Statistics and Application 统计学与应用, 2022, 11(3), 686-693
Published Online June 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.113073

Hans汉斯

大数据时代隐私披露意愿的统计分析

吴华清，梁佳慧，张　钰，刘仲阳

曲阜师范大学统计与数据科学学院，山东　曲阜

收稿日期：2022年5月29日；录用日期：2022年6月20日；发布日期：2022年6月29日

摘　要

随着大数据时代的发展，APP中频繁发生的隐私泄露问题影响用户的隐私披露意愿，而用户隐私披露意愿又与APP发展有关。为了帮助APP更好发展，本文将对APP用户隐私披露意愿的情况和影响因素进行研究，为其提供理论依据。首先，使用Python对用户隐私披露情况进行可视化，发现多数用户认为APP过度收集个人信息。其次，基于隐私计算理论和沟通隐私管理理论，从用户情感和行为角度确定影响因素，设计收集问卷，构建结构方程，得到感知有用性、习惯性、社交媒体信任性对披露意愿有正向影响，感知风险性、隐私控制性、隐私关注性有负向影响。最后，建立随机森林模型探究影响用户隐私披露意愿的核心因素，以准确率为标准，感知风险性和隐私关注性是核心因素；以GINI指数为标准，习惯性和感知有用性是核心因素。

关键词

大数据技术，隐私披露意愿，影响因素

Statistical Analysis of Factors Influencing Willingness to Disclose Privacy in the Era of Big Data

Huaqing Wu, Jiahui Liang, Yu Zhang, Zhongyang Liu

School of Statistics and Data Science, Qufu Normal University, Qufu Shandong

Received: May 29th, 2022; accepted: Jun. 20th, 2022; published: Jun. 29th, 2022

Abstract

With the development of the big data era, the frequent privacy leakage problems in APPs affect users' willingness to disclose their privacy, which in turn is related to APP development. To help

文章引用: 吴华清, 梁佳慧, 张钰, 刘仲阳. 大数据时代隐私披露意愿的统计分析[J]. 统计学与应用, 2022, 11(3): 686-693. DOI: 10.12677/sa.2022.113073

APPs develop better, we study the situation and influencing factors of APP users' willingness to disclose privacy to provide a theoretical basis for it. First, we use Python to visualize user privacy disclosure and find that most users believe that the app over-collects personal information. Secondly, based on privacy computing theory and communication privacy management theory, we identify the influencing factors from the perspective of users' emotions and behaviors, design a collection questionnaire, construct structural equations, and obtain that perceived usefulness, habituation, and social media trust have positive effects on disclosure intention, and perceived riskiness, privacy control, and privacy concern have negative effects. Finally, we build a random forest model to explore the core factors, with perceived riskiness and privacy concern as the core factors in terms of accuracy, and habituation and perceived usefulness as it in terms of GINI index.

Keywords

Big Data Technology, Privacy Disclosure Willingness, Influencing Factors

1. 引言

近年来，互联网技术的发展促使各种社交平台不断涌现。用户习惯于在社交平台上与他人沟通以获取信息，例如以微博、微信为代表的社交软件，以春雨医生、阿里健康为代表的医疗服务平台等。研究表明，用户在网络环境中的自我披露相较于现实世界更加容易[1]。然而，用户在进行个人信息披露的同时也不得不承担信息泄露的风险。频发的隐私泄露问题一定程度上影响了用户的隐私披露行为意愿，而社交平台的发展离不开用户的隐私披露行为。以在线医疗平台为例，患者只有提供完善的信息才能获得准确诊断，充分发挥在线医疗平台的作用[2]。因此，为了促使社交平台更好地发展，解决用户对隐私泄露的担忧与主动进行信息披露之间的矛盾显得尤为重要。

本文基于隐私计算理论、沟通隐私管理理论和随机森林理论，从情感和行为两个角度出发，研究用户隐私披露意愿的影响因素，以期为用户隐私安全问题的解决提供对策。

2. 理论基础

隐私计算理论和沟通隐私管理理论

隐私计算理论是指用户在进行信息披露时会衡量这一行为的预期收益与风险，如果预期收益等于或大于风险，那么用户会进行隐私披露；反之，就不会进行隐私披露[3]。已有学者将隐私计算理论与其他理论结合起来研究隐私披露问题。韩普和黄燕杰[4]将隐私计算理论和从众效应理论相结合构建了在线健康社区隐私悖论模型。袁向玲和牛静[5]基于 APCO 理论和隐私计算理论研究社交媒体用户的隐私政策对自我表露的影响机制。

