$ method with three state-of-the-art methods, e.g., $\text{DistSR}_{\text{LSH}}$ [27], WSRec [28], and ICF (item-based CF). We compare the performances of the four methods in terms of recommendation accuracy (via RMSE) and efficiency. Each set of experiments is executed 100 times and we record their average results. The experiment hardware and software configurations are as follows: 2.80 GHz processor, 8.0 GB RAM, Windows 10, and JAVA 8.

### 5.2 Experiment results

#### *5.2.1 Profile 1: accuracy comparison with competitive methods*

Through comparing the predicted QoS values and real QoS values, we can test the accuracy of the recommended results for different methods. Here, we use RMSE (the smaller the better) to measure the accuracy. The parameters are set as follows: the size of user set, i.e., $m = 142$, the size of service set, i.e., $n = \{500, 1000, 2000, 3000, 4000\}$, threshold $P = 3$. Experiment results are demonstrated in Fig. 1.

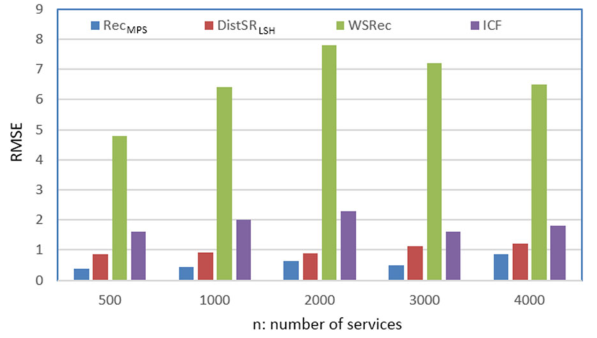

**Fig. 1** Accuracy comparisons (w.r.t. *n*) ($m = 142$). $\text{DistSR}_{\text{LSH}}$, $\text{Rec}_{\text{MPS}}$, WSRec, ICF

As Fig. 1 shows, the RMSE value of $\text{Rec}_{\text{MPS}}$ is smaller than those of the rest three methods, which indicates that $\text{Rec}_{\text{MPS}}$ can achieve higher recommendation performance in terms of accuracy. The reason is (1) WSRec and ICF are mainly collaborative filtering-based neighbor search methods and hence cannot avoid too many or too few (even null) returned neighboring users or neighboring services, while too many or too few (even null) returned neighbors for recommendation decision-makings may fluctuate or decrease the recommendation accuracy; (2) $\text{DistSR}_{\text{LSH}}$ is more suitable for protecting the QoS values that are real number instead of the historical service invocation records (Boolean values) that we focus on in this paper. While our proposed $\text{Rec}_{\text{MPS}}$ method is designed for protecting the Boolean service invocation records that are sensitive to users; therefore, high recommendation accuracy can be guaranteed.

#### *5.2.2 Profile-2: efficiency comparison with competitive methods*

Next, we test the time costs of different methods as efficiency and scalability are also important factors that influence the recommendation performances and user satisfaction. Concrete parameters are set as follows: $m = 142$, $n = \{500, 1000, 2000, 3000, 4000\}$, threshold $P = 3$. Experiment results are presented in Fig. 2.

As can be observed from Fig. 2, the time costs of four methods approximately rise with the growth of $n$; this is because more computational time is needed to find out an optimal service that is preferred by the target user from a bigger volume of candidate services. Another observation is that $\text{DistSR}_{\text{LSH}}$ and $\text{Rec}_{\text{MPS}}$ methods outperform WSRec and ICF methods in terms of efficiency and scalability as additional privacy-preservation strategies are recruited in the former two methods. Although $\text{Rec}_{\text{MPS}}$ does not perform better than $\text{DistSR}_{\text{LSH}}$ in terms of efficiency, the time cost of $\text{Rec}_{\text{MPS}}$ is still acceptable in most cases (generally, smaller than 1 s as shown in Fig. 2).

