移动互联网环境中基于用户行为的隐私保护机制研究

姓　　名：陈福臻

学　　号：1333880

所在院系：电子与信息工程学院

职业类型：工程

学科专业：计算机应用技术

指导教师：程久军 教授

副指导教师：

二〇一六年三月

同济大学
TONGJI UNIVERSITY

A dissertation submitted to

Tongji University in conformity with the requirements for

the degree of Master of Engineering

Research on privacy protection mechanism based on user behavior in mobile Internet

Candidate: Fuzhen Chen

Student Number:1333880

School/Department: College of Electronics and Information Engineering

Discipline: Computer Science and Technology

Major: Computer Technology

Supervisor: Professor Jiujun Cheng

March, 2016

移动互联网中基于用户行为的隐私保护机制研究

陈福臻　同济大学

学位论文版权使用授权书

本人完全了解同济大学关于收集、保存、使用学位论文的规定，同意如下各项内容： 按照学校要求提交学位论文的印刷本和电子版本；学校有权保存学位论文的印刷本和电子版，并采用影印、缩印、扫描、数字化或其它手段保存论文；学校有权提供目录检索以及提供本学位论文全文或者部分的阅览服务；学校有权按有关规定向国家有关部门或者机构送交论文的复印件和电子版；在不以赢利为目的的前提下，学校可以适当复制论文的部分或全部内容用于学术活动。

学位论文作者签名：

年　　月　　日

同济大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。 对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

摘要

移动互联网的快速发展在给人们带来极大便利的同时也让用户大量的个人隐私数据暴露于互联网中，隐私保护问题日益严重。此外，移动互联网由于移动性和高精度定位等特点，也会导致数据多维和异构等问题。现有移动互联网环境下隐私保护机制未考虑高维数据的特点以及数据失真问题。本文针对以上问题，结合用户行为特点，研究移动互联网环境下用户行为访问控制模型及其动态 *k*-匿名保护方法，并将其应用于社区物联网服务平台，从而为用户提供更好的隐私保护，减少数据失真，提高移动互联网的服务质量。具体研究内容包括：

(1) 研究隐私保护方法的概念以及关键技术，深入剖析移动互联网环境下隐私保护面临的问题，阐述了移动互联网环境下用户行为信息结构，给出了基于新浪微博开放平台 API 和网页爬取相结合的分布式用户行为信息采集方案、采集流程和采集结果，并对数据进行分析和预处理，从而为移动互联网环境下基于用户行为的隐私保护机制的研究提供数据保障。

(2) 研究移动互联网环境下基于用户行为的访问控制模型的相关定义，提出了移动互联网环境下基于用户行为的访问控制模型，同时给出了模型的构建、训练和更新方法，并通过仿真实验验证了模型的有效性以及准确性。

(3) 通过基于用户行为的访问控制模型获取用户隐私需求，给出了移动互联网环境下的动态 *k*-匿名方法，并通过仿真实验对该方法的正确性进行了验证。

(4) 依托于国家科技支撑计划项目，设计并实现应用于社区物联网创新服务平台的隐私保护系统，从而进一步验证了基于用户行为的隐私保护机制的正确性和合理性。

关键词：移动互联网，隐私保护，访问控制，*k*-匿名，用户行为

ABSTRACT

The rapid development of mobile internet bring convenience for people's living, but at the same time also let a large of user’s individual privacy data exposure to the Internet, privacy issues are becoming more and more serious. In addition, mobile Internet has the characteristics of mobility, high precision positioning, this characteristics make the problem of data heterogeneous and data multidimensional. current privacy protection method can’t handle high dimensional data very well, so it unable to provide very good privacy protection for user’s privacy data in the mobile internet, at the same time the current privacy protection method have a more serious problem of data distortion. This paper aims at the above problem, combine the characteristics of user behavior in the mobile internet, take research on privacy protection mechanism based on user behavior in mobile Internet. Including user behavior based on access control model and the dynamic *k*-anonymous based on this access control model. And applies the privacy protection mechanism to the community service platform, make platform provide better privacy protection for user, improve the quality of service of mobile internet. The main content of this paper as follows:

(1) Research of the concepts and key techniques of privacy protection, take a depth analysis of the problem of privacy protection in the mobile internet environment. Give the structure of the user behavior information in the mobile internet. Design a distributed user behavior data collection scheme based on Sina Weibo open platform and Web Crawl. And process and analyzed the data after the data were collected, as to provide data support for the research of privacy protection mechanism based on user behavior in mobile Internet environment.

(2) Study of the related definitions of access control model based on user behavior in mobile internet environment. Propose the access control model based on user behavior in mobile internet environment. At the same time, give the construction methods, training method and updating method of the model. Verified the validity and accuracy of the model through experiment.

(3) Acquired the user privacy requirements through access control model based on user behavior, and then give the dynamic *k*-anonymous in the mobile internet environment. Verified the correctness of the method by simulation experiment.

(4) Relying on the national science and technology support project, design a privacy protection system and applied to community innovation service platform. To further verify the correctness and rationality of the privacy protection mechanisms.

Key Words: mobile Internet, privacy protection, access control, *k*-anonymous, user behavior

目录

第1章 引言

1.1 课题的研究背景与意义

随着数字通信与信息技术的快速发展、3G 与 4G 网络的迅速普及以及各种智能终端（手机、平板电脑、智能电视、可穿戴设备）的出现，移动互联网呈现出井喷式的发展，极大的方便了人们的生活，带来了更为丰富多彩的生活方式。移动互联网既具有互联网的开放的特征，同时又具有移动网络实时、便捷、位置交互等特点，这些特点使得用户规模不断扩大，用户间的信息交流更加频繁，也带来了更多的信息共享，形成了大量的在线数据。网络中的不法分子对用户隐私数据的非法收集利用也十分的普遍，从而给人们的生活带来困恼，隐私保护问题已经成为移动互联网蓬勃所发展带来的最大困扰之一。另外，当前的隐私保护方法在提供隐私保护的同时总会带来严重的数据失真，使得整体服务质量较差。因此，在移动互联网环境下，如何在提供高质量服务的同时保护用户的隐私信息已成为目前的研究热点之一。

隐私保护[1]就是使个人或集体等实体不愿意被外人知道的信息得到应有的保护，比如用户出生日期、家庭住址、身份证号、薪酬、疾病等。数据库拥有者如果将用户的隐私信息直接发布出去，那将有可能对用户的生命财产安全造成重大的损失。因此，数据库拥有者在数据发布之前就进行一定的隐私保护处理，将有助于对用户的隐私信息提供保护。

当前隐私保护的主要研究点在于如何设计隐私保护规则使得在用户隐私不泄露的情况下更好的对数据进行利用、为用户提供更好的服务。目前的隐私保护方法主要有数据扰动、数据加密、限制发布和访问控制。其中限制发布由于不会对原始的数据进行更改而获得了较多的关注，*k*-匿名是限制发布的一种主要方法，该方法的核心思想是通过匿名化处理，使得发布数据集中每组的 *k* 条记录相互之间无法区分。但由于移动互联网下的数据带有更多位置信息，且位置信息可能时刻产生变化，使得攻击者可以通过多次查询数据而获取用户的活动轨迹，从而确定用户的其他敏感信息。另外，单一的 *k*-匿名隐私保护方法会使得合法用户与非法用户获取到的均是匿名化处理后的数据，这样的数据带有较大的数据失真。另外一种主要的隐私保护解决方案就是访问控制，特别是基于用户行为的访问控制。通过区分不同的行为与用户，给予不同的用户不同匿名程度的数据，从而为合法

用户的合法行为提供更高质量的服务。但是，当前的基于用户行为的访问控制是针对传统互联网进行设计的，并没有考虑到移动互联网下的用户行为，对移动互联网下的用户行为特性，特别是高精度位置特性考虑不够周全。同时访问控制模型缺乏自我学习的能力，无法根据外界环境的变化不断完善自身。总体上，当前的隐私保护方法所存在的不足具体如图 1.1 所示。

