Privacy Protection for Recommendation System: A Survey

Weiming Huang, Baisong Liu* and Hao Tang
Ningbo University, Ningbo 315211, China
*Corresponding author’s e-mail: liubaisong@nbu.edu.cn

Abstract. Recommendation system have become one of the most well-liked and accepted way to solve overload of information or merchandise. By collecting user's personal data for processing, suitable lists of information or merchandise are provided to the potential consumers. For online business, recommendation systems have become an extremely effective revenue driver and developed rapidly. Although recommendation systems are great beneficial, directly exposing privacy data to the recommender may lead to leakage of privacy and cause risks. Therefore, quality of recommendation and privacy protection are both important metrics in recommendation. In this paper we present a review investigating development in recommendation systems with privacy protection, including the definition of privacy, classification of privacy leakage, taxonomy of privacy, measuring of privacy risk, policies for privacy protection, approaches of privacy protection and models of privacy protection. We also speculate on the future direction.

1. Introduction
Recommender systems have attracted lots of attention since they alleviate the information or merchandise overload problem for users. As it is known, recommender systems aim at recommending information items or social elements that are likely to be of interest to users. By collecting characteristics of the individual, explicitly or implicitly, such as gender, occupation, location, age, click rate, rating of merchandise, and so on. Automated recommendation systems are essential for users to discover information and merchandise they love and for supplier to reach appropriate audience. Meanwhile, improper collection, storage and transmission of data of users, increase the probability of leakage of users’ sensitive data. When sensitive data exposed to the malicious even in a short period of time, bad things were primarily caused and the victims suffered irreversible injury. Therefore, in recent years, a number of researchers had sought to make a better system which could maintain the accuracy of recommendation while preserving user privacy. In this case, there is a clear need for the designer to understand what kind of data should be taken into account as sensitive data, in which way privacy leakage emerge, and how to evaluate the risk.

The rest of this paper is organized as follows. in section 2, basics of structure and techniques of recommendation systems are firstly presented. Then, we propose definitions of privacy in recommendation system and present a classification of privacy. Finally, we list evaluation method for privacy risks, as we try to figure out multiple ways by which such leakage occurs, accordingly. In Section 3, we describe the remarkable work for privacy protection in this field. And this section is divided into two parts: policies, approaches; we describe these parts explicitly. in section 4, we will discuss the future research should be noticed, and in section 5, we conclude the paper.
2. Recommender systems and privacy protection

2.1. Structure and technique for recommender system

Recommendation systems have drawn an increasingly broad range of interest since early 1990s. Nowadays, recommendation systems are widely deployed for online services. Areas where recommendation systems are widely used include e-commerce websites, movies websites, vide websites, music websites, social networking operators, reading websites, location-based services operators, personalized email, and advertising. By collecting characteristics of the individual, explicitly or implicitly, such as gender, occupation, location, age, click rate, rating of merchandise, and so on, automated recommendation systems are essential for users to discover information and merchandise they love and for supplier to reach appropriate audience. Although different websites may use different recommender technologies, in general, almost all recommender applications are composed of three parts: the display page in the foreground, the log system in the background and the recommendation system algorithm. Therefore, this paper tries to focus on these three parts when introducing different personalized recommendation systems. Based on the algorithm, recommender systems can be divided into four categories. The first approach adopted is based on collaborative filtering, which works by evaluating user preference through exploiting user feedback data so as to compare the similarity of new users or users' preference, and give recommendation lists of similar users in history profile. The second one is based on content, by calculating the characteristics of the content or goods that users have browsed, collected or purchased, generate lists of content or goods with similar characteristics. The third method is based on humanistic and social information. By analyzing the information provided by users when they log in or register, a preliminary recommendation list is generated, which is usually used in the case of 'cold start' with few user information. And the last one provides recommendations by learning external knowledge. Recently, motivated by the recent advances in deep learning, we lay out a vision of how deep learning techniques can be used in recommendation system. Although still experimental, deep learning has been found to be particularly useful in recommendation systems. The following Table 1 lists the four types of recommendation systems.

| Names                          | algorithms                                      |
|--------------------------------|-------------------------------------------------|
| Collaborative filtering        | Giving recommendation lists of similar users in history profile for a new one |
| Content-based                  | Generating lists of content or goods with similar characteristics |
| Based on humanistic and social information | Providing recommendations by analyzing the information provided by users when they log in or register |
| Knowledge-based                | Providing recommendations based on external knowledge |

