Research on the Association of Mobile Social Network Users Privacy Information Based on Big Data Analysis

Pingshui Wang[1, *], Zecheng Wang[1] and Qinjuan Ma[1]

Abstract: The issue of privacy protection for mobile social networks is a frontier topic in the field of social network applications. The existing researches on user privacy protection in mobile social network mainly focus on privacy preserving data publishing and access control. There is little research on the association of user privacy information, so it is not easy to design personalized privacy protection strategy, but also increase the complexity of user privacy settings. Therefore, this paper concentrates on the association of user privacy information taking big data analysis tools, so as to provide data support for personalized privacy protection strategy design.

Keywords: Big data analysis, mobile social network, privacy protection, association.

1 Introduction

As the main social network platform in the era of big data, the security and privacy of mobile social network directly affect the enthusiasm of mobile social network users to participate in network activities. The privacy protection of network users deserves great attention from social all circles.

Nowadays, with the wide application of Web 2.0 technology, mobile social network, as a new interactive mode of the Internet, is attracting more and more attention. It has become a new media platform with the largest number of users and the greatest impact on communication, such as Facebook, Twitter, Renren, Weibo, Weixin and so on. It provides a convenient service for people to chat, make friends and share information in time and attracts a large number of users to participate. The advent of the era of big data has exacerbated the risk of privacy leakage in social networks. Because mobile social networks are open, shared and connected, with the help of powerful search engines, users' privacy information is more likely to be snooped, collected and illegally used. Meanwhile, with the help of large data analysis tools, the users' related information could be mined from users' general information, which may also cause the leakage of user privacy and bring a certain security threat to the relevant individuals and organizations. Therefore, how to make social networks better protect the privacy of data owners while satisfying user communication and pattern knowledge discovery has become a hot issue in recent years.

[1] School of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu, 233030, China.

[2] School of Business Administration, Anhui University of Finance & Economics, Bengbu, 233030, China.

[*] Corresponding Author: Pingshui Wang. Email: 120081049@aufe.edu.cn.

JIHPP. doi:10.32604/jihpp.2019.05942

At present, research on privacy protection of social network users mainly focuses on privacy protection of social network data publishing and access control of social network [Liu, Wang and Yang (2014); Sun, Yu, Kong et al. (2014); Tai, Yu, Yang et al. (2011) ; Yang, Huang, Li et al. (2018)]. There are many researches on privacy preserving technology for social network data dissemination mainly using anonymous processing technology [Campan and Truta (2008); Gong, Lan, Pei et al. (2016); Hay, Miklau, Jensen et al. (2008); Lan and Tian (2014); Wang, Xie, Zheng et al. (2011); Wang, Xiong, Pei et al. (2018); Wang, Zhang, Feng et al. (2014)], so that published social network data can meet the needs of data analysis, but also can protect user privacy is not leaked. The research of social network access control technology focuses on the design of social network access control model [Carminati, Ferrari and Perego (2006); Nurmamat and Kaysar (2010); Wang, Wang, Guo et al. (2018); Zhou, Jiang and Sun (2010)] to solve the problem of authorized access to social network data. However, there is little re-search on the relationship between user privacy information in the existing literature, which makes it difficult to design personalized privacy protection strategies, and in-creases the complexity of user privacy protection settings. In this paper, data mining and large data analysis tools are used to analyze the individual and group attributes of mobile social network users and extract the association of user privacy information, so as to provide data support for the design of personalized privacy protection strategy.

2 The related concepts

2.1 Mobile social network

Mobile social network (MSN) is a special social group formed by using mobile terminal devices through Internet applications such as Facebook, Twitter, Renren, Blog, Weibo, Weixin, QQ, etc. Its essence is to provide a mobile communication platform for sharing interests, hobbies, status and activities. With the development of mobile devices and the new generation information technology such as Internet, cloud computing, big data and artificial intelligence, mobile social network has penetrated into all aspects of people's daily work, study and life with the characteristics of real-time, openness, mobility, personalization, and become the main platform of people's ideological communication, emotional exchange, data communication and information sharing. It also brings people a zero distance social experience.