#### *5.2.3 Profile 3: accuracy of $\text{Rec}_{\text{MPS}}$ with respect to n and P*

We investigate the relationship between accuracy of $\text{Rec}_{\text{MPS}}$ method and parameters $n$ and $P$. Parameter

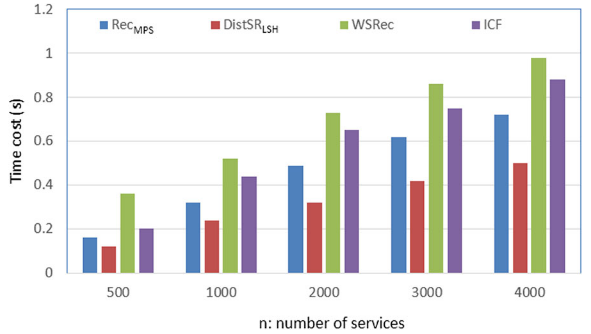

**Fig. 2** Efficiency comparisons (w.r.t. *n*) ($m = 142$)**.** DistSR$_{LSH}$, Rec$_{MPS}$, WSRec, ICF

settings are listed as below: $m = 142$, $n = \{500, 1000, 2000, 3000, 4000\}$, threshold $P = \{2, 4, 6, 8, 10\}$. Experiment results are demonstrated in Fig. 3. As reported in the figure, the RMSE values drop (i.e., accuracy values rise) approximately with the growth of $n$; this is because more candidate services often mean a higher probability of finding an optimal service that fits the preferences of a target user. Another observation from Fig. 3 is the RMSE values rise (i.e., accuracy values drop) approximately with the increment of $P$; the reason is probably as below: a larger threshold $P$, i.e., more returned neighbors of a target user may decrease the similarity between the target user and his/her neighbors and thereby reduce the recommendation accuracy.

### 5.2.4 Profile 4: accuracy of Rec$_{MPS}$ with respect to n and P

This profile tests the relationship between time cost of Rec$_{MPS}$ method and parameters $n$ and $P$. Parameter settings are listed as below: $m = 142$, $n = \{500, 1000, 2000, 3000, 4000\}$, threshold $P = \{2, 4, 6, 8, 10\}$. Experiment results are reported in Fig. 4. As Fig. 4 indicates, the efficiency of Rec$_{MPS}$ generally decreases with the growth of $n$ and $P$, as more candidate services or more returned neighbors of a target user often bring additional computational time to find the optimal service for recommendations.

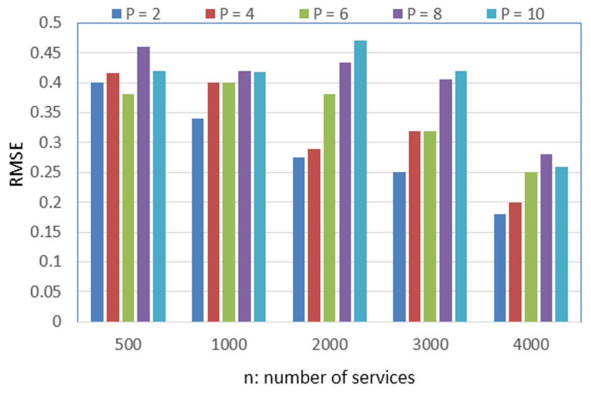

**Fig. 3** Accuracy of Rec$_{MPS}$ w.r.t. *n* and *P* ($m = 142$)**.** $P = 2$, $P = 4$, $P = 6$, $P = 8$, $P = 10$

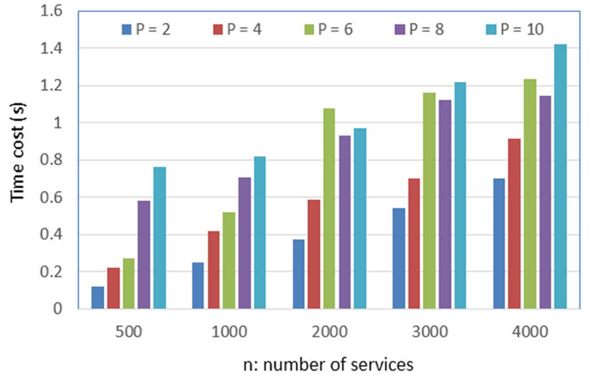

**Fig. 4** Efficiency of Rec$_{MPS}$ w.r.t. *n* and *P* ($m = 142$). $P = 2$, $P = 4$, $P = 6$, $P = 8$, $P = 10$

### 5.3 Further discussions

In this subsection, we discuss more details about the experiments and results.

