Awareness and Control of Personal Data Based on the Cyber-I Privacy Model

Tang, Li

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
10
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2015-03-24
(URL)
https://doi.org/10.15002/00011583

Awareness and Control of Personal Data Based on the Cyber-I Privacy Model

Li Tang
Graduate School of Computer and Information Sciences
Hosei University
Tokyo 184-8584, Japan
li.tang.k3@stu.hosei.ac.jp

***Abstract*—Cyber-Individual (Cyber-I), is a counterpart of Real-Individual (Real-I) in cyberspace. It will gradually approximate to its Real-I and be able to provide better personalized digital services. The accuracy of this approximation is successively improved by continuously acquiring and utilizing personal data or information related to a person. However, the collection, processing and access of sensitive personal data may bring a great privacy problem since people may not like their personal data being collected or kept by others. The problem has been receiving increasing attention from both ordinary people and technical researchers who try to offer privacy controls with following users' privacy settings. However, these privacy settings are application-oriented and have to be done manually by users for different applications. To provide a generic and user-centric privacy protection mechanism, this paper proposes a Cyber-I privacy model (CIPM) that is a systematic description about a user's privacy preference, policy and rules, which are generated semi-automatically according to each user's characteristics. Advantages of the user-centric CIPM are twofold: (1) reflecting a user's privacy needs to different applications; (2) adapting to a user's privacy demand changes. Moreover, a platform for CIPM initialization and update is developed, and the privacy protection of personal data is realized through not only control but also awareness based on the CIPM.**

Keywords—Cyber-I privacy model; personal data; awareness; control; preference; policy; rule; adaptation

I. INTRODUCTION

Due to the development of ICT technologies, we have entered into a newly cyber-physical integrated hyper world, which is characterized with digital explosions in data, connectivity, service and intelligence [1]. The appearance of Cyber-Individual, short for Cyber-I, is an effort to digitally clone real individual (Real-I) in the explosive hyper world, and it is able to gradually approximate to its Real-I [2]. The accuracy of the approximation will be continuously improved by collecting and utilizing increasing personal data, which is any information related to a person [3].

In order to build such a comprehensive Cyber-I, the collection of personal data in the initial stage may be acceptable due to relatively less data amount. However, with the increasing collection of more personal data, as well as further data processing and access, people may not feel at ease, and may concern greatly about the high risk of privacy invasions. Potential privacy concerns may cover various kinds of data manipulations including data collection, processing, and access.

Privacy can be described as "the privilege of users to control for themselves when, how and to what extend information about them is communicated to others" [4]; "the ability of an individual to control the terms under which their personal information is acquired and used" [5]; and "control over information disclosure" [6]. The control is a common word or a core function in the above descriptions, but the premise of effective control must be awareness, which is an ability to perceive or be conscious of the personal data status. Both awareness and control are essential in the process of personal privacy protection.

In the existing privacy protection mechanisms, however, privacy controls are only realized in some way of setting up privacy-related items to applications manually by users. Awareness of personal data status and adaptiveness to each user's privacy requirement are not fully accomplished in a practical way. To offer people enough awareness and control so as to effectively protect the privacy of personal data, a Cyber-I privacy model (CIPM) as a core of this research is proposed for a systematic description about a user's privacy preference, policy and rules. Based on the CIPM, users can receive exact expressions in privacy settings, prompt awareness services, and sufficient control of personal data.

The rest of the paper is organized as follows. In Section II, related work about privacy protection is described. In Section III, we first discuss privacy concerns in personal data manipulations, and then propose the Cyber-I privacy model, i.e., CIPM. The CIPM initialization and adaptation are described in Sections IV and V, respectively. Section VI shows the details of awareness and control functionalities. The evaluation of CIPM is described in Section VII, and research summary and future work are given in the last section.

II. RELATED WORK

Many researches on privacy protection have been done especially in ubiquitous environments. The representative architectures or systems are Privacy-respecting Context-aware Architecture Prototype (PCAP) [7], Privacy in Context-aware Computing Environments (PCCE) [8], Privacy-aware Computing (PaC) [9], and Distributed Dynamic Privacy-enhanced User Modeling Framework (DDPUMF) [10].

