User-Centric Privacy Preservation in Data-Sharing Applications

Feng Gao[1], Jingsha He[2], and Shufen Peng[1]
[1] College of Computer Science and Technology, Beijing University of Technology
Beijing China
[2] School of Software Engineering, Beijing University of Technology
Beijing China
maple0371@emails.bjut.edu.cn
jhe@bjut.edu.cn

Abstract. During data sharing process across people, users do not want the information that contains privacy to be shared with everyone else; some people may only want to share it with someone they are familiar with. To solve this issue of privacy preservation in data-sharing applications, we propose a novel user-centric method. Our main contributions include the followings. (1) Users can select key words or characters for their own privacy information, which is a user-centric way to protect privacy. (2) During the process of data sharing, data substitution can be used to ensure privacy preservation as well as high successful rate of data sharing. (3) Our method can be used in different data-sharing applications in a flexible way. Simulation results show that our method can achieve our privacy preservation goal.

Keywords: user-centric, privacy preservation, data sharing, trust, P2P network, online social network

1 Introduction

The evolution of peer-to-peer (P2P) networks has triggered large scale distributed applications. The main application domain is data sharing across a very large number of highly autonomous participants. The popular scenarios of data sharing in the peer-to-peer (P2P) networks focus on massive file sharing. Advanced scenarios such as online communities (e.g., medical communities) need to share private or sensitive

data frequently. The other typical data-sharing application is online social networks, which is immensely popular, claiming over 200 million users. In online social networks, the information that users share may include privacy, such as personal information or private photographs that should not be misused [2].

Privacy is a concept combining law, sociology and psychology. The dimension of privacy refers to user's culture, education level, preference and so on. Therefore, the definition, the sensitive degree and the range of privacy information is diversiform. According to the multiplicity of privacy understandings, it should be important to research the privacy preservation mechanism that is user-centric way.

To solve the issue of privacy preservation in data-sharing applications, we propose a novel user-centric method in this paper that includes following contributions. Firstly, user can select key words or characters by themselves for their own privacy information which means a user-centric way to protect privacy. Secondly, if the data under sharing contains too much privacy according to user's privacy policy and trust model, data substitution can be used to ensure privacy preservation. The substitution also brings higher successful rate of data sharing. Thirdly, our method can be applied in different data-sharing applications in a flexible way.

The rest of this paper is organized as follows. In section 2, we introduce some related work. In section 3, we explain our user-centric privacy preservation method that includes formally description and the execution model. In section 4, we perform some simulations. The results show that our method can achieve our privacy preservation goal. Finally, we conclude this paper in Section 5.

2 Relate Work

Privacy preservation is an important issue in data-sharing applications. Jawad et al. [1] propose a P2P data privacy model which combines the Hippocratic principles and the trust notions to support P2P systems. Di Crescenzo and Lipton [2] describe an evolving access control mechanism in social networking to provide non-trivial quantifiable guarantees for formally specified requirements of utility and privacy. Wei Q and Yansheng L [3] introduce a practical solution to defend against background knowledge attacks with considering the privacy disclosure in social network data publishing. Kobsa A. and Teltzrow M. [4] proposed a user interface design approach in which the privacy practice of a website was explicated in a contextualized manner.

These methods for privacy preservation in data-sharing applications focus on certain computing application, none of them can offer privacy preservation in all scenarios for data sharing.

Selective partial encryption offers protection for compressed image data [5], audio data and video data. Although this method can protect privacy by encrypting partial data which the user does not want to be shared, the receiver can recognize that the sender does not trust him from the partial encryption data. Therefore, this method is unsuitable for data sharing with privacy preservation.

3 User-Centric Privacy Preservation Method

3.1 Formally Description

In this section, we describe the concepts and expressions used in our privacy preservation method.

Data object We consider the user's data as objects. These objects make up to a logical database. Let $D_1, D_2, \ldots D_m \in D$ be the user's data objects (they can be image, text, file and so on), $\{D_1, D_2, \ldots D_m\}$ is the user's logical database, and D is the space of data object.

Privacy range The range of privacy information is defined by the customized setting and the default setting. The customized setting means the user can choose the key words and the characters of their privacy information. The set $K \{k / k \in (k_1, k_2, \ldots k_n)\}$ and the set $C \{c / c \in (c_1, c_2, \ldots c_o)\}$ denote the choice of the user's key words and characters for their privacy information respectively. The default setting means that makes all the general acknowledged privacy information into privacy range.