沟通隐私管理理论认为个体相信自己有控制个人隐私信息的权利，并设立隐私边界来管理隐私信息[6]。用户在权衡信息披露的收益和风险后选择披露或是隐藏个人信息[7]。研究表明，用户认为自己控制信息的能力越强，对控制隐私边界的自主性越强，其在网络上的保护意识就越低，就更愿意披露自己的个人信息[8]，用户在注册账号时事先阅读隐私协议，只有用户认为该协议能充分保护自己的隐私安全，用户才会愿意进行个人信息的披露。因此，用户对社交平台的信任可能是影响用户隐私披露意愿的重要

因素。

3. 隐私披露意愿的词云图

基于 APP 用户反馈的使用大数据，进行词云图分析发现：越来越多的用户在面对隐私披露问题时感觉到 APP 过多收集个人信息，APP 不恰当使用他们的个人信息等等，由下图 1 可见一斑。

Figure 1. Cloud map of users' perception of privacy disclosure words
图 1. 用户面对隐私披露感受词云图

4. 调查问卷

4.1. 问卷设计及数据收集

本文使用问卷调查方法收集数据。问卷内容设计主要分为两个部分，第一个部分内容是被调查对象的人口统计变量。第二个部分内容采用李克特五级量表，设置 6 个潜在变量，如表 1 所示，包括感知有用性、感知风险性、习惯性、隐私控制、社交媒体信任、隐私关注性，它们具体表现为 20 个测量变量，具体可见调查问卷。测量变量的选项分别为非常不同意、不同意、不一定、同意、非常同意。

提出各潜在变量的假设检验如下：

H1：感知有用性正向影响隐私披露意愿

H2：感知风险性负向影响隐私披露意愿

H3：隐私控制正向影响隐私披露意愿

H4：隐私关注性负向影响隐私披露意愿

H5：习惯性正向影响隐私披露意愿

H6：社交媒体信任正向影响隐私披露意愿

由《北京市消协发布手机 APP 个人信息安全调查报告》发布[13]，79.23%的人认为手机 APP 上的个人信息不安全，本文设置绝对误差为 5%，置信度为 95%，标准差为 0.17，经计算得到符合以上标准的最小样本量为 255。考虑到问卷有效情况，本调查发放 320 份问卷。经过平台比较，本次调查选择使用见数 Credamo 线上问卷平台，设计问卷和收集数据。平台搭建了大数据分析功能，可以保证筛选最有效的数据，这保障了问卷数据的质量。最终得到有效问卷 300 份，有效率为 93.75%。本次调查范围为全国，

Table 1. Potential variables and explanations
表 1. 潜在变量及解释

名称	潜在变量解释
感知有用性	感知有用性是指用户在 APP 中披露隐私，而 APP 给用户带来各方面的价值，如情感、金钱、服务价值。当在 APP 中披露隐私，用户获得价值，并且在同时刻不会有实质损失时，会提高用户的隐私披露意愿。
感知风险性	感知风险性是指在 APP 中披露隐私时，用户对该行为可能造成的损失的估计。APP 中可能存在盗窃隐私等，造成损失。程慧平等人在探究社交媒体用户隐私披露意愿影响因素时也证实了感知风险性对隐私披露意愿有负向影响[9]。
隐私控制	隐私控制是指用户在 APP 中披露个人隐私的控制能力。当隐私在可控范围内，不易泄露时，用户能确保个人信息的安全性，其保护意识降低，隐私披露意愿升高。郭海玲等人也证实了感知信息控制会正向影响信息披露意愿[10]。
隐私关注性	隐私关注性是指用户对于 APP 存在隐私泄露的安全问题所产生的内部担忧。警惕心高的用户对于隐私关注度会高，所以当需要披露个人信息时，用户会关注个人隐私，防范意识增强，即会降低隐私披露意愿。
习惯	习惯是逐渐养成，不易改变的。不同 APP 经常向用户要取信息，用户养成习惯性心理，将个人信息提供给 APP。S. Mouakket 等人也确认习惯对社交媒体用户隐私信息披露有正向影响[11]。用户在 APP 中习惯性发布信息，如写博客或发照片等行为，也会提高用户的隐私披露意愿。
信任	信任被定义为预期对方即使存在选择机会，仍将选择符合自己期望的行为[1]。社交媒体信任主要是：对于 APP 平台的信任和对于 APP 中用户的信任。当两者信任提高，根据隐私计算理论，用户会重新计算隐私披露行为的收益与风险，从而更积极地参与到社交媒体互动中[12]。