2.2. Definition and classifications of privacy in recommender system

Research on recommendation system has a long tradition, and privacy protection in recommendation system has been of considerable interest to the community in recent years. The word privacy has a subtly various definitions, one of the most widely accepted definition of privacy should be ‘an individual’s claim to control the terms under which personal information identifiable to the individual is acquired, disclosed or used.’ [1] The above definition of privacy emphasizes the control of the information owner over it, rather than the inability to use the complete seal. Therefore, considering the use of information in the network. In the case of use, document [2] defines privacy as a collection of information related to individuals with the right not to be collected, retained or punished by others.
And it is enough to be made public at a given time, in a particular way, to a certain extent, in accordance with the wishes of the owner.

2.3. Evaluations method of privacy risks
The previous section has shown that possibility of privacy leak is very high in the process of recommendation system work, so how to evaluate the severity of privacy leak or the risk of privacy leak? In order to calculate the severity of privacy risk, a lot of work has been made. And most remarkable works related to these methods are showed in Table 2.

| Algorithms                                      | Refs.         |
|------------------------------------------------|---------------|
| Computing and scoring risk                      | [3][4][5][6][7]|
| Computing the leakage of users’ sensitive info  | [8][9][10]   |

2.4. Ways by which privacy leakage occurs
There are three main stages in the process of personalized service that may cause users' privacy concerns.

The first stage is user modeling. In this stage, users' personalized information preferences and needs are acquired and user models are established. The methods of obtaining personalized preferences and needs include explicit ways provided by users actively, such as basic attributes such as user name, user input keywords, user preference feedback, etc. And implicit ways such as tracking user's browsing behavior (such as number of clicks, browsing time, marking bookmarks, etc.) by reasoning after tracking user's behavior. After obtaining personalized information, the user model is obtained. After that, the user's personalized information needs should be represented by appropriate data structure to facilitate system processing and use, and then matching calculation with the model. Then, the user model should be updated according to the change of user's personalized needs.

Privacy concerns that arise during this phase include improper access, collection, monitoring, analysis, merging, transmission, storage data, and so on.

The second stage is calculation. At this stage, similarity calculation should be carried out to make later recommendation. Privacy concerns include [11] improper analysis, improper merging of data, improper transmission of data, identity fraud, etc.

And the last stage is generating results of recommendation. This stage is to get the recommended results after matching calculation. Privacy concerns arising from this stage include improper analysis, improper merging of data, improper transmission of data, and misleading recommendation.

From the above definition, it can be seen that in the three stages of personalized service, personal privacy is involved, such as personal basic information collected, personal interests and preferences, personal browsing behavior and content, and personal information is processed, transmitted, stored and calculated without user permission.

2.5. Approaches and model for privacy protection
Current research on privacy protection can be classified into two groups as strategies and methods[12], and in the following part, they will be discussed individually.

2.5.1. Strategies
Privacy protection strategy is a comprehensive description of the website on the user's personal information collection, use and other aspects of the protection strategy, the purpose is to make users clear about the privacy protection strategy of the website, so the privacy protection strategy of the website usually uses a lengthy text description to present in front of users. The expression of privacy protection strategy is to use machine-readable way to express the privacy protection strategy of the website.
In order to provide a public privacy policy standard for websites, several standards and specifications have been proposed by the World Wide Web Alliance and large companies and widely adopted by the industry. Among them, privacy preference setting strategy P3P [13], enterprise internal privacy strategies EPAL and XACML [14].

P3P enables Web sites to have a standard, machine-readable, XML-described privacy protection policy file, including the description grammar and semantics of privacy information collection, storage and use. Web users can set privacy preference parameters in the P3P software (user agent) according to their personal needs, and the web site visited will send the XML policy reference file to the user agent; User agent can automatically or semi-automatically match the privacy policy of the web site with the user's privacy preference parameters, and if matched, explain the user's privacy preference and the hidden site. Private policy accords with each other, and users can access the web site smoothly; otherwise, users have the right to decide whether to give up access to the web site or to modify their personal privacy preference parameters to continue their access to the web site. This process usually takes the form of dialog boxes to facilitate users to make choices. The P3P strategy can be used for the whole website, or a part of the website.