Privacy policy ontology We abstract the user's privacy policy as privacy policy ontology and its attribute as trust evidence with constraint which need to be satisfied in semantic way ([6]).

Privacy information ontology According to the customized setting, the default setting and privacy policy ontology, privacy information ontology created in semantic way ([6]) correspondingly. Privacy information ontology describes the needed trust evidences when disclosing privacy.

Privacy mapping function P The mapping function $P: D \to \{0,1\}$ denotes the user data D whether contains privacy based on privacy information ontology. Let $d_1, d_2, \ldots . d_r \in \{0,1\}$ express the value of function P where $d_j = P(D_j)(j = 1,2,\ldots r)$. And $d_j = 1$ means the user's data D_j involves privacy while $d_j = 0$ means D_j does not involve privacy.

Privacy entropy We describe the entity's trust that should be achieved when disclosing a piece of privacy as T_a , the trust that the entity has already been achieved before the privacy disclosure as T_b .Conditional probability $P = prob(T_a / T_b)$ denotes the probability of achieving T_a under condition T_b . Set $T_e\{e_1, e_2, \ldots e_s\}$ denotes the needed trust evidences when disclosing one piece of privacy information. Conditional probability $P_i = prob(T_{ai} / T_{bi})$ represents the i^{th} trust evidence's conditional probability when disclosing the privacy information. Let $p_1, p_2, \ldots p_s$ denote the original values of conditional probability p_i , preprocess P_i by formula 1.

$$P_i = \frac{p_i}{p_1 + p_2 + \ldots . p_s} \tag{1}$$

We use H to denote privacy entropy which means privacy information quantity:

$$H = -k(\sum_{i=1}^{s} P_i \log P_i) \qquad (\sum_{i=1}^{s} P_i = 1 , k = p_1 + p_2 + \ldots . + p_s) \tag{2}$$

Trust mapping function T The mapping function $T: D \to (a,b)$ denotes that for certain trust, the interval of privacy entropy can be afforded in data sharing. This mapping function related to the user's trust model and privacy policy.

Substitution data object Let $D_1', D_2', \ldots . D_t' \in D$ be the substitution data objects (they can be image, text, file and so on), $\{D_1', D_2', \ldots . D_t'\}$ as logical substitution database, and D' is the substitute data object space.

3.2 Privacy Preservation Method

The architecture of our user-centric privacy preservation method in data sharing is illustrated in Fig.1. And this method includes privacy setting module and privacy processing module.

Privacy setting module, in this module the users can choose their own key words ($K \{k / k \in (k_1, k_2, \ldots . k_n)\}$) and characters ($C \{c / c \in (c_1, c_2, \ldots . c_o)\}$) for their privacy with the default setting which makes the all general acknowledged privacy information into privacy range.

Privacy processing module, this module process the user's sharing data by four steps.

Step1 Scan the information that the user want to share by using privacy mapping function P, for the sharing data set D_s $(D_1, D_2, \ldots D_t)(t \le m)$, if $d_l = P(D_l) = 1(l = 1,2,\ldots m)$, record d_l to set d' $d' \in \{d_k' / d_k' = d_j, d_j = P(D_j) = 1\}, (k \in (1, w), j \in (1, k))$.

Step2 For each D_j that $P(D_j) = d_j, (d_j \in d')$, compute $H(D_j)$, then let H' records the sum of the $H(D_j)$ based on formula (2).

$$H' = \sum_{j=1}^{w} H(D_j) \left(H(D_j) = -k(\sum_{i=1}^{n} P_i \log P_i) \right) \quad (3)$$

Step3 Divide the trust interval to define that for the certain trust how much privacy information can be shared. The total of interval can be set by the user or default as a certain constant. Denote there is u trust interval T, each interval controls v terms privacy information Q, for $t_i \in T, q_j \in Q$, t_i controls q_j privacy information. For trust interval t_i, the lower bound of mapping function $T : D \to (a, b)$ says a_i compute by formula (4).

$$a_i = \frac{1}{u}(i-1) \sum_{i=1}^{u} \sum_{j=1}^{v} H(q_j)(i = 1,2,\ldots u) \quad (4)$$

And the upper bound says b_i compute by formula (5).

$$b_i = \frac{1}{u} \cdot i \cdot \sum_{i=1}^{u} \sum_{j=1}^{v} H(q_j)(i = 1,2,\ldots u) \quad (5)$$

So the interval (a_i, b_i) denotes the privacy entropy that the i^{th} trust interval can be afforded in data sharing.