男女比例接近，主要为 20~29 岁和 30~39 岁，本科学历及以上占 90%左右，接触 APP 时间大约在 2~20 年。

4.2. 信度与效度检验

对测度模型进行信度和效度检验，分析结果如表 2 所示。各因子的信效度衡量指标测度分别为：Cronbach's α 和组合信度(CR)以及平均方差提取量(AVE)，如下表所示，结果表明 Cronbach's α 均大于 0.7，组合信度(CR)均大于 0.5，平均方差提取量(AVE)均大于 0.3，因此表明所分析的数据具有较好的信度与效度。

Table 2. Standard load of each factor, Cronbach's α, CR, AVE
表 2. 各因子标准负荷、Cronbach's α、CR、AVE

潜变量	观察变量	标准负荷	Cronbach's α	CR	AVE
感知有用性	PU1	0.809	0.852	0.8254	0.6122
	PU2	0.732			
	PU3	0.804			
习惯性	XG1	0.645	0.734	0.7545	0.5088
	XG2	0.815			
	XG3	0.668			
感知风险性	PR1	0.732	0.849	0.6428	0.4746
	PR2	0.643			
隐私控制性	IC1	0.531	0.701	0.7755	0.544
	IC2	0.833			
	IC3	0.810			

Continued

社交媒体信任性	SMT1	0.629	0.888	0.5878	0.3237
	SMT2	0.555			
	SMT3	0.517			
隐私关注性	PC1	−0.607	0.703	0.6202	0.354
	PC2	−0.644			
	PC3	−0.528			

5. 隐私披露意愿的影响因素及强度

5.1. 结构方程模型的适用性测度

模型的适用性检验，结果如表 3 所示，可以通过对所分析的指标值的取值与指定的推荐值进行比较，通过对比，可以得到，除了 NNFI 指标值 0.899，略小于推荐值 0.9 以外，其他指标值均符合推荐值的范围之内。表明整体上，模型的适用性高。

Table 3. Applicability measure of the model
表 3. 模型适用性测度

指标	指标含义	推荐值	分析值
χ^2	卡方值	愈小愈好	346.399
χ^2/df	卡方值比自由度	<3.0	2.325
GFI	拟合优度	>0.9	0.891
AGFI	调整的拟合优度	>0.8	0.847
RMSEA	近似误差的均方根	<0.08	0.067
NNFI	非规范拟合指数	>0.9	0.899
NFI	规范拟合指数	>0.9	0.94
CFI	拟合优度指数	>0.9	0.939

5.2. 假设检验结果分析

Amos 结构方程模型的分析如下图 2，从图中结果可以得到，满足假设：感知有用性、习惯性、社交媒体信任性对隐私披露意愿均存在显著的正向影响，影响的路径系数分别为：0.236、0.246、0.078。感知风险性、隐私控制性及隐私关注性对隐私披露意愿存在显著的负向影响，分别为：–0.042、–0.144、–0.335。

5.3. 基于随机森林模型的研究

5.3.1. 数据分析与模型拟合

本节将具体分析各个因素对隐私披露意愿的影响程度大小。基于上述效度检验筛选后的变量，将关于隐私披露意愿的 3 个问题量化后取平均值，以此作为因变量，其余变量作为自变量，各自变量标按照强度由大到小依次赋值 5，4，3，2，1，而后进行随机森林模型的拟合。