EPAL is developed by IBM for the purpose of describing the internal privacy policies used to process user data that is mastered by the enterprise. The strategy described in this language can be applied and automatically complied with by enterprise management system, and it is also a document in XML format. EPAL uses rules to describe privacy policies and EPAL vocabulary to describe rules. Each privacy rule includes rules, data users, data usage methods, data categories, use purposes, etc. It may also include rules of use conditions and compliance. Privacy rules are degraded by priority. When in use, the system matches EPAL privacy rules with requests sent, and the matching results determine whether requests are allowed or rejected.

XACML, like EPAL, is a language for describing privacy policies within an enterprise. It also uses XML standard language. However, XACML has good universality, reusability, portability, distributability and extensibility [15]. It makes XAML not only have access to control policies, but also have more privacy policies than EPAL. So it has superiority.

In summary, we can see that the privacy protection policy framework has been developed and widely adopted. However, privacy protection policy framework in the actual use of the process is not well implemented by only through announcement and lack of real technical measures to ensure the smooth implementation. Users also have no measures to determine whether the site can be implemented in accordance with its privacy policy.

2.5.2. Methods
Methods used in privacy protection of recommendation system can be divided into two categories: one is based on statistics, the other is based on cryptography. In recommendation system, statistical method is a practical method which takes account of the calculation cost. By eliminating the sensitive features in data, it can hide the user's privacy. Encryption technology is a more secure but computational resource-consuming method. By using homomorphic encryption technology, data can be completely protected without losing information, which can ensure the recommendation quality of recommendation system.

2.5.2.1. statistical methods
Statistical method is a mature privacy protection method with high computational efficiency. Usually, by removing features, confusion, adding noise and so on, the sensitive information in the data file storing user's behavior is concealed and the cost of malicious data collectors and attacker is increased.

A variety of anonymisation algorithms with dissimilar anonymisation processes have been proposed by different authors. Models like k-anonymity, l-diversity and t-closeness are the most approved and accepted methods that deliver appropriate outcomes in anonymisation. K-anonymity [16-18] and l-diversity [19] are the main accepted models on privacy to quantify the degree of privacy, for sensitive information revelation against record linkage attack and attribute linkage attacks,
respectively. Supplementary secrecy models such as t-closeness[18] and m-invariance [19] are also presented for numerous attack in privacy scenarios. Numerous anonymising processes are applied to maximize the advantage of anonymise data-sets, as well as suppression [20], generalization [21, 22], anatomisation [23], slicing [24], disassociation [25]. And most remarkable works related to these methods are showed in Table 3.

Table 3 A comparison of works by statistics methods

| Techniques          | Methodologies                                                                 | Refs.          |
|---------------------|------------------------------------------------------------------------------|----------------|
| k-anonymity         | It guarantees the information for each individual stays undistinguished for the other k-1 individuals in a data-set | [16][17][18]   |
| l-diversity         | It guarantees that the values of sensitive parameters are dissimilar in each equivalence class. | [19][20]       |
| T-closeness         | They indicate that the sensitive information distribution within each QI compared with its distribution in the original dataset should be close. | [21][22]       |
| Differential privacy| This is done by adding proper noise to a small sample of user’s usage pattern. | [23]           |
| Slicing             | By slicing the dataset.                                                      | [24]           |
| User clustering      | Hiding sensitive information by clustering individuals into group.           | [26][27]       |

2.5.2.2. cryptography methods

The concept of homomorphic encryption is proposed in document [28]. The special nature of homomorphic encryption enables us to directly perform some operations on ciphertext instead of plaintext operations to achieve the same results, which does not affect the confidentiality of plaintext data. Homomorphic encryption methods will play an important role in cloud computing and multi-party secure computing.

Multi-party confidential computing is the core technology of privacy protection in information society and one of the research hotspots in the international cryptography field in recent years. Multi-party confidential computing includes two or more participants, but it is collectively called multi-party confidential computing in academia. For two participants, it is sometimes called bilateral confidential computing. Multi-party confidential computing enables private data to be owned. Many participants can cooperate to use these private data for computation without revealing the confidentiality of their private data, thus enabling people to maximize the use of private data without compromising the privacy of the data. Recommendation system is a typical multi-party computation. And most remarkable works related to these methods are showed in Table 4.