PCAP provides an interface for users to specify their privacy preferences, which are then uploaded to and stored at privacy agent (PA) residing somewhere on the network. The PA is the core module function of personal data protection. The PCCE based architecture obtains the privacy rules from the GUI defined by users, and then transfers the rules into a context privacy policy language format to implement controls of personal data manipulation. In PaC architecture, a broker handles the communication between all services using an extended agent communication language. A context-aware filter running on the client allows a user to set his/her preferred level of quality of privacy (QoP), which is the tradeoff between the amount of privacy the user is willing to concede and the value of the services that can be provided by an application. DDPUMF relies on an LDAP-based user modeling server, to which users should first define the privacy demands by themselves, and then evaluate them from the privacy conditions expressed in a privacy policy language.

Basically, the existing privacy protection mechanisms have provided users with many controls of personal data privacy. However, privacy settings are all done by users manually and independently for different applications. Considering the real privacy requirements of users and the possible ways of privacy invasions, several problems still exist as follows. (1) Users have to do the same kind of privacy settings, no matter how much they concern about privacy. (2) Privacy settings are application-oriented, and users may be tired of such similar privacy controls. (3) Static privacy settings cannot adapt to a user's dynamic privacy requirements.

III. PERSONAL DATA AND CYBER-I PRIVAY MODEL

Cyber-I is built from the personal data (PD) collected about its Real-I as shown in Fig. 1. The PD can be roughly divided into three categories, *lifelog data* about a person's life records, *Web data* about a person's information existing on the Web, and *model data* about a person's characteristic descriptions. During the Cyber-I's growth process, PD will be continuously collected, and thus become increasing and huge.

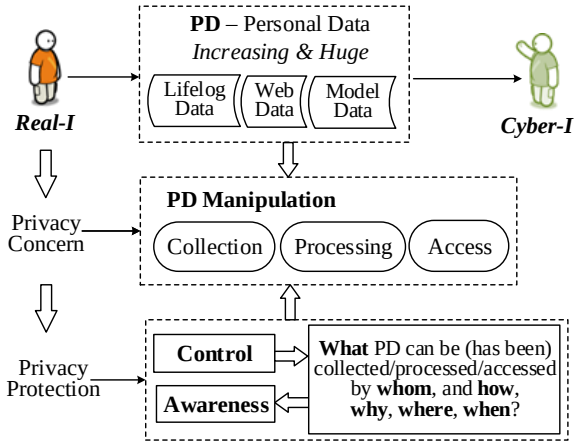

Fig. 1. Personal data and privacy concerns

Manipulations in data collection, processing and access will result in privacy invasions/risks concerned greatly by users. The collection related privacy concern mainly refers to sensitive data collection without users' awareness. The processing related privacy concern is about dealing with personal data in an unreasonable way. The access related privacy concern is possibly resulted from existing sensitive personal data accessed by vicious people. In order to reduce these privacy concerns, users should be provided with sufficient control on their own personal data, and prompt awareness about their data status and manipulation situation.

Considering the above privacy concerns about personal data as well as necessary data control and awareness, we adopt a generic privacy protection architecture in which the CIPM, i.e., Cyber-I privacy model, is core as shown in Fig. 2. Initial CIPM, which is the initial description about a user's privacy preference, policy and rules, is generated based on the user's privacy characteristics and inputs or selections. Adaptive CIPM means that a user's privacy preference, policy and rules can be updated for adaptation to situation changes according to event detections and analyses of personal data access logs. The CIPM is used to implement awareness and control of sensitive personal data. All kinds of data accesses will be recorded, and a user can be aware of and control what personal data can be and has been collected/processed/accessed by whom, and how, why, where, when. The detailed CIPM initialization, adaptation, and utilization will be explained in the next three sections, respectively.