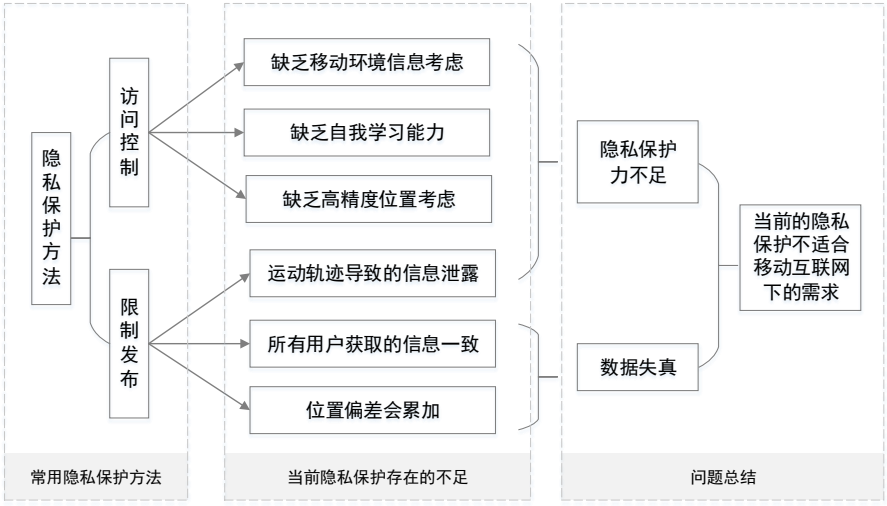


图 1.1 当前隐私保护方法存在的不足

综上所述，传统的隐私保护机制并不能很好的保护移动互联网下的用户隐私信息，如何结合移动互联网下的用户行为特性，设计一套适合于移动互联网的隐私保护机制，在最大限度地保护用户隐私的同时减小数据失真是一个值得研究的问题。

1.2 国内外研究现状

近年来，随着移动互联网的快速发展，用户越来越多的参与到移动互联网中，也越来越多的将自己的隐私信息呈现于互联网中，所面临的安全风险也随之大大增加，用户也越来越多的开始关注自己的隐私信息。基于这样的背景，对于信息的保护，特别是移动互联网下的隐私信息保护正成为近几年的研究热点。

1.2.1 用户隐私保护方法

在传统的网络中，针对用户的隐私泄露问题，已经有了相对成熟的隐私保护方法，概括的说，主要有基于数据扰动的隐私保护方法（又称基于数据失真的隐私保护方法）、基于数据加密的隐私保护方法、基于限制发布的隐私保护方法以

及基于访问控制的隐私保护方法。

1. 基于数据扰动的隐私保护方法

数据扰动（data perturbation）[2]就是发布失真的数据使得数据获得者无法根据该数据完全还原出原始的数据，但这个失真的数据集又保留有数据获得者希望保留的特性。数据扰动虽然可以防止隐私数据被完全泄露，却依然会引起部分泄露，更重要的是数据扰动会造成原始数据的失真，因此该技术近几年来发展较为缓慢，当前的数据扰动方法主要可以分成两类。

(1) 基于添加噪声的数据扰动方法[3-4]，该方法通过向原始数据添加随机噪声，然后基于贝叶斯、期望最大化算法（Expectation Maximization）等方法重构模型，消除噪音影响。该方法会对原始数据进行改变，造成数据的失真。

(2) 基于数据变换的数据扰动方法[5-6]，主要采用的方法有特征值分解和离散余弦变换矩阵，使得数据获取者在新构建的数据模型上进行数据的获取与研究。其核心思想是找出原始数据中支持敏感规则的敏感事务，通过数据项的删除或增加，从而将敏感规则的置信度降低到指定的阈值以下[7]。

2. 基于数据加密的隐私保护方法

基于数据加密的隐私保护方法主要是通过加密技术隐藏敏感数据，主要应用于分布式环境中，例如分布式安全查询、几何计算等。

(1) 安全多方计算技术(Secure Multi-Party Computation, SMC)[8-9]。SMC 包含有多个互不信任的参与方，其目的是解决参与方之间隐私保护的协同计算问题，并确保每个参与方独立输入，计算正确，同时每个参与方都不泄露任何输入值给其他参与方。虽然 SMC 的安全性很高，但是关于 SMC 的计算复杂度却非常的高，实际应用较为困难，当前关于 SMC 的主要研究工作集中于降低计算开销、优化分布式计算协议。

(2) 分布式匿名化。分布式匿名化就是各站点先对原始数据进行处理后，然后在此基础上进行数据的发布操作。但实现较为复杂，计算复杂度高，当前该领域的研究较少。

3. 基于限制发布的隐私保护方法

限制发布[10]是在数据发布时有选择性的发布一部分原始数据并隐藏其中的敏感数据以达到隐私保护的目的，其核心思想在于对数据进行匿名化处理。具体的，数据的匿名化处理一般采用抑制和泛化两种操作，抑制是不发布某些数据项；泛化是以一种更加概括、抽象的描述来表示数据项。关于限制发布的研究集中在匿名化原则的设计以及针对特定匿名化原则的更高效的的匿名算法。

隐私保护所对应的原始数据一般都是以数据库中表格的形式存储的。表中每

一条记录（或每一行）对应一个个人，包含多个属性值，这些属性值包括标识符、准标识符和敏感属性。数据的匿名化一般就是对准标识符进行泛化处理，用更加概括的值来代替原本的精确数值，泛化过程一般采用语义树（泛化树）进行操作，当前关于限制发布的最典型方法就是 *k*-匿名和 *l*-多样性。下图 1.2 为数据表中包含年龄 20-90 岁时的泛化树形状。

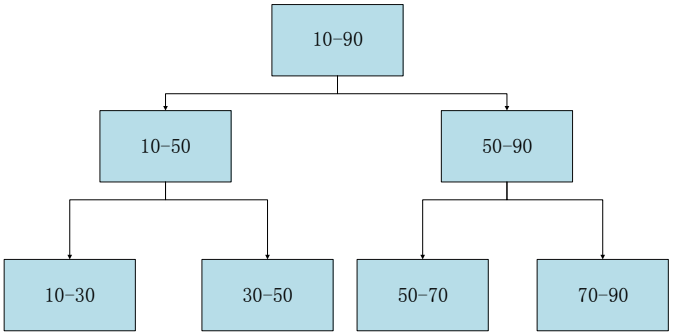


图 1.2 泛化后的树状图

4. 基于访问控制的隐私保护方法

访问控制通过对用户提出的对系统资源的访问请求进行评估，判断是接受或者拒绝该请求，以此来保护信息资源不被非法使用和访问，并使合法的用户在合法的时间内获得有效的系统访问权限，阻止非法的用户、合法用户的非授权行为访问受保护的网络资源。

针对访问控制的研究包括部署于路由器的访问控制[11]、部署于防火墙的访问控制[11-12]以及部署于 Web 服务器的访问控制，其中部署于路由器和防火墙的访问控制由于访问策略的配置限制相对而言比较简单，研究较少，当前关于访问控制的研究基本是关于 Web 服务器的访问控制，其基本研究思路是通过构建访问控制模型，分析用户的操作，动态的判断用户访问的访问权限，然后根据访问权限给予访问的内容，限制不能访问的内容。

1.2.2 数据发布中的 *k*-匿名隐私保护

定义 1.1[11]. *k*-匿名(*k*-anonymity)：给定数据表$T(A_1, A_1, \dots, A_n)$，QI是与T相关联的准标识符，当且仅当在$T[QI]$中出现的每个值序列在$T[QI]$中至少出现 k 次，则T满足 *k*-匿名。其核心思想就是将数据表 T 分成 N 组，每组里面都包含有 k 条记录，当这 k 条记录被发布时，其中的准标志符属性被泛化，使得发布的数据中这 k 条记录的准标识符完全相同。从而保证攻击者无法从这 k 条记录中定位到具体的主体或用户，以此保护主体或用户的隐私信息。图 1.3（a）和图 1.3（b）分别是原始数据和经过 *k*-匿名化处理的数据。

表a 原始数据

序号	邮编	年龄	国籍	疾病
1	13053	28	俄罗斯	心脏病
2	13068	29	美国	心脏病
3	13068	21	日本	感冒
4	13053	23	美国	感冒
5	14853	50	印度	癌症
6	14853	55	俄罗斯	心脏病
7	14850	47	美国	感冒
8	14850	49	美国	感冒
9	13053	31	美国	癌症
10	13053	37	印度	癌症
11	13068	36	日本	癌症
12	13068	35	美国	癌症

表b 经过k-匿名处理后的数据

序号	邮编	年龄	国籍	疾病
1	130**	<30	*	心脏病
2	130**	<30	*	心脏病
3	130**	<30	*	感冒
4	130**	<30	*	感冒
5	1485*	≥40	*	癌症
6	1485*	≥40	*	心脏病
7	1485*	≥40	*	感冒
8	1485*	≥40	*	感冒
9	130**	3*	*	癌症
10	130**	3*	*	癌症
11	130**	3*	*	癌症
12	130**	3*	*	癌症

图 1.3 匿名化处理之后的数据

k -匿名的实现方式大都是通过泛化或者隐匿来实现的，其基本思想是将 *k* -匿名问题视为一种聚类问题，将数据对象分成若干类或簇，同一个簇当中的数据记录之间既有很高的相似性。当前针对泛化的方法主要有全域泛化算法和局域泛化算法。

全域泛化算法是对一个属性的所有值进行泛化，在泛化之前，对所有的准标志符（*QI*）计算泛化格（也称泛化层次树），形成一个泛化系列集。如图 1.4 所示为对一个含有邮编、出生日期、性别三个准标志符的数据表所对应的泛化层次树。当前全域泛化算法的最经典实现是 Incognito[12]。