Table 4 A comparison of works by cryptography methods

| Categories                        | Tech.                  | Refs.          |
|-----------------------------------|------------------------|----------------|
| Neighborhood-based collaborative filtering | Partially homomorphic encryption | [29][30][31] |
| Machine learning-based collaborative filtering | Partially homomorphic encryption | [32][33][34][35] |
| Machine learning-based collaborative filtering | Fully homomorphic encryption | [36]           |
| Content-based recommendation       | Partially homomorphic encryption | [37]           |
3. Future directions

(1) Efficiency
The efficiency of recommendation system depends largely on the time complexity and space complexity of the algorithm. Currently, cryptography-based privacy protection methods have advantages in security but consume more computing resources, which can not meet the immediacy of large recommendation systems. Statistical and confusion-based methods are often ineffective in dealing with malicious attackers who hold contextual information. Therefore, improving the system efficiency of cryptography is an important research direction.

(2) Malicious behavior discovery
For all kinds of cyber attacks against privacy, early detection and warning for victims is crucial. It is very useful to analyze and study the various malicious behavior patterns of privacy disclosure, and then to make the corresponding strategies.

(3) Adequacy
At present, most websites are compatible with privacy protection technologies. But privacy protection technologies such as P3P still have some incompleteness, which can not adequately express users' privacy protection needs. At the same time, these privacy protection technologies lack the corresponding technical and legal measures to ensure the implementation of their privacy protection strategies. Whether the expressed privacy protection needs can be realized is uncertain for users. Therefore, it is a future research direction to study the corresponding privacy description language which can fully describe users' privacy requirements and related technical guarantee measures. At the same time, it is also a research direction to study the methods and measures to facilitate users to understand and grasp whether the needs of personal privacy protection can be realized.

4. Conclusions
In this paper, we surveyed the literature related to privacy preserving recommendation services. We first presented the system architecture of personalized recommendation services, commonly adopted recommendation techniques, and privacy issues posed by personalized recommendation services. Then we described existing privacy-preserving techniques for recommendation services, which are classified into two broad categories: privacy-preserving policies and privacy-preserving methods. And a comparison of the existing works on privacy-preserving recommendation was made. Finally, we provided some discussion on future research directions.

References
[1] WANG Yang, KOBSA A. Technical solutions for privacy-enhanced personalization[J]. Intelligent User Interfaces, 2009, 5704(949):353-376.
[2] ZANG Cheng. Research on Key Techniques of Privacy Preservation in Personalized Search[D]. Zhejiang University, 2008 : 15-16.
[3] Renner, Christoph. Privacy in online social networks[D]. Swiss Federal Institute of Tech., Zurich, 2010 : 11-13.
[4] Liu K, Terzi E. A framework for computing the privacy scores of users in online social networks [J]. ACM Transactions on Knowledge Discovery from Data (TKDD). 2010;5(1):6.
[5] Maximilien EM, Grandison T, Liu K, et al. Enabling privacy as a fundamental construct for social networks. In: Computational Science and Engineering, 2009. CSE’09. International Conference on. vol. 4. IEEE,. p. 1015–1020.
[6] Srivastava A, Geethakumari G. Measuring privacy leaks in online social networks. In: Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on. IEEE,. p. 2095–2100.
[7] Domingo-Ferrer J. Rational privacy disclosure in social networks. In: International Conference on Modeling Decisions for Artificial Intelligence. Springer; 2010. p. 255–265.
[8] Aghasian E, Garg S, Gao L, et al. Scoring Users Privacy Disclosure Across Multiple Online Social Networks. IEEE Access. 2017;.
[9] Nepali RK, Wang Y. Sonet: A social network model for privacy monitoring and ranking. In: 2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops. IEEE; p. 162–166.

[10] Talukder N, Ouzzani M, Elmagarmid AK, et al. Privometer: Privacy protection in social networks. In: Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on. IEEE; p. 266–269.

[11] Wang H, Lee M K O, Wang C. Consumer Privacy Concerns about Internet Marketing[J]. Communications of the ACM, 1998, 41(3):63-70.

[12] Zhang L, Li XY, Lei J, et al. Mechanism design for finding experts using locally constructed social referral web [J]. Parallel and Distributed Systems, IEEE Transactions on. 2015;26(8):2316–2326.

[13] Cranor L F. P3P: Making privacy policies more useful[J]. IEEE Security and Privacy Magazine, 2003, 1(6):50-55.