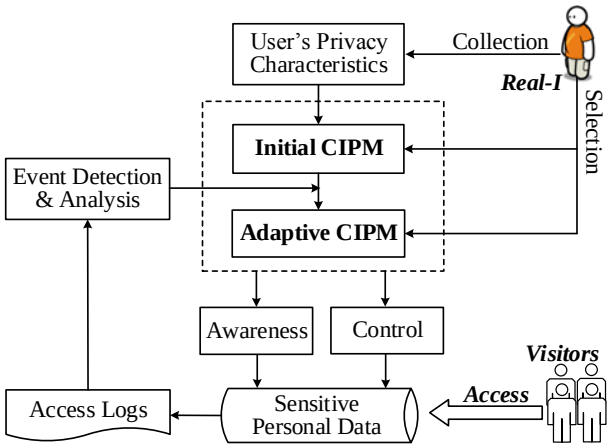

Fig. 2. The overall architecture of privacy protection about personal data

IV. THE INITIALIZATION OF CIPM

Different from conventional privacy protection systems in which users' privacy preferences and control rules are set up manually by the users, one of our core thoughts is to calculate each user's privacy preference degree according to the user's characteristics, map the degree into a suitable privacy policy, and generate the corresponding privacy rules only with the user's simple selections on given options. The whole process of CIPM initialization is shown in Fig. 3, and the details in the calculation, mapping and generation are illustrated in the following subsections.

A. Calculation of Privacy Preference

People with different characteristics may have different privacy concerns. The factors that may influence users' privacy concerns can be divided into intrinsic and extrinsic ones. The intrinsic factor is about a user's basic information including age,

gender, nationality, etc. The extrinsic factor is related to a user's background or experience, and it can be further divided into life-related factor, such as career and life style, and network-related factor, such as network use experience and privacy intruded history.

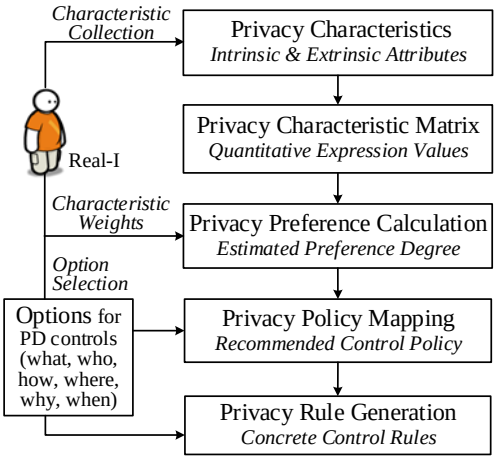

Fig. 3. The initialization process of CIPM

There may be more privacy related factors, and the total number of all possible factors is denoted as *n*, namely, the general factors are with *n* dimensions. Suppose that each of the dimensions/factors includes *m* characteristic attributes, and each attribute will be quantized as a value ranging from *0* to *1* according to its relevance extent to privacy concern. Therefore, all privacy relevant characteristics of a user can be represented as an *n*m* matrix *D* as given in equation (1).

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1m} \\ d_{21} & d_{22} & \dots & d_{2m} \\ . & . & \dots & . \\ d_{n1} & d_{n2} & \dots & d_{nm} \end{bmatrix} \quad (1)$$

The privacy relevance extents to the same factor may vary for different people. The extents can be represented by weights for corresponding factors, which may be pre-defined or specified by each user. For example, suppose that the total weight of intrinsic, network-related and life-related factors is 100%, a user may specify the three weights 30%, 50% and 20% to the three factors. Generally, we use a vector $W_1 = [w_{11}\ w_{12}\ \dots\ w_{1n}]$ to represent the weights of *n* privacy dimensions with meeting the conditions that $w_{11}+w_{12}+\dots+w_{1n}=1$, and $w_{ij} \in [0, 1]$. Similarly, people's privacy relevance extents to an attribute within a privacy dimension/factor may vary as well. For instance, somebodies believe that the age in the intrinsic factor is more important than gender, while others consider the opposite. A vector $W_2 = [w_{21}\ w_{22}\ \dots\ w_{2m}]$ is used to represent the weights of *m* attributes with the conditions that $w_{21}+w_{22}+\dots+w_{2m}=1$, and $w_{pq} \in [0, 1]$. With the characteristic matrix and the two weight vectors, we can calculate a user's privacy preference degree using the equation below.

$$P = (W_2 * D^T) * W_1^T \quad (2)$$

Figure 4 shows an example in which there are three dimensions/factors and three attributes. Each characteristic attribute is converted into a score/value in [*0, 1*], and the user's characteristics are a *3*3* matrix as shown in the middle of equation (3). Once the two weight vectors are specified, the user's privacy preference is calculated using equation (3), and the result is *0.374*.