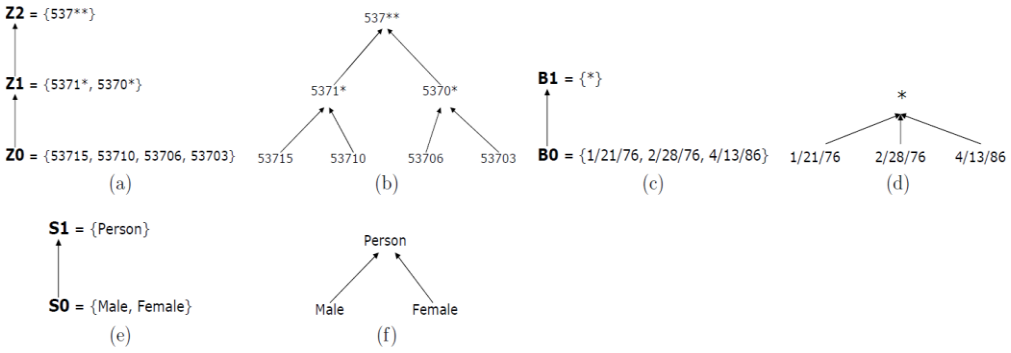


图 1.4 泛化层次树举例

局域泛化相对于全域泛化则具有更好的灵活性，局域泛化不需要实现构造泛化层次树，而是允许在泛化过程中将同一属性中不同元组所对应的属性值泛化到不同的等级中。但是，局域泛化需要有每个属性值域的顺序，同时计算起来也更加复杂。

然而如果每个准标识符对应的各条记录中的敏感属性值取值相同或某些敏

感值出现的频率很高，则仍然存在隐私泄露的风险。文献[13]对此给出了同质攻击（Homogeneity Attack）和背景知识攻击（Background Knowledge Attack）两种攻击方法，并提出了 *l*-多样性模型。

定义 1.2[13]：*l*-diversity(*l*-多样性)。如果一个准标识符组(*QI*-group)中至少包含 *l* 个具有代表性的敏感信息值，则该准标识符组满足 *l*-多样性。如果匿名数据 T^*中的所有准标识符组均满足 *l*-多样性，则T^*满足 *l*-多样性。*l*-多样性其核心思想是要求每个准标志符组中的敏感属性个数大于等于 *l*，使得确定敏感属性被披露的风险小于 *1/l*。

定义 1.3[14-16]：支持数据重发布的匿名方案。*k*-匿名和 *l*-多样性都是针对静态数据的，但通常我们的数据表是动态变化的，会有数据的删除与插入操作。对于 *k*-匿名与 *l*-多样性，攻击者通过获取多次的发布的数据，结合背景知识攻击，有很大的可能确定用户的身份。因此出现了支持数据重发布的匿名方案，该方案通过 m-invariance 算法实现对数据重发布的隐私保护。

定义 1.4[17]：时空 *k*-匿名（Spatio-temporal Cloaking）：主要是针对用户的位置信息进行匿名处理，具体的将用户的位置及时间数据进行模糊化，使得最后在该时空环境中至少包含 *k* 个用户的位置，从而使攻击者将该消息关联到某特定用户的概率小于 1/*k*。其中 *k* 称为匿名度。

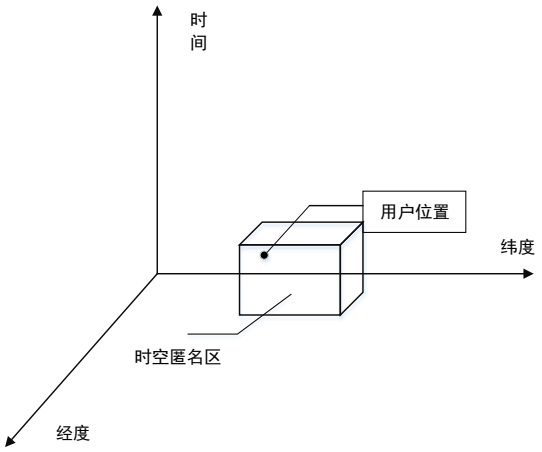


图 1.5 时空 *k*-匿名

图 1.5 所示为时空 *k*-匿名的效果图，对于用户的位置信息来说可以通过将其由点形式的值转换为一个矩形的区域来代替从而达到模糊的效果，而对于时间信息则可以通过将具体的时间推迟或者提前，从而达到泛化的效果。

但是，当前基于 *k*-匿名的隐私保护方法均是预先设定好泛化系数 *k*，导致所有用户、所有行为下看到的结果都是一致的，即使最近的一些研究中出现了个性化 *k*-匿名模型[20-21]，但目前也仅仅是针对不同的敏感属性提供不同的泛化系数，

并不能针对不同的用户、不同的行为提供个性化的数据，这就导致合法用户与非法用户获取到的数据是相同的，既造成了数据的失真，也导致隐私保护力度不足。同时，当前移动互联网环境下针对位置的 *k*-匿名方案存在着位置偏差叠加的问题，每个位置的匿名处理都会造成一定的位置偏差，当一次请求涉及多个位置时，这多个位置的偏差存在着累加的问题，从而造成超出预期的位置偏差，进一步导致了数据的失真。

1.2.3 基于用户行为的访问控制模型

1.2.3.1 移动互联网用户行为分析

随着用户规模的不断扩大以及网络信息的频繁更新，互联网数据呈现井喷似的增长，如何在繁多的数据中找出有效的信息显得尤为重要。在服务层面上，通过对用户行为的分析，可以更好的发现用户可能感兴趣的东西，从而为用户提供更好的服务。从安全层面上，通过对用户行为的分析，可以判断用户当前的行为的合法性，从而有效的抵御不法用户的非法访问行为。

用户行为分析的宗旨是通过对用户的信息行为的分析找出规律，并将这些规律与服务策略、营销策略、安全策略等相结合，从而进一步提高服务质量、安全机制[22]。其中，用户行为可以用某些特征量的统计特征或特征量的关联关系定量或定性的表示[23-24]。

用户行为分析首先需要获取大量的用户行为数据，其中的一种方式就是从日志文件中获取用户行为数据。文献[25]提出了一种基于大规模日志分析的搜索引擎用户行为分析方法，文中通过对搜狗搜索引擎在一个月内的近 5000 万条查询日志进行了分析，得出了中文用户的搜索行为习惯，从而为搜索引擎的搜索算法提供参考。

目前，关于用户行为的分析主要应用于搜索以及推荐系统，文献[26]就提出了一种基于用户行为的用户兴趣度模型，从而向用户进行个性化的推荐。用户行为分析的另一个主要应用就是网络安全，通过对用户的网络行为进行分析、建模，判断用户每一次访问的危险程度，并给出相应的防御策略。文献[27]提出了一种可信网络中用户可信行为的评估、预测与控制结构，包括了对于行为信任的可靠评估以及满足不同安全的可信预测；基于博弈理论的系统访问博弈决策；基于信任的动态的资源访问控制和异常行为的监控与防范。另外，当前的入侵检测系统也是基于用户行为分析的。文献[28]基于模式挖掘的用户行为异常检测系统就是利用数据挖掘中的关联分析和系列挖掘技术对用户行为进行模式挖掘，从而为入

侵检测系统的设计提供一种方案。

1.2.3.2 访问控制模型综述

访问控制是对信息资源进行保护的重要措施之一。R Sandhu 等人在 1994 年提出了访问控制的基本原则[29]，1996 年，R Sandhu 等人又提出了著名的 RBAC96 原则[30]，将传统的 RBAC 模型根据不同需要拆分成 4 种嵌套的模型并给出形式化定义[31]，1997 年他们更近一步，提出了一种 RBAC 管理模型 ARBAC97，从而在 RBAC 模型基础上实现了角色管理[32]。

为了进一步保证不同安全级别的信息只能被指定级别的用户访问，出现了多级访问模型，通过强制的多级访问控制策略，在一定程度上保护了信息的机密性和完整性；另外，针对分布式网络环境，有面向分布式的访问控制模型[33-34]；针对云计算平台，有针对云计算的访问控制模型[35-36]；针对资源中的时间和空间等上下文信息，出现了一系列与时空相关的访问控制模型[37-38]。

1. 基于对象的访问控制模型

在基于对象的访问控制（OBAC Model：Object-based Access Control Model）[39]中，通过将访问控制列表与受控对象相关联，并将访问控制选项设计成为用户、组或角色及其对应权限的集合，从而可以对受控对象以及受控对象的属性进行访问控制，而且派生对象可以继承父对象的访问控制设置，从而有效减轻由于信息资源的派生、演化和重组等操作所带来的工作量。