2.2 Big data

Big data refers to the data set that cannot be captured, managed and processed by conventional software tools within a certain period of time. Mobile social networks are generating new data almost every minute, and data types and scales are growing exponentially at an unprecedented rate. On the whole, they present 4V characteristics, namely, large data scale (**V**olume), fast processing speed (**V**elocity), multiple data types (**V**ariety), low value density (**V**alue). This provides a research basis for social network data analysis and researchers, and facilitates the development of relevant data analysis, pattern recognition and knowledge discovery.

2.3 Association rules

Association rules refer to interesting association and rules between attributes hidden in large data sets. They are important research contents in data mining and are widely used in the financial field [Han and Kamber (2012)]. However, in the mobile social network user attributes data set, association rules mining and large data analysis techniques can also find out the association between social network user attributes (some of which may be the user's privacy information), thus providing data support for the privacy protection policy settings of social network users.

3 Big data analysis techniques

As we all know, one of the characteristics of big data is the low value density, that is, in a large amount of data may only be small data is valuable, how to extract the value of which requires the support of big data analysis technology. There are many big data analysis techniques, such as data mining, statistical analysis, model prediction, visual analysis, and so on. The following is a brief introduction to the main technologies related to user attribute data processing in social networks.

3.1 MapReduce

MapReduce is a computing model, framework and platform for parallel processing of big data. It was first proposed by Google for parallel computing model and method of large-scale data processing. Later, it was implemented in Hadoop with open source, and its function was significantly enhanced [Lin (2017)].

Hadoop MapReduce highly abstracts complex parallel computing processes running on large-scale clusters into two functions: Map and Reduce, MapReduce uses a “divide and conquer” strategy, that is, a large-scale data set stored in a distributed file system is split into many separate pieces. It can be processed by multiple Map tasks in parallel, and the processed intermediate results can be used as input to the Reduce task to produce the desired results: < *key*, *value* > pair (such as Tab. 1).

Table 1: Map and Reduce functions

Function	input	output	specification
Map	<*k1*,*v1*>	List(<*k2*,*v2*>)	1. The small dataset is further resolved into a batch of <*key*, *value*> pairs, and processed in the input Map. 2. Each input <*k1*, *v1*> will output a batch of <*k2*, *v2*>. <*k2*, *v2*> is the intermediate result of computation.
Reduce	<*k2*, List(*v2*)>	<*k3*,*v3*>	The List (*v2*) in the intermediate result <*k2*, List (*v2*) > of the input indicates a group of value belonging to *k2*.

3.2 Association rule mining

Association rules are implicit forms such as $X \rightarrow Y$, in which X and Y are called precedence and successor of association rules respectively. Meanwhile, association rules have support and confidence degree.

Let $I=\{I_1,I_2,\cdots,I_m\}$ be a set of items [Han and Kamber (2012)]. Let D be a database of transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated to an identifier, call TID. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with *support* s, where s is the percentage of transactions in D that contains $A \cup B$. The rule $A \Rightarrow B$ has *confidence* c in the transaction set D. That is,

$$sup(A \Rightarrow B) = P(A \cup B) = \frac{|A \cup B|}{|D|} \tag{1}$$

$$conf(A \Rightarrow B) = P(B \mid A) = \frac{|A \cup B|}{|A|} \tag{2}$$

Where $|A|$ is named as the *support count* of the set of items A in the set of transactions D, as denoted by $sup\_count(A)$. Item A occurs in a transaction T, if and only if $A \subseteq T$. Rules that satisfy both a *minimum support threshold* (min_sup) and a *minimum confidence threshold* (min_conf) are called strong. A set of items referred to as an *itemset*. An itemset that contains k items is a k-itemset. Itemsets that satisfy min_sup is named as *frequent itemsets*. All strong association rules result from frequent itemsets.