(1) In subsection 5.A, only a QoS dimension (Boolean values) is recruited. However, we argue that our method can be easily extended to the more complex application scenarios with multiple dimensions [29–38] as well as their weights [39–45] by repeating the Simhash-based index building process multiple times.

(2) In the experiment test, only one type of decision-making data (i.e., Boolean service invocation records) is considered to judge whether two users are similar neighbors. In the future, we will extend our method to accommodate the diversity of data types in the big data environment, e.g., discrete values [46–52], continuous values [53–57], fuzzy values [58], and so on.

(3) In Rec$_{MPS}$ method, the sensitive service invocation records (Boolean values) are converted into less-sensitive user indices based on Simhash technique, through which the privacy-preserving recommendation goal is achieved. However, the privacy-preservation effects of Rec$_{MPS}$ method are not quantified in the experiments due to the inherent shortcoming of Simhash. Therefore, further attention should be paid to the quantitative measurement and analyses of privacy-preservation capability of our proposal.

(4) In the experiments, we only test the recommendation accuracy (through RMSE) and efficiency separately. However, as the experiment results in Fig. 3 and Fig. 4 show, there is a trade-off relationship between these two key recommendation criteria. Therefore, it is necessary to investigate the trade-off between them in the future. Moreover, there may exist a three-party trade-off relationship among the recommendation accuracy, time

cost, and capability of privacy-preservation, which still need further analyses in future experiments.

## 6 Conclusions and future work

Recommender systems have become a promising tool to help people to quickly extract valuable information from big data. Typically, through analyzing the Boolean service invocation records, a collaborative recommender system can find and recommend appropriate new services to a target user. However, existing methods focus more on recommendation accuracy or efficiency, while often fail to protect the sensitive information contained in Boolean service invocation records distributed across different cloud platforms. Inspired by this observation, we introduce multi-probe Simhash strategy in the information retrieval domain into the recommendation process and bring forth a privacy-aware recommendation method based on historical service invocation records (Boolean values), i.e., $\text{Rec}_{\text{MPS}}$. At last, we design several experiments on the popular service quality dataset, i.e., WS-DREAM. Experimental findings show the advantages of $\text{Rec}_{\text{MPS}}$ compared with other competitive methods.

However, there are several shortcomings in $\text{Rec}_{\text{MPS}}$ method. First, for simplicity, we only consider one dimension for recommendation decision-makings, i.e., Boolean service invocation records; in future research, we will extend $\text{Rec}_{\text{MPS}}$ method by including more dimensions and possible linear correlations [59–61] and non-linear correlations [62–69]. Besides, the recommendation basis in $\text{Rec}_{\text{MPS}}$ is unique, i.e., Boolean service invocation records. Therefore, we will improve $\text{Rec}_{\text{MPS}}$ method by integrating the diverse recommendation data. At last, $\text{Rec}_{\text{MPS}}$ cannot always make successful recommendations as Simhash is a probability-based technique; therefore, we will tackle this issue in future research work.

### Abbreviations

CF: Collaborative filtering; DP: Differential privacy; MPS: Multi-probe Simhash; QoS: Quality of service

### Acknowledgements

Not applicable.

### Funding

Part work of this paper is supported by the Research on Data Mining Technology for the Healthcare Big Data (No. JY2016KJ034Y) and the Natural Science Foundation of China (No. 61872219).