2. 基于任务的访问控制模型

基于任务的访问控制模型（TBAC Model，Task-based Access Control Model）[40]是以面向任务的观点，从任务的角度来建立安全模型，并在任务处理的过程中提供动态实时的安全管理的主动安全模型。TBAC 一般适用于工作流、分布式处理、多点访问控制的信息处理以及事务管理系统中的决策制定等[46]。

3. 基于角色的访问控制模型

在基于角色的访问控制模型（RBAC Model，Role-based Access Model）[31,41]中，角色（Role）是一定数量的权限的集合。角色一般作为用户与权限之间的代理层，所有的授权操作都是给予角色而不直接给用户。有效的克服了传统访问控制模型灵活性不足的问题，相对而言是一种更加有效的安全策略。

4. 基于属性的访问控制模型

在基于属性的访问控制模型（ABAC Model，Attributed based Access Control Model）[42]中，其基本元素包括请求者、被访问资源、访问方法和条件。但是将这些元素统一用属性来表示，摆脱了基于角色的限制。其框架示意图如图 1.6 所

示[47]。

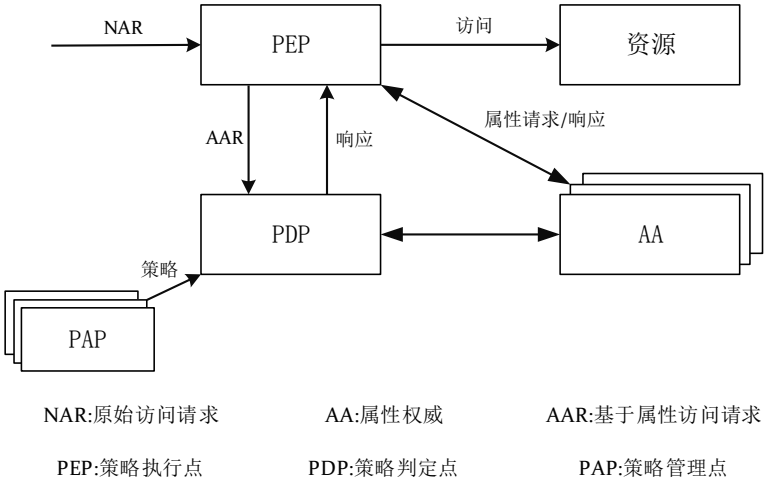


图 1.6 ABAC 框架示意图

5. 基于行为的访问控制模型

基于行为的访问控制模型（ABAC Model，Action-based Access Control Model）[43]会综合角色、时态状态和环境状态的相关安全信息，对特定角色在某种环境下某段时间内实现的某个功能所需的权限进行定义，动态的为用户当前的行为配置不同的访问权限。基于行为的访问控制模型如图 1.7 所示[43]。

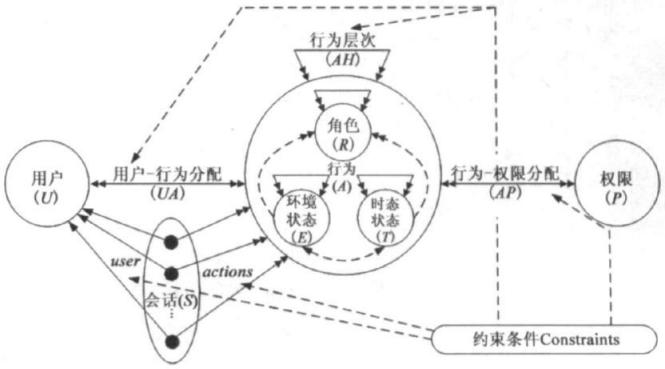


图 1.7 基于行为的访问控制模型

6. 传统的访问控制模型

传统的访问控制模型主要有自主访问控制（DAC）[44]、强制访问控制（MAC）[45]。其中自主访问控制（Discretionary Access Control, DAC）是一种接入控制服务，通过用户主体对系统资源的接入授权，包括在文件和文件夹中设置许可。用户主体可以对自己所创建的文件进行访问，并可将其授予其他用户或收回授予给其他用户的访问权限。而强制访问控制（Mandatory Access Control, MAC）是系统强制主体或者用户服从某一种访问控制策略。是由系统对用户所创建的对象，

按照规定的规则强制用户必须遵从的权限。

在当前基于访问控制的隐私保护方法的研究中，访问控制策略，访问控制机制和访问控制模型是研究和建立访问控制技术的几个主要的研究方向。针对访问控制模型，当前主要的研究点包括基于角色的访问控制模型、面向分布式的访问控制模型、基于信任的访问控制模型[48-49]和基于行为的访问控制模型。

在实际的研究与应用中，为了满足信息管理多模式的需要，结合现有的访问控制模型与多级安全模型已经成为访问控制模型的发展趋势之一。

在访问控制模型中，模型是一种利用数字、公式、结构化存储等来表示授权策略的虚拟动态模型。一般地，基于访问控制模型的隐私保护方法包括模型建模、行为匹配、系统响应等几个阶段。而访问控制模型是隐私保护方法的基础和核心，为了更好的提供隐私保护，需要对用户的访问行为进行分析，获取用户的行为特征信息，然后建立合适的基于用户行为的访问控制模型，通过模型准确的表示用户的行为安全等级，从而控制用户的访问权限。对于访问控制模型，一般需要包括模型构建、模型训练、模型更新等过程，如图 1.8 所示。

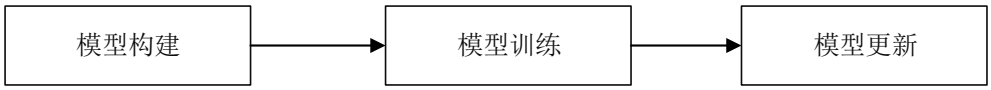


图 1.8 模型的构建过程

模型的构建一般是指通过数学分析方法，从定量的角度分析和研究一个实际问题，通过对对象的调查研究、分析、规律的总结，用数学的符号和语言来对模型进行表示。模型训练则是在模型构建的基础上，将模型应用于实际，首先通过已有的数据对构建的模型进行验证与实验，然后通过机器学习的相关方法对模型进行反复的反馈学习，直至达到某种程度的收敛。模型更新则是在模型的实际应用中，通过应用的效果不断的进行自我的调整，使得模型能够更好的迎接外界的环境变化。

当前的访问控制模型大部分是基于传统互联网思维研究的，在面对移动互联网下实时性、位置性的、多变性的外界环境时隐私保护效果相对较差。即使是文献[43]中提及的适应于移动计算的基于行为的访问控制，也并没有很好的考虑移动互联网的高精度位置性和信息的实时性、多变性。当前的访问控制模型并不能很好的为移动互联网环境下的隐私数据提供隐私保护。

1.3 论文主要工作

本文针对现有隐私保护方法在移动互联网环境下隐私保护力不足以及数据

失真的问题，给出了移动互联网环境下基于用户行为的隐私保护机制，包括基于用户行为的访问控制模型，以及基于该模型的动态 k -匿名方法。通过对用户身份与用户行为的判别，为不同的用户提供不同的访问服务，合法用户的合法行为可以获得高保真的数据，而非法用户或者非法行为则只能获取失真的数据甚至被拒绝服务，从而在保护隐私数据的同时提高服务质量。主要工作包括以下几点。

(1) 对隐私保护方法的概念、原理以及分类进行了梳理与概括，分析了各类隐私保护技术的研究现状以及关键技术点，总结其各自优缺点，深入剖析传统隐私保护方法在移动互联网环境下面临的问题，明确了研究适合于移动互联网环境的隐私保护方法的重要性和必要性。

(2) 深入分析移动互联网环境下的用户行为特点，根据移动互联网下的用户行为维度高、更新频繁、数据量大、精确位置的特点，阐述了移动互联网环境下的用户行为信息结构，给出了基于新浪微博开放平台 API 和网页爬取相结合的分布式用户行为信息采集方案、采集流程和采集结果，并对采集到的数据进行分析和预处理，从而为移动互联网环境下基于用户行为的隐私保护机制的研究提供数据保障。

(3) 研究移动互联网环境下基于用户行为的访问控制模型的相关定义，提出了移动互联网环境下基于用户行为的访问控制模型，给出基于相似性计算和遗传算法的模型训练方法以及基于记忆原理和遗传算法的模型更新方法，并通过仿真实验验证了模型的有效性以及准确性。

(4) 通过基于用户行为的访问控制模型获得访问安全等级，利用映射函数获得用户隐私需求，给出移动互联网环境下的动态 k -匿名方法，包括动态 k-匿名方法和基于四叉树的动态位置 k-匿名方法，并通过仿真实验对该方法的正确性进行了验证。