In general, association rule mining can be viewed as a two-step process:

- *Finding all frequent itemsets*: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, Min_Sup.
- *Generate strong association rules from the frequent itemsets*: By definition, these rules must satisfy minimum support and minimum confidence.

4 Association analysis of user privacy information in mobile social network

Users in the social network provide a lot of real personal information, including personal data, education and work experience, contact information, photos, speech and online activities. Moreover, chat information, video information, picture information in the mobile social network has increased dramatically, showed by kinds of forms such as structured, semi-structured, unstructured, and so on. The huge amount of information conforms to the typical “4V” characteristics of big data. Traditional data analysis tools are unable to deal with such complex and large-scale social network data, and need to use special big data processing tools to deal with it effectively.

In order to use big data analysis technologies to analyze the association of user privacy information in social networks, we randomly selected some user attributes data of a certain social network as a sample. There are 50,000 users in the sample, each of whom contains the attributes such as name, sex, birthday, blood type, occupation, hobbies, mobile phone, and mailbox. And each attribute contains the option of whether to publish or not. We mainly analyze whether the data of each attribute is publicly related, so as to simplify the privacy settings of related attributes in user account registration.

4.1 Privacy analysis of single attribute data

Assuming that the support degree is 60%, the following results are obtained by statistical analysis of single attribute data of sample data (such as Tab. 2):

Table 2: Privacy statistics for single attribute data

Attribute	Number of users that do not open attribute data to public	support degree	Is private or not?
Name	32,165	64%	is
Sex	8,418	17%	
Birthday	44,207	88%	is
Blood type	21,262	43%	
Career	10,184	20%	
Hobbies	8,509	17%	
Mobile phone	42,802	86%	is
E-mail	39,274	**79%**	is

Statistical results show that more than 60% of users regard name, birthday, mobile phone and e-mail as their personal privacy, so the system automatically sets these attributes to be closed and other attributes to be open by default when registering social network users' accounts.

Among 8,418 users who regarded sex attribute data as privacy, 16% are male and 84% are female, as shown in Tab. 3. The results show that females are more aware of sex data privacy protection than males.

Table 3: Privacy statistics for sex attribute data

Attributes	Number of males that don’t open attribute data to public	Number of females that do not open attribute data to public
Sex	1,356(16%)	7062(84%)

4.2 Privacy association analysis of double-attribute data

Among the 8,418 users who treat sex attribute data as privacy, users regarding other open attribute data as privacy are shown in Tab. 4. Therefore, the system can automatically complete the default settings of related attributes by real-time detecting the privacy settings of gender attributes when registering user accounts in social networks, so as to simplify user operations and protect user's related attribute data.

Table 4: Privacy statistics for dual attribute data (including gender attribute)

Attribute	Number of users that do not open attribute data to public	Is private or not (confidence degree 60%)?
Sex, blood type	6,216（74%）	is
Sex, occupation	5,478（65%）	is
Gender, hobbies	3,613（43%）	

4.3 Privacy association analysis of multi-attribute data

Among the 8,418 users who regard sex and blood group attributes as privacy, users regarding other open attribute data as privacy are shown in Tab. 5. Similarly, in the social network user account registration, the system can automatically complete the default settings of related properties.

Table 5: Three-attribute data (including sex, blood type) privacy statistics

Attribute	Number of users that don't open attribute data to public	Is private or not (confidence degree 60%)?
Sex, blood type, hobby	5,011（60%）	is
Sex, blood type, occupation	3,062（36%）	

In addition, we can also combine the public attributes with the private attributes for multi-attribute privacy association analysis to find out the relationship between the public attributes, the private attributes, so as to provide a reference for the design of user privacy protection strategy.