首先基于全部的特征，让决策树数目遍历区间[1, 500]中的全部整数，拟合 500 个随机森林模型，并相应的对训练数据进行预测，得到的误差图如图 3 所示。

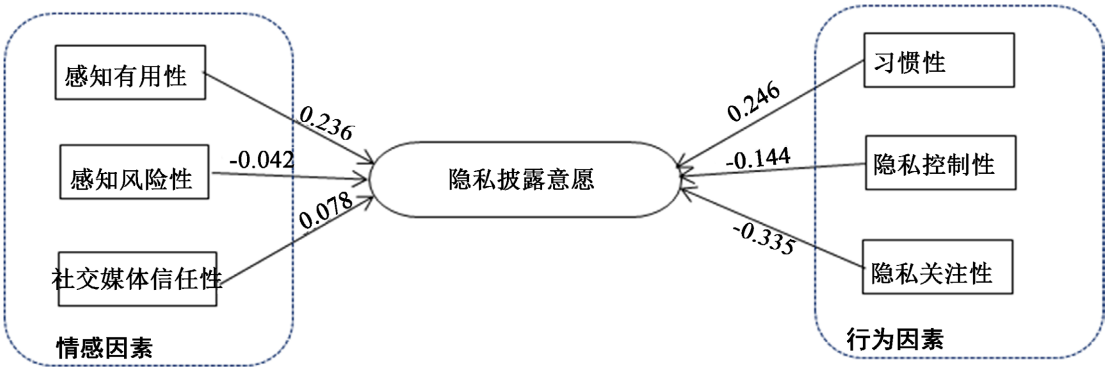

Figure 2. Amos path analysis diagram
图 2. Amos 路径分析图

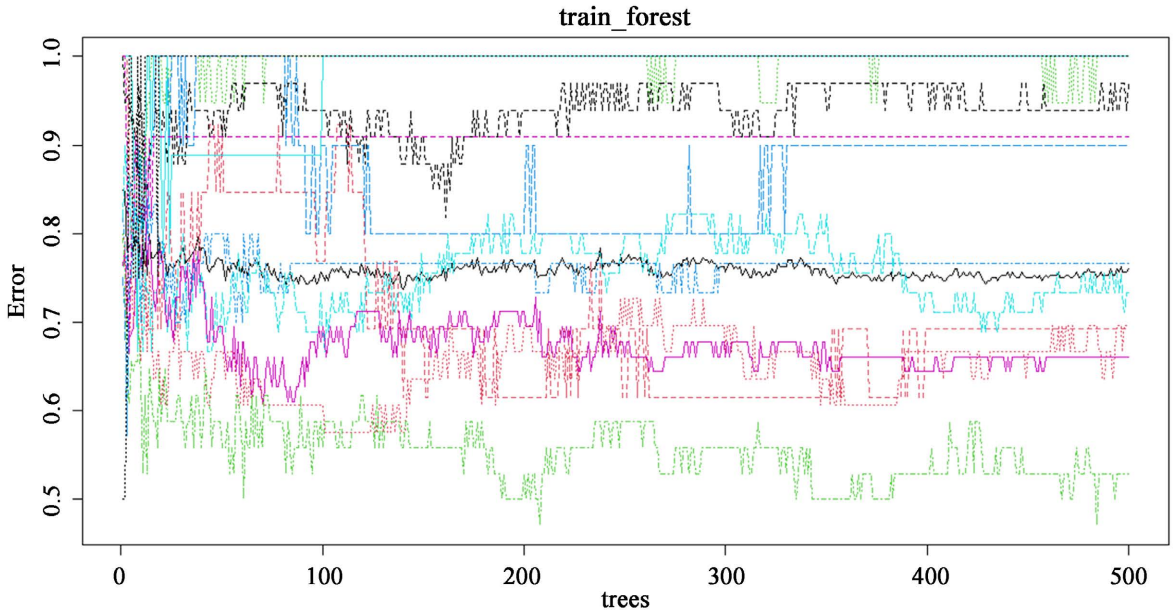

Figure 3. Error diagram of random forest model with different number of decision trees
图 3. 不同决策树数目的随机森林模型误差图

计算得到最低误差对应的决策树数目为 140。接下来，决策树数目固定为 140，对每棵决策树的特征数进行考察。得到的不同特征树对应的均方误差如下表 4：

Table 4. Error diagram of random forest model with different feature numbers
表 4. 不同特征数的随机森林模型误差图