(5) 依托于国家科技支撑计划项目，设计并实现应用于社区物联网创新服务平台的隐私保护系统，从而进一步验证了基于用户行为的隐私保护机制的正确性和合理性。

1.4 论文组织结构

本文一共分为六章。

第一章，讲述了本文的课题研究背景和意义，以及课题在国内外研究现状，总结各个技术的优缺点，并说明主要研究内容以及结构安排。

第二章，阐述了移动互联网环境下的用户行为信息结构，设计了一套基于新

浪微博开放平台 API 和网页解析相结合的分布式数据采集方案，同时对采集到的数据进行了分析、预处理、特征提取等，为后续基于用户行为的访问控制模型的进一步研究奠定了基础。

第三章，研究移动互联网环境下基于用户行为的访问控制模型的相关定义，提出了移动互联网环境下基于用户行为的访问控制模型，并根据爬取的数据进行模型的构建、训练以及更新，最后通过仿真实验验证了模型的有效性和构建算法的准确性。

第四章，通过基于用户行为的访问控制模型获取用户隐私需求，给出了移动互联网环境下基于 *k*-匿名的动态隐私保护方法，针对不同的隐私需求提供个性化的 *k*-匿名方案。并通过实验验证匿名算法的准确性以及有效性。

第五章，依托于国家科技支撑计划项目，设计并实现应用于社区物联网创新服务平台的隐私保护系统，从而进一步验证了基于用户行为的隐私保护机制的正确性和合理性，展现其实际应用价值。

第六章，总结本文主要研究内容及成果，并给出了下一步的工作方向。

1.5 本章小结

本章主要介绍了隐私保护的研究背景，从理论和实际应用两个方面阐述了用户隐私保护的研究意义。详细介绍了用户隐私保护的国内外研究现状和研究方法，分析各种技术的优劣点，并重点介绍了典型的 *k* -匿名隐私保护方法和基于用户行为的访问控制隐私保护方法，分析其中的关键技术点，指出这两种隐私保护方法在移动互联网环境下所面临的问题，明确了结合移动互联网移动性、高精度定位等特征进行用户隐私保护研究的必要性和重要性。最后给出了论文的主要工作和组织结构。

第2章 移动互联网中用户行为信息采集及分析

本章首先给出了移动互联网环境下的用户行为信息结构表示；然后设计了基于新浪微博开放平台和网页爬取相结合的分布式用户行为数据爬取方案；紧接着对抓取到的数据进行预处理，去除不合理的数据；最后对预处理完成后的数据进行数据特征提取，为基于用户行为的隐私保护机制的研究提供数据保障。

2.1 移动互联网环境下的用户行为信息结构

移动网络环境中的用户行为信息主要包括移动网络环境下的用户信息和移动网络环境下的行为信息。其中，用户信息是对用户静态属性的表述，行为信息则是对用户的一系列动态操作的表述。

2.1.1 用户信息

用户的个人信息主要包括用户的一些基本信息，例如姓名、性别、出生日期、昵称、所在地、个人简介等。用户的每个属性信息都可以看成是二维平面的一个点，纵轴是不可变的属性，包括出生日期、姓名、性别等。横轴是可变的用户属性，包括所在地、年龄、昵称、关注用户数、被关注用户数、是否认证等。因为当前所在地会随着用户居住地的变化而变化，年龄会随着时间变化，关注用户数与被关注用户数则是基于用户的操作动态变化的。

用户除了个人的基本信息外，还会有用户的标签信息。用户标签是用户个人信息特征的载体，用于描述用户当前的性格特征、爱好倾向。一方面，用户可以通过标签发现与自己同行同好、同性格或共同经验的人。另一方面，服务提供商通过用户标签，可以用用户提供更符合用户兴趣的服务，也可以通过用户标签判断用户当前行为的可信度，从而对用户的访问行为进行控制。

定义 2.1：用户U，$U=\{U_1,U_2,\cdots,U_n\}$，$U_i(0\le i\le n)$表示一个具体的用户。

2.1.2 好友关系

好友关系属于复杂网络，是一张巨大的带权有向图，包含节点、关系、群组、社区等基本要素，每个用户都是图中的一个节点，如果用序号U_i来表示每一个用

户节点，则用户的好友关系可以定义为：

$$F_{ij} = (U_i, U_j, W) \tag{2.1}$$

其中U_i，U_j分别代表用户 i 和用户 j，W 表示用户U_i对于U_j的好友紧密程度，W 的定义如下：

$$w = \begin{cases} 0 & user\ i\ not\ follows\ user\ j \\ 1 & user\ i\ follows\ user\ j \\ 2 & user\ i\ follows\ user\ j\ and\ user\ j\ follow\ user\ i \end{cases} \tag{2.2}$$

对于由 n 个用户组成的好友关系，可以用图的邻接矩阵来表示，共有 n 个顶点和n^2条弧，因此，好友关系的邻接矩阵可以定义为：

$$A = \begin{pmatrix} F_{11} & \cdots & F_{n1} \\ \vdots & \ddots & \vdots \\ F_{1n} & \cdots & F_{nn} \end{pmatrix} \tag{2.3}$$

2.1.3 行为信息

行为信息是指用户的一系列操作的集合，可以是一次访问，也可以是一次简单的信息修改。在移动互联网环境中，用户的每一次操作都是一次行为，每一次行为中会包含何时（when）、何地（where）、何种操作（what）三种信息。同时可能会包含有使用终端型号、操作时长、操作频率等其他客观环境信息。

时间（when），对于同样的操作，在不同时间段所具备的属性是不一样的，白天时间可能更多的进行工作相关的操作，晚上的时候则会更多的进行娱乐相关的操作。记时间为s_t。

位置（where），位置是移动互联网的基本属性，每一个操作都会有对应的位置信息，不同的位置下对应的安全属性也是不一样的，例如，在常用的地点（办公室、教师、住宿地址）进行操作会有更高的安全属性，而在非常用地点的操作则相对有更高的风险，记位置为s_l。

操作（what），操作是一系列动作的集合，可以是一次信息访问、也可以是一次信息修改，对于微博用户，操作可以是查看微博、查看微博详情、访问他们微博主页、关注其他用户等。记操作为s_w。

其他客观环境形态，记为s_e，$s_e = (e_1, e_2, \cdots e_m)$，$s_e$是指用户行为发生时对应的外部客观环境，例如使用平台（windows, android 等），网络环境（内网、VPN，公网）等。

定义 2.2：行为ub

行为是一系列操作的集合，用于表述一次完整的操作，可以用一个四元组表示，记为$ub = (s_t, s_l, s_w, s_e)$。其基本构成如图 2.1 所示。

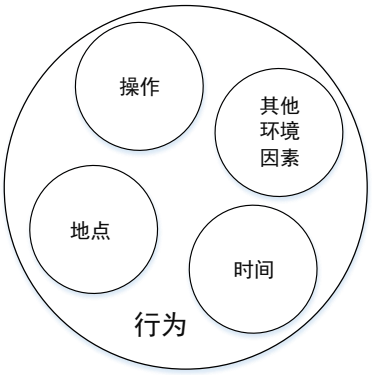


图 2.1 行为的基本构成

2.2 用户行为信息采集方案

用户的行为信息主要来自两个方面，一方面从实验室已有的社区物联网创新服务平台项目中获取，另一方面则是通过网络爬虫从网络中爬取，如图 2.2 所示，而网络爬虫主要的数据来源则是新浪微博。

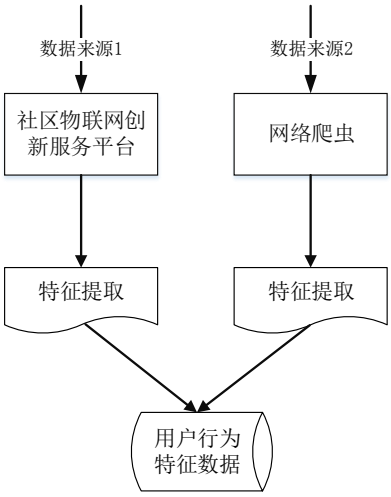


图 2.2 数据获取方式

新浪微博使用人数众多，而且随着移动化的浪潮，越来越多的用户通过微博手机客户端来发表、浏览微博，每条微博信息所含有的字段如下，其中含有所需要的来源、位置、时间、操作等信息，而且微博用户又是一个庞大的用户关系网，通过对微博用户以及微博用户行为的分析，可以有效的构建自己的用户行为模型。当前的微博开放平台所提供的用户信息所包含的字段如图 2.3 所示。