5 Conclusions

As one of the technological products of Web 2.0, mobile social network has become the main platform for people to disseminate information and communicate through the Internet. The emergence of big data tools exacerbates the risk of user privacy leakage in mobile social networks. In recent years, the security and privacy protection of mobile social network users have become a hot issue in academia and industry. However, the existing research seldom pays attention to the relationship between users' privacy information, which brings inconvenience to the design of user privacy protection strategy and increases the complexity of user privacy protection settings. In order to provide data support for the design of personalized privacy protection strategy, this paper analyzes the relationship between privacy information of mobile social network users through big data analysis tools. In the next step, we will build an authorization model to support the personalized privacy preferences of mobile social network users to achieve a more flexible and practical definition of privacy policy, and conduct simulation experiments and comparative analysis to comprehensively solve the problem of user privacy leakage in mobile social network applications.

Acknowledgment: We thank the anonymous reviewers and editors for their very constructive comments. This work was supported by the National Social Science Foundation Project of China under Grant 16BTQ085.

References

Campan, A.; Truta, T. (2008): A clustering approach for data and structural anonymity in social networks. *Proceedings of the 2nd ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD*, pp. 33-54.

Carminati, B.; Ferrari, E.; Perego, A. (2006): Rule-based access control for social networks. *On the Move to Meaningful Internet Systems: OTM'06 Workshops, Springer Berlin Heidelberg, LNCS* 4278, pp. 1734-1744.

Gong, W.; Lan, X, Pei, X.; Yang, L. (2016): Privacy protection methods based on k_degree anonymity in social networks. *Journal of Electronics*, vol. 44, no. 6, pp. 1437-1444.

Han, J.; Kamber, M. (2012): *Data Ming: Concepts and Techniques*. Beijing: China Machine Press, pp. 1-40.

Hay, M.; Miklau, G.; Jensen, D.; Towsley, D. (2008): Resisting structural identification in anonymized social networks. *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 102-114.

Lan, L.; Tian, L. (2014): Preserving social network privacy using edge vector perturbation. *Proceedings of the International Conference on Information Science and Cloud Computing Companion*, pp. 188-193.

Lin, Z. Y. (2017): *Big data technology principles and Applications (Second Edition)*. Beijing: People's Education Press.

Liu, X.; Wang, B.; Yang, X. (2014): Survey on privacy preserving techniques for publishing social network data. *Journal of Software*, vol. 25, no. 3, pp. 576-590.

Nurmamat, H.; Kaysar, R. (2010): Representation of RBAC model with negative authorization in OWL and conflict detection. *Computer Engineering and Applications*, vol. 46, no. 30, pp. 82-85.

Sun, C.; Yu P.; Kong, X.; Fu, Y. (2014): Privacy preserving social network publication against mutual friend attacks. *Transactions on Data Privacy*, vol. 7, no. 2, pp. 71-97.

Tai, C.; Yu, P.; Yang, D.; Chen, M. (2011): Privacy-preserving social network publication against friendship attacks. *Proceedings of the SIGKDD*, pp. 1262-1270.

Wang, M. J.; Wang, J.; Guo, L. H.; Harn, L. (2018): Inverted XML access control model based on ontology semantic dependency. *Computers, Materials & Continua*, vol. 55, no. 3, pp. 465-482.

Wang, R.; Zhang, M.; Feng, D.; Fu, Y. (2014): A clustering approach for privacy-preserving in social networks. *Lecture Notes in Computer Science*, pp. 193-204.

Wang, X.; Xiong, C.; Pei, Q. Q.; Qu, Y. Y. (2018): Expression preserved face privacy protection based on multi-mode discriminant analysis. *Computers, Materials & Continua*, vol. 57, no. 1, pp. 107-121.

Wang, Y.; Xie, L.; Zheng, B.; Lee, K. (2011): Utitily-oritented k-anonymization on social networks. *Database Systems for Advanced Applications: 16th International Conference*, pp. 78-92.

Yang, Z.; Huang, Y. F.; Li, X.; Wang, W. Y. (2018): Efficient secure data provenance scheme in multimedia outsourcing and sharing. *Computers*, *Materials & Continua*, vol. 56, no. 1, pp. 1-17.

Zhou, X.; Jiang, X.; Sun, K. (2010): Research and improvement of RB-RBAC model. *Information Security and Communication Secrecy*, no. 4, pp. 100-102.