### Availability of data and materials

The recruited experiment dataset WS-DREAM is available at wsdream.github.io/.

### Authors' contributions

QW and GZ finished the algorithm and wrote the paper in English. WW finished the experiments. TS put forward the idea of this paper. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. [2]Section of Information, Chinese Service Center for Scholarly Exchange (CSCSE), Beijing, China. [3]Medical Information Engineering School, Jining Medical University, Rizhao, China.

Received: 1 March 2019 Accepted: 10 April 2019
Published online: 25 April 2019

### References

1. X. Wang et al., A cloud-edge computing framework for cyber-physical-social services. IEEE Commun Mag **55**(11), 80–85 (2017)
2. L. Qi et al., Weighted principal component analysis-based service selection method for multimedia services in cloud. Computing **98**(1), 195–214 (2016)
3. L. Ren et al., Multi-scale dense gate recurrent unit networks for bearing remaining useful life prediction. Futur Gener Comput Syst **94**, 601–609 (2018)
4. L. Qi et al., "Time-location-frequency"-aware internet of things service selection based on historical records. International Journal of Distributed Sensor Networks **13**(1), 1–9 (2017)
5. X. Wang et al., A tensor-based big data-driven routing recommendation approach for heterogeneous networks. IEEE Network. https://doi.org/10.1109/MNET.2018.1800192
6. Y. Xu et al., Privacy-preserving and scalable service recommendation based on simhash in a distributed cloud environment. Complexity **2017**, 3437854–3437859 (2017)
7. L. Qi et al., Data-sparsity tolerant web service recommendation approach based on improved collaborative filtering. IEICE Transactions on Information and Systems, vol E100D **9**, 2092–2099 (2017)
8. C. Yan et al., Privacy-aware data publishing and integration for collaborative service recommendation. IEEE ACCESS **6**, 43021–43028 (2018)
9. L. Qi et al., A two-stage locality-sensitive hashing based approach for privacy-preserving mobile service recommendation in cross-platform edge environment. Futur Gener Comput Syst **88**, 636–643 (2018)
10. X. Wang et al., A big data-as-a-service framework: State-of-the-art and perspectives. IEEE Transactions on Big Data **4**(3), 325–340 (2018)
11. W. Tang et al., An offloading method using decentralized P2P-enabled mobile edge servers in edge computing. J Syst Archit **94**, 1–13 (2019)
12. Z. Zheng et al., Investigating qos of real world web services. IEEE Trans Serv Comput **7**(1), 32–39 (2014)
13. X. Wang et al., A distributed HOSVD method with its incremental computation for big data in cyber-physical-social systems. IEEE Transactions on Computational Social Systems **5**(2), 481–492 (2018)
14. L. Qi et al., Time-aware IoE service recommendation on sparse data. Mob Inf Syst **2016**, 4397061–4397012 (2016)
15. K. Chung et al., Categorization for grouping associative items using data mining in item-based collaborative filtering. Multimed Tools Appl **71**(2), 889–904 (2014)
16. X. Wang et al., A spatial-temporal QoS prediction approach for time-aware web service recommendation. *ACM Transactions on the Web* **10**(1, Article No. 7), 1–25 (2016)
17. L. Qi et al., Time-aware distributed service recommendation with privacy-preservation. Inf Sci **480**, 354–364 (2019)
18. K.K. Fletcher et al., A collaborative filtering method for personalized preference-based service recommendation. IEEE International Conference on Web Services, 400–407 (2015)
19. W. Dou et al., HireSome-II: Towards privacy-aware cross-cloud service composition for big data applications. IEEE Transactions on Parallel and Distributed Systems **26**(2), 455–466 (2015)
20. X. Zheng et al., Location-privacy-aware review publication mechanism for local business service systems. IEEE International Conference on Computer Communications, 1–9 (2017)