返回值字段	字段类型	字段说明
created_at	string	微博创建时间
id	int64	微博ID
mid	int64	微博MID
idstr	string	字符串型的微博ID
text	string	微博信息内容
source	string	微博来源
favorited	boolean	是否已收藏，true：是，false：否
truncated	boolean	是否被截断，true：是，false：否
in_reply_to_status_id	string	（暂未支持）回复ID
in_reply_to_user_id	string	（暂未支持）回复人UID
in_reply_to_screen_name	string	（暂未支持）回复人昵称
thumbnail_pic	string	缩略图片地址，没有时不返回此字段
bmiddle_pic	string	中等尺寸图片地址，没有时不返回此字段
original_pic	string	原始图片地址，没有时不返回此字段
geo	object	地理信息字段 详细
user	object	微博作者的用户信息字段 详细
retweeted_status	object	被转发的原微博信息字段，当该微博为转发微博时返回 详细
reposts_count	int	转发数
comments_count	int	评论数
attitudes_count	int	表态数
mlevel	int	暂未支持
visible	object	微博的可见性及指定可见分组信息。该object中type取值，0：普通微博，1：私密微博，3：指定分组微博，4：密友微博；list_id为分组的组号
pic_ids	object	微博配图ID。多图时返回多图ID，用来拼接图片url。用返回字段thumbnail_pic的地址配上该返回字段的图片ID，即可得到多个图片url。
ad	object array	微博流内的推广微博ID

图 2.3 微博信息相关字段

目前使用的微博信息获取方法主要有两种，一种是基于新浪微博 API（新浪微博开放平台）的数据获取方法，使用该方法可以快速的获取 JSON 格式的数据；另一种是使用传统的网页解析方案[50]，该方法实现难度较大，但所受的限制比较小。

2.2.1 基于微博开放平台的数据获取方法

通过新浪微博开放平台 API 可以直接获取 JSON 格式的数据，这种数据获取方式简洁明了，获取成本较低，不需要额外的数据解析即可获得关键的数据，可以获得的数据包括用户信息、微博内容、评论、用户关系等。但是开放平台出于安全或者盈利的考虑对于接口请求次数和频率有一定的限制，每个小时只能请求一定的次数，这意味着一个应用每小时调用接口的频次受到新浪平台的严格限制。当频次达到上限时，在该段时间内就无法获取新的信息。加上接口可以获得的数据内容完全取决于新浪提供的 API，缺乏自由定制的空间，可扩展性较差。

数据的抓取首先需要在新浪微博进行应用注册，获取专属的 App Key 和 App Secret，然后获取 AccessToken 进行 Oauth2.0 身份认证。最后直接通过新浪微博的 SDK 即可获取对应的数据。其抓取流程如图 2.4 所示

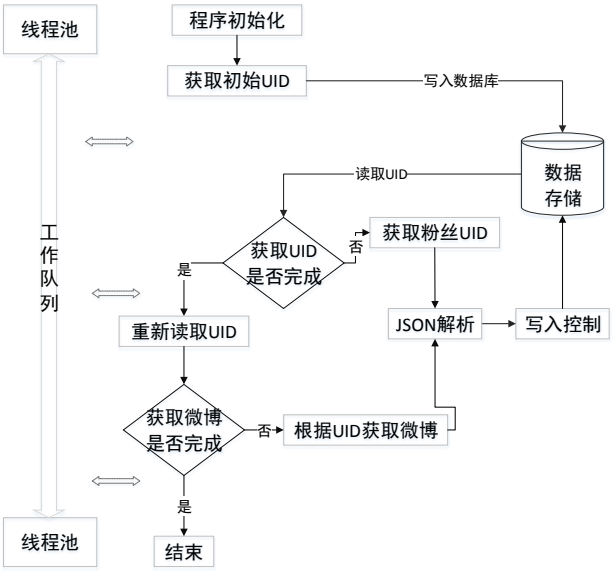


图 2.4 数据抓取程序流程

2.2.2 基于网页解析的数据获取方案

虽然基于 API 的数据获取方式简单便捷，但其诸多限制也导致了一些数据无法获取，因此还需要引入网络爬虫与网页解析技术来获取更多的信息。网页爬取其实就是模拟用户在网页的操作流程，对于新浪微博，其操作基本是建立在用户登录的基础上的，具体的，对于加 V 的用户，用户可以在不登录的情况下直接获取微博数据，对于普通用户，用户则必须登录后才可以查看其信息，并且查看其发布的微博信息。当然，有一个例外，如果访问新浪微博的移动端网页版的话，可以在不需要登陆的情况下直接获取用户的微博信息，其主要流程就是根据每个用户的 *id* 获取其对应微博的 *containerId*，然后根据这个 *containerId* 就可以获取该用户的全部微博信息，但由于新浪微博的限制，该方法获取的微博消息不包含有位置信息。因此，采用的是第一种登陆后获取微博信息的方法，其具体的流程主要包括以下部分：

(1) 用户登录，从而获取新浪存储到用户电脑上的 *cookie*。

(2) 通过几个公众账户，通过 API 获取关注该账户的用户，从而获取大量的微博用户 *Id*。

(3) 通过用户 *Id* 获取用户的微博内容信息，并对获取到的网页利用 *httpParams* 进行内容解析。

(4) 获取用户的关注者 *Id*，如果该 *Id* 不在当前的 *Id* 池中，则将该 *Id* 加入到用户的 *Id* 池中。

(5) 重复执行步骤(2)、(3)、(4)，直到符合程序终止条件时终止程序。

2.2.3 分布式用户行为信息采集方案

由于整个研究需要较多的数据，而不管基于 API 的数据获取方案还是基于网页的数据获取方案，其多少都会受到新浪的一些政策限制。同时 IO 性能、网络资源和操作系统的限制也会影响数据的爬取速度，总体而言，单点单线程采集数据获取速度较慢，缺乏自由定制的空间。

因此，设计一个分布式数据采集系统来采集用户行为数据，具体的数据包括用户基本信息、用户的好友关系、用户的行为操作信息。并在分布式系统中同时采用基于 API 的数据获取方法和基于网页解析的数据获取方法。支持大规模的并发操作，降低偶然网络错误导致的采集问题，降低新浪的政策限制对数据采集产生的影响。然后将采集下来的数据存放到特定的数据库中。

系统的总体架构如图 2.5 所示，主要包括数据存储、进行爬虫调度的控制节点，运行爬虫程序的爬虫节点。其中爬虫节点主要包括用于爬虫的机器以及运行在上面的爬虫程序，有的机器运行基于开放平台 API 的数据采集程序，有的机器运行基于网页解析的爬虫程序，具体运行哪个程序根据需要动态设定。控制节点主要是负责系统的调度，负责系统的负载均衡，同时给各个爬虫节点分配任务。数据存储就是将采集下来的数据存到到数据库中。

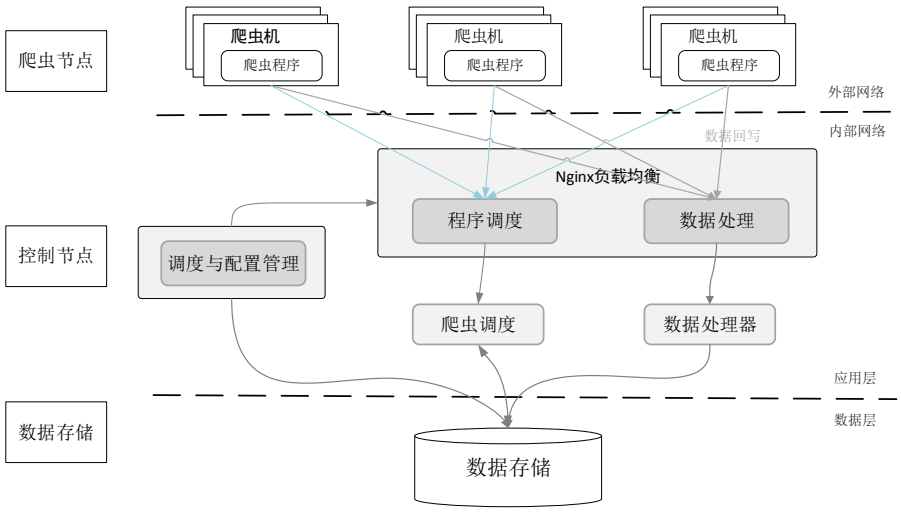


图 2.5 分布式数据采集系统

2.3 数据采集结果与数据预处理

2.3.1 采集结果

采集到的数据主要有用户信息，用户好友关系，用户的行为数据，为了方便后续的处理，采集下来的数据存储在 MySQL 数据库中。其中，用户信息主要包括用户 *UID*、用户昵称、用户所在地、用户性别、用户头像、微博数、粉丝数、关注数等，存在表 *weibouser* 中。好友关系主要是定义用户之间的关系，在存储时以队列的形式存储用户的关系矩阵，存在表 relations 中，用户行为数据主要包括用户 *UID*、时间、位置（包括经纬度）、操作类型（读、写、执行）、详细的操作日志等，存储在表 *behavior* 中。

2.3.2 数据预处理

采集到的数据中可能含有一部分的肮脏的数据，比如微博上的“僵死”用户（注册后从未使用过的账户）、不完整的行为数据记录、重复数据等。因此，为了方便后续的处理，需要对数据进行简单的预处理[51]，以得到满足条件的数据。