Step4 For the sharing data set, if $H' \in (a_i, b_i)$, share the data with the entity. If $H' \notin (a_i, b_i)$, substitute some sharing data from substitution database, until the H' satisfied $H' \in (a_i, b_i)$, then share that new data set.

Our user-centric privacy preservation method in data-sharing applications includes 9 steps as follows.

1. The user input the data that he wants to share to privacy preservation model; 2. Creates privacy policy ontology according to privacy policy and based on it also with the user setting and the default setting, privacy setting module creates the privacy ontology; 3. Scan input data set by using privacy mapping function P then record the information which involved privacy; 4. Compute the entropy of privacy information;

5. Divide the trust interval; 6. According to entity's trust, judge the privacy entropy whether in the interval or not. If privacy entropy is in it, share the data set, otherwise do step 7; 7.Substitute some data from substitution data base; 8. Do 4 until the privacy entropy is in relevant interval; 9. Output the new data set which means privacy preserved.

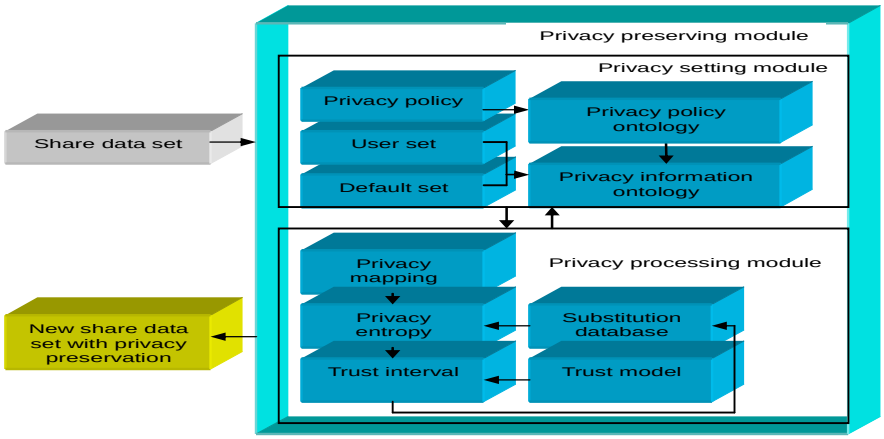

Fig. 1. Architecture of Privacy Preservation Model

4 Simulations and Analysis

We have performed evaluation of our method using the .NET Framework 2.0 platform and implemented a simulator written in C#. The experiment was carried out on a laptop computer with an Intel Pentium (R) Core(TM) 2 Duo 1.83 GHz CPU and 1G bytes of memory.

In our simulation, we choose the parameters as follows.

The data set $\{D_1, D_2, \ldots D_m\}(m = 500)$ is the user's logical database, there is $u(u = 5)$ trust interval T, each interval controls $v(v = 10)$ terms privacy information Q. The space of privacy information is 50 terms and the data set $\{D_1', D_2', \ldots D_t'\}(t = 15)$ is logical substitution database. Privacy entropy uses the data form our previous work [7], the sum of privacy information entropy is between 0 and 3.5 in our randomly selecting data simulations.

In case 1, we select 10 terms data from 500 randomly by using our privacy entropy computing method but without substitution method to describe the situation of data sharing. We run our simulation 50 rounds, the x-axis records the time of round, and

y-axis denotes the trust interval. There are 5 trust intervals which can afford privacy entropy $[0,0.7)$ $[0.7,1.4)$ $[1.4,2.1)$ $[2.1,2.8)$ $[2.8,3.5]$ respectively. In our randomly selection, if privacy entropy of sharing data is in arbitrary interval, record as one successful data sharing. The result of case 1 is illustrated in Fig.2. The coordinate (x, y) means the y^{th} simulation involved x privacy. Take point $(40,0.7)$ as example, it means the 40^{th} simulation data sharing scenario is successful if entity's trust can afford privacy entropy 0.7. That is to say, if the entity's trust is in the certain trust interval which can afford privacy entropy between 0.7 and 1.4, this entity can share data with user. And the entity whose trust is in the certain trust interval which can afford privacy entropy $[1.4,2.1)$, $[2.1,2.8)$ and $[2.8,3.5]$ respectively also can share data with user. So our simulations show the situation of once data selecting, each trust interval that entity's trust in whether can share this selected data or not.