21. W. Gong et al., Privacy-aware multidimensional mobile service quality prediction and recommendation in distributed fog environment. Wirel Commun Mob Comput **2018**, 3075849–3075848 (2018)
22. L. Qi et al., A distributed locality-sensitive hashing based approach for cloud service recommendation from multi-source data. IEEE Journal on Selected Areas in Communications **35**(11), 2616–2624 (2017)
23. T. Ma et al., KDVEM: A k-degree anonymity with vertex and edge modification algorithm. Computing **70**(6), 1336–1344 (2015)
24. J. Zhu et al., A privacy-preserving qos prediction framework for web service recommendation. IEEE International Conference on Web Services, 241–248 (2015)
25. D. Li et al., An algorithm for efficient privacy-preserving item-based collaborative filtering. Futur Gener Comput Syst **55**, 311–320 (2016)
26. C. Li et al., Differentially private trajectory analysis for points-of-interest recommendation. IEEE International Congress on Big Data, 49–56 (2017)
27. L. Qi et al., Privacy-preserving distributed service recommendation based on locality-sensitive hashing. IEEE International Conference on Web Services, 49–56 (2017)
28. Z. Zheng et al., QoS-aware web service recommendation by collaborative filtering. IEEE Trans Serv Comput **4**(2), 140–152 (2011)
29. M. Wang et al., Robust group non-convex estimations for high-dimensional partially linear models. Journal of Nonparametric Statistics **28**(1), 49–67 (2016)
30. X. Wang et al., Variable selection for high-dimensional generalized linear models with the weighted elastic-net procedure. J Appl Stat **43**(5), 796–809 (2016)
31. P. Wang et al., Some geometrical properties of convex level sets of minimal graph on 2-dimensional riemannian manifolds. Nonlinear Anal **130**, 1–17 (2016)
32. X. Wang et al., Adaptive group bridge estimation for high-dimensional partially linear models. J Inequal Appl **2017**(158), 1–18 (2017)
33. X. Wang et al., Restricted profile estimation for partially linear models with large-dimensional covariates. Statistics & Probability Letters **128**, 71–76 (2017)
34. H. Tian et al., Bifurcation of periodic orbits by perturbing high-dimensional piecewise smooth integrable systems. J Differ Equ **263**, 7448–7474 (2017)
35. P. Wang et al., The geometric properties of harmonic function on 2-dimensional Riemannian manifolds. Nonlinear Anal **103**, 2–8 (2014)
36. M. Wang et al., Adaptive lasso estimators for ultrahigh dimensional generalized linear models. Statistics & Probability Letters **89**, 41–50 (2014)
37. M. Wang et al., A note on the one-step estimator for ultrahigh dimensionality. J Comput Appl Math **260**, 91–98 (2014)
38. G. Tian et al., Variable selection in the high-dimensional continuous generalized linear model with current status data. J Appl Stat **41**, 467–483 (2014)
39. S. Yang et al., The weight distributions of two classes of p-ary cyclic codes with few weights. Finite Fields and Their Applications **44**, 76–91 (2017)
40. Y. Wang et al., Uniform estimate for the tail probabilities of randomly weighted sums. Acta Mathematicae Applicatae Sinica **30**(4), 1063–1072 (2014)
41. S. Yang et al., A class of three-weight linear codes and their complete weight enumerators. Cryptogr Commun **9**, 133–149 (2017)
42. J. Cai, An implicit sigma (3) type condition for heavy cycles in weighted graphs. Ars Combinatoria **115**, 211–218 (2014)
43. S. Yang et al., Complete weight enumerators of a family of three-weight linear codes. Des Codes Crypt **82**, 663–674 (2017)
44. S. Yang et al., Complete weight enumerators of a class of linear codes. Discret Math **340**, 729–739 (2017)
45. S. Yang et al., A construction of linear codes and their complete weight enumerators. Finite Fields and Their Applications **48**, 196–226 (2017)