[14] Stufflebeam W H, He Q, Jain N. Specifying privacy policies with P3P and EPAL:lessons learned[C]// Acm Workshop on Privacy in the Electronic Society. 2004.

[15] Wolter C, Schaad A, Meinel C. Deriving XACML Policies from Business Process Models[C]// International Conference on Web Information Systems Engineering. 2007.

[16] Sweeney L. k-anonymity: A model for protecting privacy [J].International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002;10(05):557–570.

[17] Wong RCW, Li J, Fu AWC, et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. p. 754–759.

[18] Fung B, Wang K, Yu PS. Anonymizing classification data for privacy preservation [J]. Knowledge and Data Engineering, IEEE Transactions on. 2007;19(5):711–725.

[19] Machanavajjhala A, Kifer D, Gehrke J, et al. l-diversity: Privacy beyond k-anonymity [J]. ACM Transactions on Knowledge Discovery from Data (TKDD). 2007;1(1):3.

[20] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE; 2007. p. 106–115.

[21] Li N, Li T, Venkatasubramanian S. Closeness: A new privacy measure for data publishing [J]. IEEE Transactions on Knowledge and Data Engineering. 2010;22(7):943–956.

[22] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. IEEE; 2007. p. 106–115.

[23] Xiao X, Tao Y. Anatomy: Simple and effective privacy preservation. In: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment; 2006. p. 139–150.

[24] Dwork C. An ad omnia approach to defining and achieving private data analysis. In: Privacy, Security, and Trust in KDD. Springer; 2008. p. 1–13.

[25] Li T, Li N, Zhang J, et al. Slicing: A new approach for privacy preserving data publishing[J]. IEEE transactions on knowledge and data engineering. 2012;24(3):561–574.

[26] Terrovitis M, Mamoulis N, Liagouris J, et al. Privacy preservation by disassociation [J]. Proceedings of the VLDB Endowment. 2012;5(10):944–955.

[27] Canny J. Collaborative filtering with privacy[C]// IEEE Symposium on Security & Privacy. 2002.

[28] Canny J F. Collaborative filtering with privacy via factor analysis[C]// International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2002.

[29] R Rivest, L Adleman, M Dertouzos. On Data Banks and Privacy Homomorphisms [A]. Foundations of Secure Computation[C]. Liverpool: Academic Press, 1978. 169 - 177.

[30] Erkin Z, Veugen T, Toft T, et al. Generating Private Recommendations Efficiently Using Homomorphic Encryption and Data Packing[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3):0-0.
[31] Badsha S, Yi X, Khalil I. A practical privacy-preserving recommender system. Data Sci Eng 2016;1(3):161–77.

[32] Badsha S, Yi X, Khalil I, Bertino E. Privacy preserving user-based recommender system. In: Proceedings of the 37th IEEE International Conference on Distributed Computing Systems; 2017 Jun 5–8; Atlanta, GA, USA. Los Alamitos: IEEE Computer Society Press; 2017. p. 1074–83.

[33] Nikolaenko V, Ioannidis S, Weinsberg U, Joye M, Taft N, Boneh D. Privacy preserving matrix factorization. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security; 2013 Nov 4–8; Berlin, Germany. New York: Association for Computing Machinery, Inc.; 2013. p. 801–12.

[34] Nikolaenko V, Weinsberg U, Ioannidis S, Joye M, Boneh D, Taft N. Privacy preserving ridge regression on hundreds of millions of records. In: Proceeding of the 2013 IEEE Symposium on Security and Privacy; 2013 May 19–22; Berkeley, CA, USA. Los Alamitos: IEEE Computer Society Press; 2013. p. 334–48.

[35] Hu S, Wang Q, Wang J, Chow SSM, Zou Q. Securing fast learning! Ridge regression over encrypted big data. In: Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA; 2016 Aug 23–26; Tianjin, China. Los Alamitos: IEEE Computer Society Press; 2016. p. 19–26.

[36] Kim S, Kim J, Koo D, Kim Y, Yoon H, Shin J. Efficient privacy-preserving matrix factorization via fully homomorphic encryption: Extended abstract. In: Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security; 2016 May 30–Jun 3; Xian, China. New York: Association for Computing Machinery, Inc.; 2016. p. 617–28.

[37] Jiang J, Zheng Y, Yuan X, Shi Z, Gui X, Wang C, et al. Towards secure and accurate targeted mobile coupon delivery. IEEE Access 2016;4:8116–26.