数据处理时，首先是进行数据的清理[52]，数据清理的基本流程如图2.6所示。其中，脏数据主要包括空值、数据不合法、数据重复、噪声数据等。清洗规则包括属性清洗、重复记录清洗、数理统计技术、数据挖掘技术等。最终得到满足质量要求的数据。

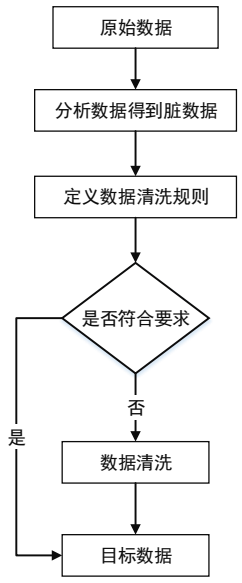


图 2.6 数据清理流程

对于用户信息，主要进行用户信息的筛选，删除掉信息不全的用户，去掉“僵死”用户。对于用户关系，主要的预处理流程如下：

(1) 遍历用户好友关系表，如果某一用户对应的 *UID* 不在用户信息表中，则删除该条关系

(2) 删除重复的好友关系。

(3) 如果两个用户互相关注，则置 *relation* 字段为 2（互相关注）。

(4) 数据聚集，对于同一用户的好友关系，在排序上连续，方便后续查询。

处理完用户的好友关系之后，还需要对用户的微博行为信息进行处理，对于用户的行为信息，其主要的预处理流程如下：

(1) 去掉不含有位置信息的记录，去掉信息不完整的记录。

(2) 根据 *weiboId* 去掉重复的记录。

(3) 对于只有经纬度而没有详细地理名称的记录，通过调用微博 API 获取详细的地理名称。

(4) 根据操作用户 *UID*，如果该用户不在用户信息表中，则更新用户信息表，并进行用户信息的预处理判断。

2.4 用户行为信息特征提取

用户行为特征提取主要就是从用户当前的操作中提取出关键的信息，具体可以提取到的信息如图 2.7 所示。对于每一次用户行为，首先可以从 *Session* 或者用户的 *Token* 中获取用户身份，从而获取用户 *UID*。对于每一次行为信息，则是从中提取时间、地点和对应的操作。对于每一个原子操作，我们再一次进行了特征提取，提取出操作类型和操作详情，操作类型主要分为读、写、执行。操作详情则是具体的操作，比如发布微博、关注他人、读取图片等。

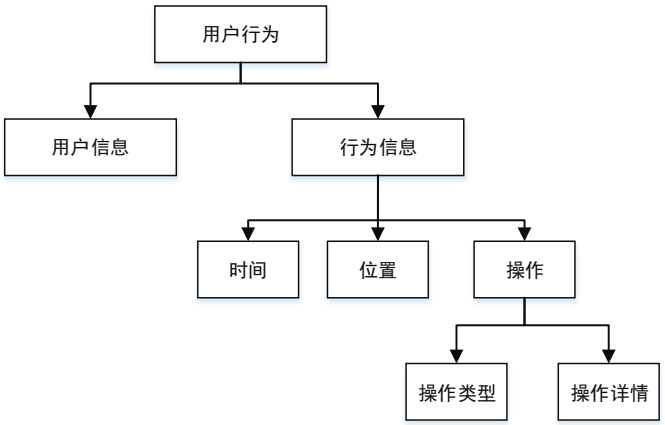


图 2.7 从用户行为中提取到的信息

2.4.1 时间特征提取

对于时间，为了方便统计，将时间分成若开时间片，然后进行聚类。根据新浪公布的新浪微博用户各个时间段的使用时间统计[53]，如图 2.8（a）为工作日的用户使用情况，图 2.8（b）为周末的用户使用情况。然后结合用户在不同时间段的地域分布，比如凌晨基本在家，白天则更多的在公司等原则，得到用户的使用时间片划分，具体的划分如表 2.1 所示。

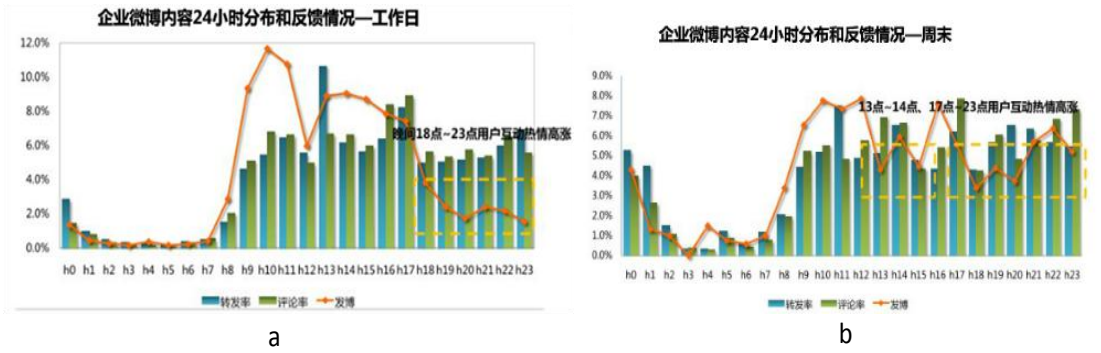


图 2.8 微博用户 24 小时的使用情况

表 2.1 用户时间段划分

时间段	描述
00:00-08:00	清晨，家庭
08:00-12:00	上午，公司
12:00-14:00	中午，公司
14:00-18:00	下午，公司
18:00-24:00	晚上，家庭

在时间段划分完成之后，由于每次行为的时间点都是分散的，分别处理的话计算维度太高，计算代价太大，所以需要对用户的时间属性进行分类，降低时间维度，将用户的行为活动时间划分到对应的时间片中。

2.4.2 位置特征提取

其次，针对用户行为的位置信息，由于位置是二维平面上的一系列点的集合，对这些点进行计算，计算复杂度和维度也都很高，需要对这些点按照位置距离进行聚类，形成一个个聚集点，例如家庭地点、公司地点等。因此，采用 k-means[54] 进行位置信息的聚类实现。

k-means 算法是很典型的基于距离的聚类算法，k-means 算法以 k 为参数，

把 n 个对象分成 k 个簇，使簇内具有较高的相似度，而簇间的相似度较低。对位置信息进行聚类时采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似性就越大，是典型的目标函数聚类方法的代表，以数据点到类别中心的某种距离和作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。

本文基于 K-means 算法的思想和原理，给出了基于 K-means 的位置处理算法。同时，为了避免 K-means 聚类算法的初始点选择不稳定，由于随机选取引起聚类结果的不稳定问题，本文在 K-means 的 k 个初始值选取前进行了一定的预处理，优先选择代表用户常用地点的 k 个值（包括家庭、公司、常去公共区域等），然后再进行 K-means 聚类得到 k 个聚簇。因此，该方法主要包括两个部分，如图 2.9 所示。

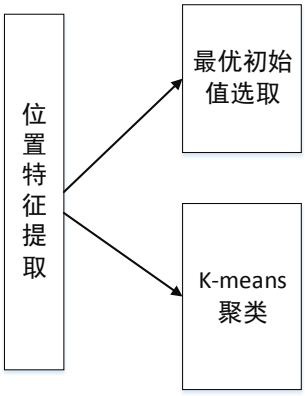


图 2.9 位置特征提取

首先，最优初始值选取采用基于最小最大原则的聚点选择方法，其基本原理是：假设要将样品分成 k 个类别，则首先使相距最远（余弦相似度最小）的两个样品 x_{i1} 和 x_{i2} 为前两个聚点，而其余聚点的选取则可以用递推公式表达。若已经选择了 m 个聚点 $m<k$，则第 m+1 个聚点选取的原则为：

$$\min\{d(x_{i_m}=1,x_{i_r}),r=1,2,\cdots m\}=\max\{\min[d(x_j,x_i),r=1,2,\cdots,m],j\neq i_1,\cdots,i_r\}$$

其次，*k-means* 算法的处理过程如下：首先，随机地选择 k 个对象，每个对象初始地代表了一个簇的平均值或中心；对剩余的每个对象，根据其与各簇中心的距离，将它赋给最近的簇；然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。具体实现如下所示：

(1) 给定大小为 n 的用户位置数据集 $(x_1,x_2,...,x_n)$，令 $I=1$ ，选取 k 个初始聚类中心 $Z_j(I)$，$j=1,2,...,k$ 。

(2) 计算每个位置数据与聚类中心的距离 $D(x_i,Z_j(I)),i=1,2,...,n,j=1,2,...,k$，如果满足 $D(x_i,Z_k(I))=min\{x_i,Z_k(I),j=1,2,...,n\}$，则 $x_i \in r_k$，r_k 表示区域 k。

(3) 使用公式（2.4）计算误差平方和准则函数 J_c，$x_k^{(j)}$ 表示属于 r_j 的位置数据，n_j 表示位置数据属于 r_j 的位置总数。

$$J_c(I)=\sum_{j=1}^{k}\sum_{k=1}^{n_j}\left\|x_k^{(j)}-Z_j(I)\right\|^2 \tag{2.4}$$

(4) 判断：若 $\left|J_c(I)-J_c(I-1)\right|<\xi$ 则算法结束；否则 $I=I+1$，计算 k 个新的聚类中心 $Z_j(I),\ j=1,2,...,k$：

$$Z_j(I)=\frac{1}{n}\sum_{i=1}^{n_j}x_i^{(j)},\ j=1,2,...,k \tag{2.5}$$

算法 2.1 基于 K-means 的位置聚类算法

输入： $(x_1,x_2,...,x_n)$，用户位置数据集.