Then we do case 2 that select 20 terms data from 500 randomly and show result in Fig.3. We can see that the points which mean successful data sharing in Fig.3 is universal lower than the points in Fig.2. That is to say in case 1 the demand of trust to share selected data is more strictly.

In case 3 and case 4, we select 10 data terms and 20 data terms respectively from the user data space randomly with using our substitution method to describe the situation of data sharing. Fig.4 show the result of case 3 and Fig.5 illustrates the result of case 4 respectively.

We denote "successful rate of data sharing" as successful data sharing to all of data sharing scenarios. We then record the successful rate of data sharing and illustrate in table 1. We can see that without the substitution method, successful rate of data sharing is 37.2% and 34.8% respectively according to our simulation scenario. That is clear that in process of data sharing with privacy considering the success rate is low. With substitution method, the successful rate of user data sharing is close to 54.4% and 54.8% respectively. That accounts for our method is effective in data sharing with privacy preservation. Also we can learn that the more data selection for share the smaller successful rate by comparing case 1 to case 2.

Fig.6 illustrates information quantity of data sharing compare case 1 to case 3. By using substitute method, the information quantity is less than the data sharing without substitute method which means privacy enhanced.

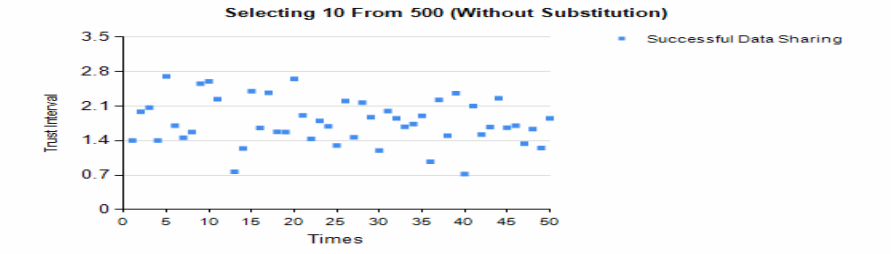

Fig. 2. The situation of data sharing selecting 10 from 500 (without substitution)

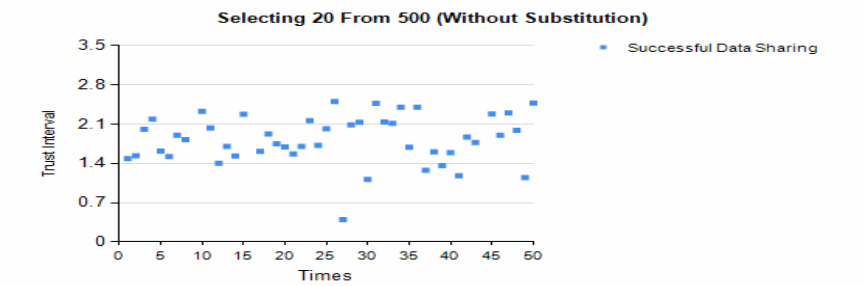

Fig. 3. The situation of data sharing selecting 20 from 500 (without substitution)

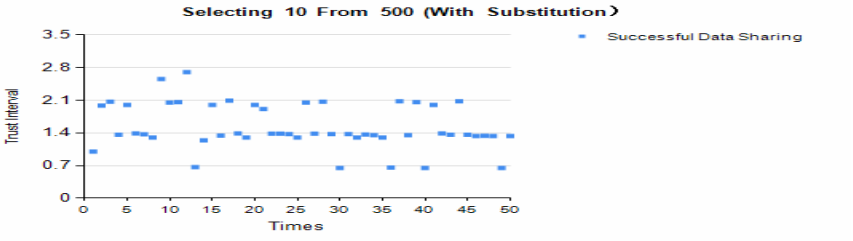

Fig. 4. The situation of data sharing selecting 10 from 500 (with substitution)