特征数	1	2	3	4	5	6
均方误差	0.8341	0.8375	0.8230	0.8313	0.8286	0.8148
特征数	7	8	9	10	11	12
均方误差	0.8193	0.8182	0.8235	0.8138	0.8090	0.8231
特征数	13	14	15	16	17	
均方误差	0.8265	0.8203	0.8114	0.8064	0.8266	

可以看出，不同特征数的均方误差整体差异不大，当特征数为 16 时，随机森林的均方误差最小。基于此，建立一个包含 16 个特征共计 140 棵决策树的随机森林模型。

5.3.2. 特征重要性的提取

特征重要性的计算是通过将相应变量替换成一列随机的数后，计算模型准确率或 GINI 指数的降低，降低的越多，表明该变量越重要，如图 4 中所示。

图 4 中横坐标表示降低的数值，纵坐标是相应的变量。可以看到，以准确率作为标准，PC1、PR1、PR2 的重要行排在前三名；以 GINI 指数作为标准，PU3、XG2、PU2 的重要性排在前三名。

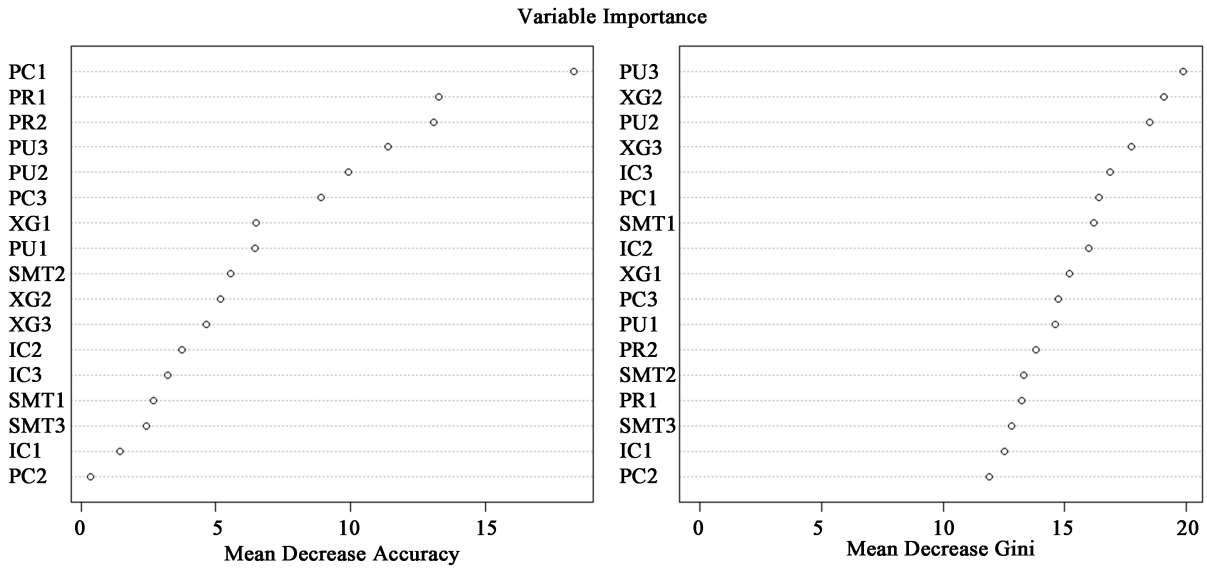

Figure 4. Rank of importance of independent variables in the random forest model
图 4. 随机森林模型中自变量的重要性排序

6. 结论

本文在隐私计算理论、沟通隐私管理理论和随机森林模型理论的基础上，从情感和行为两个角度出发，研究用户隐私披露意愿的影响因素。情感因素包括：感知有用性、感知风险性和社交媒体信任性，行为因素包括：习惯性、隐私控制性和隐私关注性。得到以下结论：