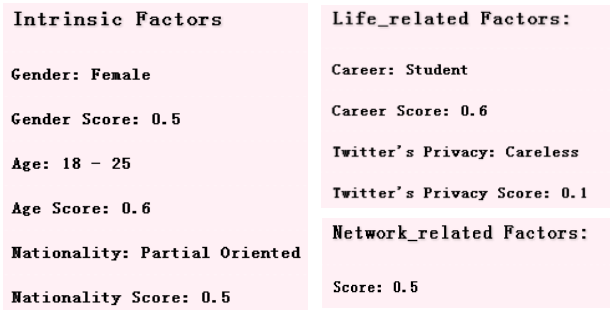

Fig. 4. Values of privacy related characteristics

$$P = \left(\begin{bmatrix}0.5 & 0.3 & 0.2\end{bmatrix} * \begin{bmatrix}0.5 & 0.6 & 0.5 \\ 0.6 & 0.1 & 0 \\ 0.5 & 0 & 0\end{bmatrix}\right) * \begin{bmatrix}0.3 \\ 0.5 \\ 0.2\end{bmatrix} = 0.374 \quad (3)$$

B. Mapping from Privacy Preference to Privacy Policy

By referring to the privacy model in [11], six variables, *what*, *who*, *how*, *where*, *why* and *when*, are involved in privacy policy in our model. There is no doubt that the number of variables in privacy policies can be extended when necessary.

The most basic privacy policy is ***w2*** policy, for the reason that our purpose is to protect personal data, which is represented by *what*, and the attackers may be other users or some applications accessing personal data maliciously, which are represented by *who*. If the personal data must be disclosed to whom, the next concern is *how* to protect the disclosed personal data. So the second privacy policy is ***w2h***, which is a combination of *what*, *who* and *how*. After taking measures to protect personal data, we may further restrict the location of a requestor with the third privacy policy, a combination of w*hat*, *who*, *how* and *where*, named ***w3h*** policy. Restricting the location of the requestor, we should further figure out the purpose of the requestor accessing the personal data, that is the meaning of *why*, which composes the fourth privacy policy, a combination of *what*, *who*, *how*, *where* and *why*, named ***w4h*** policy. Finally the strictest privacy policy contains all the six variables, named ***w5h*** policy.

When mapping a privacy preference to a privacy policy, three mapping functions or curves can be used as shown in Fig. 5. For a given value of privacy preference, such as *0.374*, a vertical straight line intersects with the three curves at points A, B and C, which are mapped to three options of possible policies. Curve (1) is more fitted for open-minded people, as they want to control privacy in a simple way. In this example, curve (1) generates the *w2* privacy policy, and the user only needs to specify two variables. While curve (3) is designed for conservative people, as they may want to control privacy in a more detailed way. In this example, curve (3) generates *w4h* privacy policy, and users need to specify five variables

respectively. It is obvious that *w4h* is more complicated than *w2*. People without distinctive features would be suitable for curve (2). In this example, with curve (2), the privacy value *0.374* is mapped to the *w2h* privacy policy. The complexity of *w2h* is between *w2* and *w4h*. This design is similar as the μ-law algorithm widely used in digital telecommunication systems.