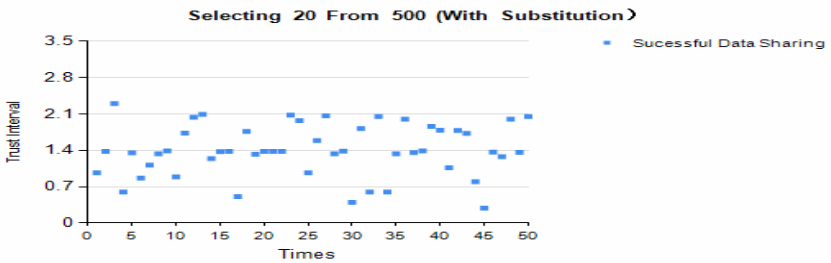

Fig. 5. The situation of data sharing selecting 20 from 500 (with substitution)

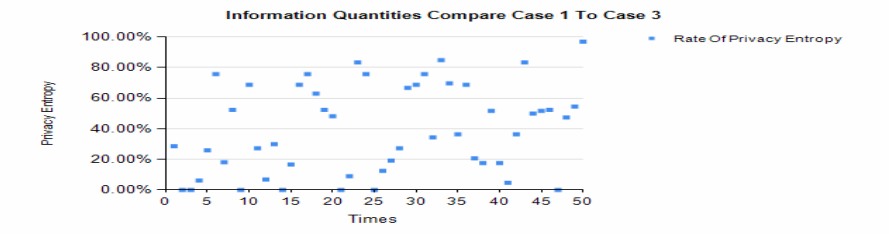

Fig. 6. Information quantities compare case 1 to case 3

Table 1. Successful Rate of Data Sharing

Simulation Case	Successful Rate of Data Sharing
Case 1	37.2%
Case 2	34.8%
Case 3	54.4%
Case 4	54.8%

5 Conclusions and Future Work

Privacy preservation is an important issue for data sharing in applications such as P2Pnetworking, online social networking and so on. Privacy is a concept combining all kinds of elements, and the dimension of privacy is diversiform.

To solve the privacy preservation problem in data sharing, we propose a novel user-centric method in this paper. In our method, the users can select key words and characters for their privacy information, which means user-centric way to protect privacy. Then during the process of data sharing, if the sharing data contains too much privacy according to user's privacy policy and trust model, our data substitution method can be used to ensure privacy preservation and as well as high successful rate for data sharing. And our method can be used in different data sharing applications in a flexible way. We did five simulation cases and simulation results show that our method can achieve our privacy preservation goal.

In the future, we will analyze the complexity and cost of our approach; refine our method to balance privacy preservation and its additional cost. We need further improve our method with real implementation and applications.

Acknowledge

The work in this paper has been supported by research funding from Beijing Education Commission (Grant No. KM201010005027).

References

1. Jawad, M., Serrano-Alvarado, P., and Valduriez, P.: Protecting Data Privacy in Structured P2P Networks, Data Management in Grid and Peer-to-Peer Systems, Linz, Austria. (2009).
2. Di Crescenzo, G., and Lipton, Richard J.: Social Network Privacy via Evolving Access Control, Lecture Notes in Computer Science (including sub series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 5682 LNCS, p 551-560, (2009).
3. Wei, Q., and Yansheng, L.: Preservation of Privacy in Publishing Social Network Data, International Symposium on Electronic Commerce and Security (2008).
4. Kobsa, A., and Teltzrow, M.: Contextualized Communication of Privacy Practices and Personalization Benefits: Impacts on Users' Data Sharing and Purchase Behavior, Lecture Notes in Computer Science, v 3424, p 329-343, Springer Verlag, (2005).
5. Spinsante, S. and Gambi, E.: Selective Encryption for Efficient and Secure Transmission of Compressed Space Images, International Workshop on Satellite and Space Communications (IWSSC), 9-11 , Tuscany, Italy, Sept. (2009).
6. Gao, F., He, J., Peng, S., Wu, X. and Liu, L.: An Approach for Privacy Protection Based-on Ontology, The 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing, Wuhan, China, April 24-25, 2010.
7. Gao, F., He, J., Peng, S., and Wu, X.: A Quantify Metric for Privacy Protection Based on Information Theory, 3rd International Symposium on Intelligent Information Technology and Security Informatics (IITSI 2010), Jinggangshan, China April 2-4, 2010.