46. H. Liu et al., Some new generalized volterra-fredholm type discrete fractional sum inequalities and their applications. J Inequal Appl **2016**(213) (2016). https://doi.org/10.1186/s13660-016-1152-7
47. P. Li et al., Some classes of equations of discrete type with harmonic singular operator and convolution. Appl Math Comput **284**, 185–194 (2016)
48. P. Li, Singular integral equations of convolution type with hilbert kernel and a discrete jump problem. Advances in Difference Equations **2017**, 360. https://doi.org/10.1186/s13662-017-1413-x
49. Y. Wang et al., Approximation for the ruin probabilities in a discrete time risk model with dependent risks. Statistics & Probability Letters **80**(17–18), 1335–1342 (2010)
50. Q. Feng et al., Generalized Gronwall-bellman-type discrete inequalities and their applications. *Journal of Inequalities and Applications* 2011, 47 (2011)
51. P. Li, Two classes of linear equations of discrete convolution type with harmonic singular operators. Complex Variables and Elliptic Equations **61**(1), 67–75 (2016)
52. Z. Zheng, Invariance of deficiency indices under perturbation for discrete Hamiltonian systems. Journal of Difference Equations and Applications **19**(8), 1243–1250 (2013)
53. C. Hou et al., Continuity of (α,β)-derivations of operator algebras. Journal of the Korean Mathematical Society **48**(4), 823–835 (2011)
54. C.Q. Ma et al., On formability of linear continuous multi-agent systems. J Syst Sci Complex **25**(1), 13–29 (2012)
55. H. Wu et al., Continuous dependence property of BSDE with constraints. Appl Math Lett **45**, 41–46 (2015)
56. L.L. Liu, Continued fractions and the derangement polynomials of types A and B. Ars Combinatoria **125**, 321–330 (2016)
57. H. Feng, The modulus of continuity theorem for G-Brownian motion. Communications in Statistics-Theory and Methods **46**(7), 3586–3598 (2017)
58. L. Wang, The fixed point method for intuitionistic fuzzy stability of a quadratic functional equation. Fixed Point Theory and Applications **2010**, Article ID 107182, 1–7 (2010)
59. G. Guo et al., Parallel tempering for dynamic generalized linear models. Communications in Statistics-Theory and Methods **45**(21), 6299–6310 (2016)
60. L.L. Liu et al., Recurrence relations for linear transformations preserving the strong q-log-convexity. Electron J Comb **23**(3), 1–11 (2016)
61. H. Li et al., Partial condition number for the equality constrained linear least squares problem. Calcolo **54**(4), 1121–1146 (2017)
62. H. Liu et al., Some new nonlinear integral inequalities with weakly singular kernel and their applications to FDEs. J Inequal Appl **2015**(209), 1–17 (2015)
63. X. Zhang et al., Entire large solutions for a Schrödinger systems with a nonlinear random operator. J Math Anal Appl **423**(2), 1650–1659 (2015)
64. Z. Zong et al., On Jensen's inequality, Holder's inequality and Minkowski's inequality for dynamically consistent nonlinear evaluations. J Inequal Appl **2015**(152), 1–18 (2015)
65. X. Hao et al., Positive solutions for nonlinear fractional semipositone differential equation with nonlocal boundary conditions. Journal of Nonlinear Science and Applications **9**(6), 3992–4002 (2016)
66. X. Hao et al., Iterative solution for nonlinear impulsive advection-reaction-diffusion equations. Journal of Nonlinear Science and Applications **9**(6), 4070–4077 (2016)
67. F. Li et al., Global existence uniqueness and decay estimates for nonlinear viscoelastic wave equation with boundary dissipation. Nonlinear Analysis: Real World Applications **12**, 1770–1784 (2011)
68. Y. Wang et al., Positive solutions for a class of fractional boundary value problem with changing sign nonlinearity. Nonlinear Anal **74**(17), 6434–6441 (2011)
69. J. Jiang, L. Liu, Y. Wu. Positive solutions to nonlinear fractional differential equations involving Stieltjes integrals conditions. Journal of Nonlinear Sciences and Applications **10**, 5351–5359 (2017)