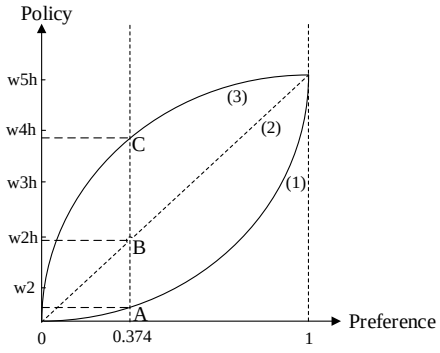

Fig. 5. Mapping from a privacy preference to a privacy policy

C. Generation of Privacy Rules

If we regard the privacy policy as the overall principle of privacy control, the privacy rules can be regarded as concrete control implementations. Based on a user's privacy policy mapped from the privacy preference, the user can select concrete items in given options associated with each of variables in the policy. For example, if a user gets *w2h* policy, he/she should choose which options can be accepted in *what*, *who* and *how* variables. Privacy rules are generated automatically based on users' selections. Contents of *what* variable are privacy subjects to be controlled, while the remaining variables within privacy policy are regarded as conditions. An example of initial CIPM generation can be seen in Fig. 6. Inside the system, the rules generated are expressed in XML format, such as the one below.

```
<rule name="w2h Rule 1">
<condition>If family-member</condition>
<condition>If cipher</condition>
<result>Disclose contact-information</result>
</rule>
```

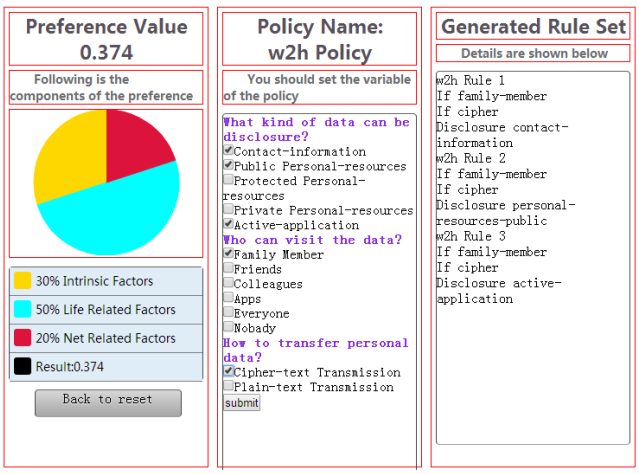

Fig. 6. An example of initial CIPM generation

V. The Adaptation of CIPM

As described in the last section, the initial CIPM is generated based on a user's characteristics and their processing, but the initial one may not match the user's demands perfectly due to various reasons. For instance, the user's characteristic information may be incomplete, or the weights specified may not be appropriate, or the user's selection of optional items could be inconsistent. Furthermore, the context or situation after generating the initial model may be changed. For example, new types of personal data will be added, or the user's privacy concern extents to privacy may be altered, or a privacy intrusion event may happen. Because of these, the CIPM must be updated for adaptation to possible changes related to the user and situation. It is expected that the adaptive CIPM can approach to a user's exact privacy features and thus provide better privacy protections along with the use and growth of the user's Cyber-I.

The adaptations in our CIPM are conducted in three forms based respectively on (1) user's feedbacks, (2) personal data access logs, and (3) contradictions between a user's settings and his/her actual behaviors. The forms and associative processes of our adaptive CIPM are depicted in Fig. 7.

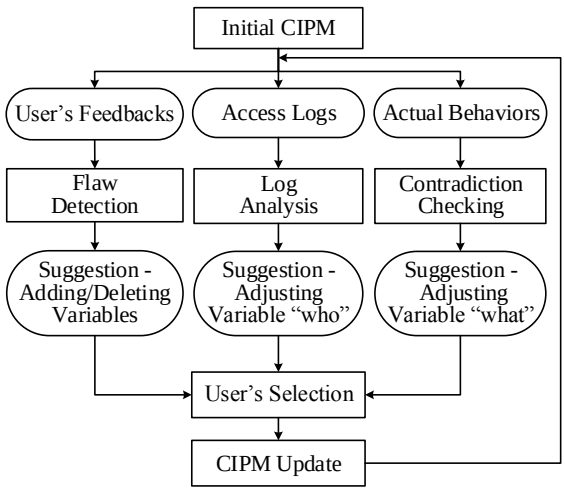

Fig. 7. The forms and process of adaptive CIPM generation

User's feedbacks about the performance in using the present CIPM-based privacy awareness and control are direct evaluations that can be used to update CIPM. One way to get the feedback is to ask a user to fill an online questionnaire in a certain period. The questionnaire includes the user's satisfactory degrees about the privacy relevance extents to his/her characteristics as well as corresponding weight values. Such feedbacks can be used to calibrate the user's privacy preference. The questionnaire also includes a user's evaluations about awareness and control effects, and even the adaptive CIPM itself. In our implementation, a user will be first asked to give an overall evaluation, and if it is negative, the user will be asked to check detailed items in a questionnaire. According to the detailed feedbacks, the user will be suggested to add or delete some variables in the current policy and/or control rules. The key in such a form of adaptation is how to effectively collect a user's feedbacks and process them for making accurate suggestions to the user.