1) 从情感角度出发，感知有用性和社交媒体信任性对隐私披露意愿有正向影响。用户在社交媒体分享个人信息以满足自身需求，他们的所得越多，隐私披露的意愿就越强烈。而社交媒体信任对隐私披露意愿的正向影响也与模型假设相同，用户对社交媒体的隐私协议越信任，就越愿意披露自己的隐私信息。感知风险性对隐私披露意愿有负向影响。社交媒体的使用涉及自身的敏感信息，比如用户有时只有提供以往病史、身体缺陷等信息，才能在医疗服务平台上获得准确的诊断，而信息越敏感，用户感知到的风险就越大，就越不愿意进行隐私披露。

从行为角度出发，习惯性对隐私披露意愿具有正向影响。如果用户有发布个人信息的习惯，那么就容易忽视这一行为的风险，从而越愿意披露隐私。隐私关注性对隐私披露意愿具有负向影响。用户使用社交媒体披露大量隐私信息，那么这一行为所带来的隐私担忧会大大降低用户的隐私披露意愿。隐私控制性对隐私披露意愿有负向影响，这与研究假设和已有研究结果不符。究其原因可能是问卷的问题设置不够合理或者问卷的数据还不够准确。

2) 随机森林模型的结果解释了各因素对隐私披露意愿的影响程度大小。以准确率为标准，感知风险性和隐私关注性对用户隐私披露意愿的影响是最重要的，说明用户对于个人信息披露这一行为所带来的风险考量和隐私忧患大大影响用户的隐私披露意愿。以 GINI 指数作为标准，习惯性和感知有用性是影响用户隐私披露意愿的重要因素，用户对披露个人信息行为所获得的益处衡量和用户是否习惯于在社交媒体发布自己的个人信息都会在很大程度上影响用户的隐私披露意愿。

参考文献

[1] 李琪, 王璐瑶, 乔志林. 隐私计算与社会资本对移动社交用户自我披露意愿的影响研究——基于微信与微博的比较分析[J]. 情报杂志, 2018, 37(5): 169-175.

[2] 柳薇, 吴丁娟. 在线医疗平台用户隐私披露行为影响因素研究[J]. 医学信息学杂志, 2021, 42(6): 16-23.

[3] 李玮祎, 徐中阳, 孟知谦. 在线健康社区用户隐私披露行为影响因素研究[J]. 医学信息, 2022, 35(9): 5-9.

[4] 韩普, 黄燕杰. 在线健康社区中用户隐私悖论行为影响因素研究[J]. 南京邮电大学学报(社会科学版), 2022, 24(2): 42-55.

[5] 袁向玲, 牛静. 社交媒体隐私政策与用户自我表露的实证研究: 一个被调节的中介模型[J]. 信息资源管理学报, 2021, 11(1): 49-58.

[6] Petronio, S. (2002) Boundaries of Privacy. State University of New York Press, Albany, 268.

[7] 刘百灵, 孙文静. 沟通隐私管理理论整体视角下移动用户信息披露决策的过程研究[J]. 管理科学, 2021, 34(6): 76-87.

[8] 吴茜, 姚乐野. 互联网用户隐私披露行为影响因素研究[J]. 现代情报, 2022, 42(6): 121-131.

[9] 程慧平, 闻心玥, 苏超. 社交媒体用户隐私披露意愿影响因素模型及实证研究[J]. 图书情报工作, 2020, 64(16): 92-104. https://doi.org/10.13266/j.issn.0252-3116.2020.16.010

[10] 郭海玲, 马红雨, 许泽辉. 社会化媒体用户信息披露意愿影响模型构建与实证——以微信用户为例[J]. 图书情报工作, 2019, 63(15): 111-120.

[11] Mouakket, S. and Sun, Y. (2019) Examining Factors that Influence Information Disclosure on Social Network Sites from the Perspective of Network Externalities. *Industrial Management & Data Systems*, **119**, 774-791.

[12] 朱侯, 刘嘉颖. 共享时代用户在线披露个人信息的隐私计算模式研究[J]. 图书与情报, 2019(2): 76-82.

[13] 北京市消协发布手机 APP 个人信息安全调查报告[J]. 广西质量监督导报, 2018(4): 6. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2018&filename=GXZL201804007&uniplatform=NZKPT&v=UiHHhgB_4ELQ_uSmRhV8hiBEKMrEHFKTQL-F5__jM0HbwHSH2rDrQh9FTFDJDGmM