Access logs are records about all accesses to a user's personal data. From the logs, we can know the access status and detect abnormal accesses. The basic information contained in the log is about what personal data pieces are accessed by whom from what possible locations at what time. From the information, data access features can be known, such as who accesses what data often. When the user has pre-specified the name/account information of his/her friends, it is possible to identify whether an access is from a familiar member or not. If a large number of accesses from an identical stranger are detected, this can be regarded as an indicator that the visitor may have some special purpose to the user, and accordingly the user should be suggested to pay attention to the visitor and make a necessary adjustment on the variable "*who*" related options. Analyzing the logs and detecting abnormal accesses are very helpful for a user to promptly know the possible risks and take proper actions so as to greatly prevent and reduce privacy invasions.

Contradictions in privacy rules mean the inconsistence between a user's privacy settings and his/her actual behaviors in manipulations of his/her own personal data. The occurrence of such contradictions may be because a user has no enough experience in privacy management, or personal data types and amounts become more and too complex to correctly manipulate in all time. Let us look at typical examples below. Suppose that a user initially sets to disclose contact information, but his/her contact information is empty in the profile kept in Cyber-I personal database. Assume that a user sets to disclose uploaded data or some personal resources, actually all uploaded personal data and resources are marked with a *private* tag, which means inaccessible by others. Given that a user chooses to disclose data generated by applications, however, he/she rarely uses these applications. Once a contradiction is detected, the system will suggest the user to check and adjust the variable "*what*".

Figure 8 shows a CIPM adaptation example in which three kinds of information about the user's feedbacks, logs' analysis results, and detected contradictions are given in the large background window. The GUI window also provides three sets of suggestions to the user. After the suggestions are committed by the user, the previous and the updated rules in CIPM are shown in the two small snapshots in the figure.

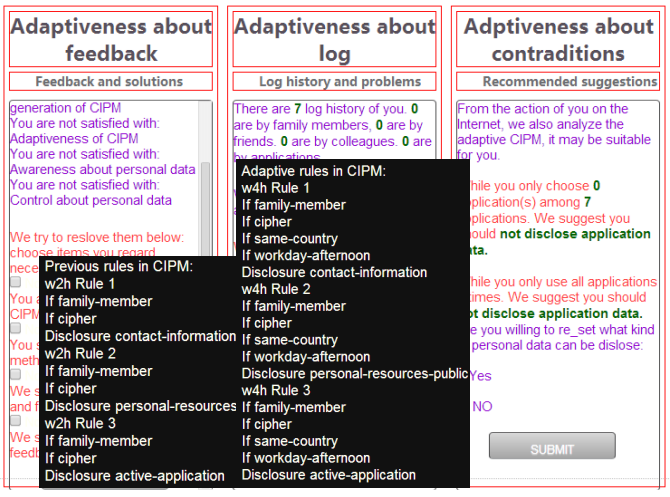

Fig. 8. Suggestions about adaptation, previous and adaptive privacy rules

VI. Awareness and Control of Personal Data

During the process of generating and using Cyber-I in our system, one of our fundamental requirements is to protect users' privacy by providing sufficient awareness and control functions on what data can be or has been accessed by whom, and how, where, why, when based on the CIPM.

Figure 9 shows three main functions of CIPM-based awareness for being aware of personal data current status, access histories and potential risks. The current data status enables a user to check what kind of personal data has been collected, and decide to delete unwanted items of own personal data. Access history mainly shows what personal data has been accessed by whom at what time and in how many times. Meanwhile, a user will receive an email to be informed about a possible privacy risk in his/her personal data based on analyzing personal data access histories. All kinds of awareness are functioned based on the user's CIPM, for example, when access histories will be informed, how potential risks will be announced, and on what kind of conditions, current data status will be forwarded to users.

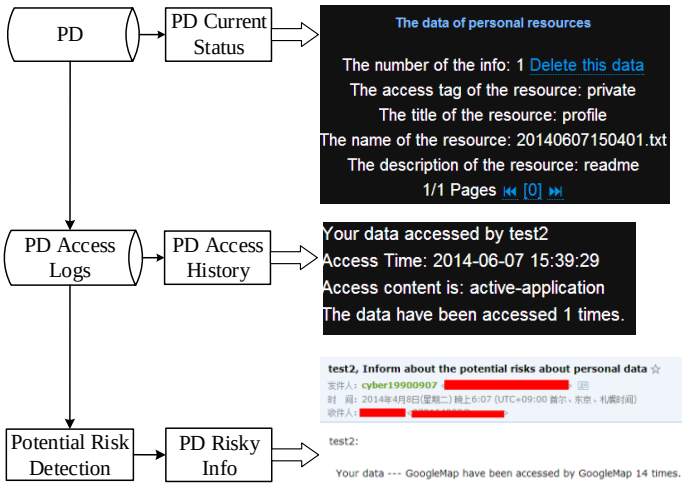

Fig. 9. Awareness of personal data, access histories and potential risks

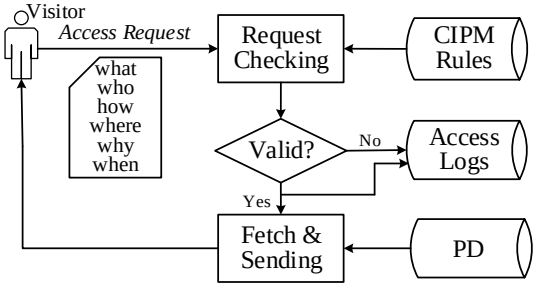

Fig. 10. The control and record of personal data access

Figure 10 shows the overall process of personal data access control to manage what kind of personal data can be disclosed, who can access, and how to handle the data access requests from other visitors. If a visitor requests to access a certain kind of personal data (PD), it will be checked whether the parameters in the request, such as *what*, *who*, *how*, etc., satisfy with the rules in corresponding CIPM. If all parameters match

with the conditions in privacy rules, the request will be regarded as valid, and the requested personal data will be taken from personal data database and sent to the requestor. Otherwise, the request will be denied. Any request to personal data access will be recorded in an access log database.

VII. EVALUATION OF CIPM-BASED PRIVACY PROTECTION

The performance of the CIPM-based privacy protection is depended up the strength to prevent *Internal Violations (IV)* and *External Violations (EV)* [12]. *IV* means that service providers may be dishonest and violate the privacy policy of their own. *EV* refers to services offered by and private data kept in service providers are attacked by malicious parties. The ability of CIPM to prevent *IV* can be reflected from users' evaluation scores about the reliability of CIPM functionalities. In our *IV* evaluation, each test user was asked to give a score M_i ranging from *0* to *10*. As for *EV* evaluation, the ability to protect a user's personal data is denoted as M_e and calculated in equation (4), in which *n* is the whole number of personal data (PD) records, and *p* is the number of these records without privacy disclosure, i.e., successfully protected by our CIPM system. The purpose of multiplying *10* is to make the value M_e range from *0* to *10*.

$$M_e = 10 * \frac{p}{n} \tag{4}$$

All evaluation results from *10* test users are shown in Tab. I. The M_e to each test user was calculated using the above equation according to the user's PD number and the protected PD number. The M_i was directly scored by each test user. For a value in both M_i and M_e, the larger the value is, the better the user's evaluation is. To visualize integrated evaluations, we draw all *10* M_i-M_e value pairs on an *IV-EV* evaluation plane as shown in Fig. 11. In the *IV-EV* plane, the best performance region is on the top-right close to *(10, 10)*, and the worst region is on the bottom-left close to *(0, 0)*. The average evaluation values of M_i-M_e (*7.3*, *8.87)* are also given in the table and shown in the figure. It can be seen that the performance of our system is good in protecting *EV* since the most M_e values are between *8* and *10*. The M_i values are very diverse among the test users and with the average value *7.3*, which illustrates that the performance in protecting *IV* is not perfect but still good. According to all these evaluation results, the whole performance of CIPM-based privacy protection is relatively good, near the best region, which is to be achieved in the future.

TABLE I. EVALUATION DATA AND RESULTS

User	PD number (n)	Protected PD number (p)	M_i (Score)	M_e
TestUser1	30	24	10	8
TestUser2	55	30	9	5.45
TestUser3	9	9	9	10
TestUser4	74	70	7	9.46
TestUser5	111	110	3	9.9
TestUser6	37	36	8	9.73
TestUser7	69	63	6	9.13
TestUser8	16	13	6	8.125
TestUser9	45	40	5	8.88
TestUser10	7	7	10	10
Average	45.3	40.2	7.3	8.87

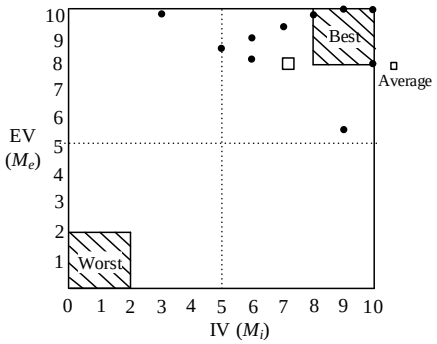

Fig. 11. Evaluation results of CIPM-based privacy protection

VIII. CONCLUSION AND FUTURE WORK

This paper proposes a user-centric Cyber-I privacy model, shortly CIPM, including its initialization and adaptation for updating the CIPM so as to reflect a user's privacy features and demands more exactly and precisely. Based on the CIPM, various functions for personal data awareness and control have been developed to protect users' privacy. The preliminary evaluations have shown the effectiveness of CIPM-based privacy protection of personal data.

This research is still at its infancy, and much work remains to improve CIPM. The future work will be carried out to refine the generation mechanisms of both initial and adaptive CIPM, provide more awareness and control service, and conduct further evaluations on the CIPM.

REFERENCE

[1] J. Ma, "Digital explosions and digital clones in cyber world," the IEEE Int'l Conf. on Intelligent Computing, Keynote Speech, 2011.

[2] J. Ma, J. Wen, R. Huang and B. Huang, "Cyber-Individual meets brain informatics," IEEE Intelligent Systems, vol. 26, pp. 30-37, 2011.

[3] J. Wen, B. Huang, and J. Ma, "Cyber-I: Vision of the individual's counterpart on cyberspace," IEEE International Conference on Dependable. Autonomic and Secure Computing, 2009, pp. 295-302.

[4] J.I. Hong and J.A. Landy, "An architecture for privacy-sensitive ubiquitous computing," Proceedings of the 2nd International Conference on Mobile Systems Applications and Services, June 2004, pp. 177-189.

[5] R.S. Cardoso and V. Issarny, "Architecting pervasive computing systems for privacy," Washington, Proceedings of the 6th Working IEEE/IFIP Conference on Software Architecture, 2007, pp. 26.

[6] J. Mary, "Protecting privacy online: Is self-regulating working?" Journal of Public Policy and Marking, 2000, pp. 20-26.

[7] H. Chen, T. Finin, and A. Joshi, "An ontology for context-aware pervasive computing environments," The Knowledge Engineering Review, 2003, pp. 197-207.

[8] A. Behrooz, "Privacy of mobile users in context-aware computing environments," The Ph.D. Dissertation, KTH Information and Communication Technology, 2011.

[9] M. Tentori, J. Favela, V. Gonzalez, and M. Rodríguez, "Towards the design of privacy-aware computing," Workshop on UbiComp Privacy. Privacy in context. Ubicomp, Tokyo Japan, September 2005, pp. 11-14.

[10] Y. Wang, "PLA-based privacy-enhancing user modeling framework and its evaluation," Journal of User Modeling and User-Adapted Interaction, vol. 23, pp. 41-82, March 1st, 2013.

[11] F. Cena, N. Dokoohaki, and M. Matskin, "Forging trust and privacy with user modeling frameworks," The 1st International Conference on Social Eco-Informatics, 2011, pp. 43-48.

[12] G. Yee, "Measuring privacy protection in Web services," Interntional Conference on Web Services, Chicago, 2006, pp. 647-654.