输出： $Z_j(I),\ j=1,2,...,k$，k 个聚类中心;

$r_j,j=1,2,..,k$，k 个聚类位置集合.

1. $I = 1$
2. Initial $Z_j(I),\ j=1,2,...,k$
3. **For** $i = 1$ to n **do**
4. **For** $j = 1$ to k **do**
5. $D(x_i,Z_j(I)) = \left|x_i-Z_j(I)\right|$
6. **If** $D(x_i,Z_j(I))==min\{x_i,Z_j(I),j=1,2,...,n\}$ **do**
7. $x_i \in r_k$
8. **End If**
9. **End For**
10. **End For**
11. **If** $I=1$ **do**

12. $J_c(I) = \sum_{j=1}^{k}\sum_{k=1}^{n_j}\left\|x_k^{(j)}-Z_j(I)\right\|^2$
13. $I = I+1$
14. **End If**
15. **For** $j = 1$ to k **do**
16. $Z_j(I) = \frac{1}{n}\sum_{i=1}^{n_j} x_i^{(j)}, j=1,2,...,k$
17. $J_c(I) = \sum_{j=1}^{k}\sum_{k=1}^{n_j}\left\|x_k^{(j)}-Z_j(I)\right\|^2$
18. **End For**
19. **If** $\left|J_c(I)-J_c(I-1)\right| <\xi$ **do**
20. **Return**
21. **Else**
22. $I = I+1$
23. **Go** 15
24. **End If**

2.4.3 操作特征提取

根据图 2.7 中的阐述，操作特征的提取包括操作类型和操作详情两部分，其中，对于操作详情的特征提取主要是通过关键词提取来实现的。

1. 操作类型提取

每一次的操作请求都会发到特定的服务器接口，根据服务器接口可以提取出用户当前的操作类型。在操作系统中，对每个文件赋予了三个权限，分别是读（r）、写(w)、执行(x)，分别用数字表示如下，比如 6 代表可读可写、3 代表可写可执行。针对每一个服务器接口，根据其接口类型，获得该操作的类型，并记录到数据库中，方便后续处理使用。具体设置如表 2.2 所示。

表 2.2 权限的数值表示方式

权限	数值表示
读（r）	4
写(w)	2
执行(x)	1

2. 关键词提取

关键词的提取主要是从全部特征中选出一个特征子集，使得处理起来更加的方便。其主要的流程如图所示，对于获取到的文本，首先进行文本分析，主要就是分词处理，获得一系列词语，然后针对获取到的词语，进行特征的提取。

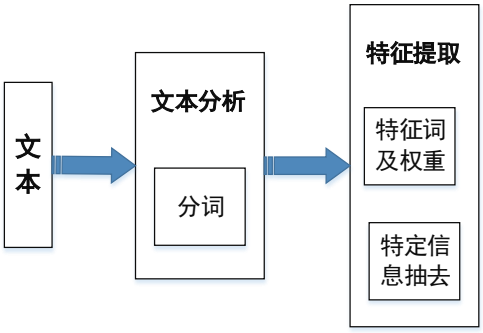


图 2.10 特征提取流程

现有的分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。本文通过字符串匹配，获取用户操作的一系列带有特征的关键词，最后，通过 *k* 近邻算法[55]（*k-Nearest Neighbour*，*KNN*）对关键词进行分类。

KNN 是一种源于思维决策的分类算法，其思路是如果一个样本在特征空间中的 *k* 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。根据判定的类别获取当前的语句特征，从而检测出用户当前的操作特征。

2.5 仿真实验与结果分析

2.5.1 实验环境与实验方法

实验采用 Java 实现，在 Intel(R) Core(TM) i5-3470 CPU @ 3.2GHz 3.20GHz、8G 运行内存的 Windows 8.1Pro 平台上运行，编译器为 Eclipse Juno(4.2)+JDK1.7。后台数据库为 MySQL5.0。实验中通过三组位置数据来验证基于 K-means 的位置处理算法的正确性与有效性。其中，这三组数据对应的二维平面上的点分别如图 2.11 所示。

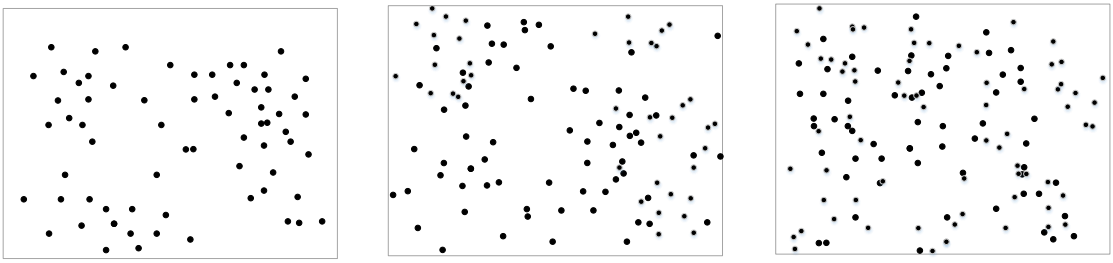


图 2.11 三组实验初始位置点

2.5.2 结果分析

在实验中将每个位置$(x_1, x_2, \ldots, x_n)$的经纬度值对应到二维平面上的点$\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}$来表示位置特征，以点之间的距离来表征位置间的距离。计算过程中选择了不同的k值进行测试，图 2.12 中 a 和 b 分别显示的是第一组实验中当$k=2$和$k=3$时的聚类结果。

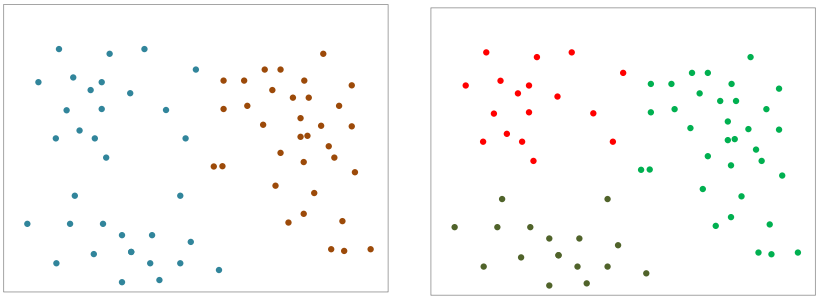


图 2.12 k 分别为 2 和 3 时的聚类效果

图 2.13 中展示的则是三组实验中 k 值和误差平方和准则函数 J_c 的关系。随着 k 的增大，J_c 的值减少，最后趋于平稳。这是因为聚类中心越多，各个位置点距离中心的平均距离越近。

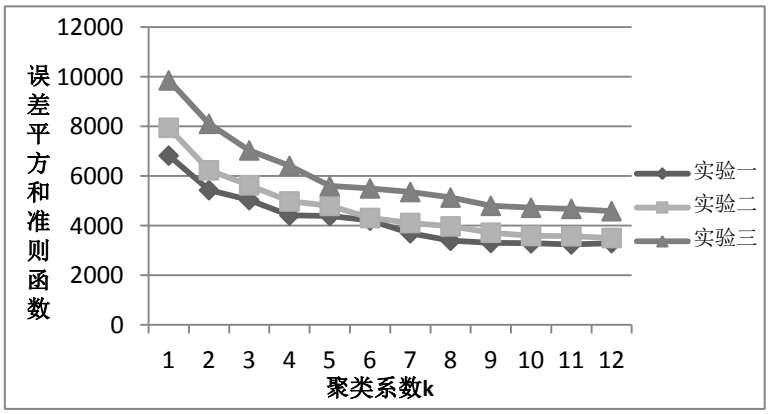


图 2.13 k 值与聚类误差的关系

2.6 本章小结

本章首先给出了移动互联网环境下用户行为信息结构表示，包括用户信息、好友关系、行为信息等；然后设计了基于新浪微博开放平台和网页爬取相结合的分布式用户数据爬取方案，用于抓取后期实验所需要的数据；紧接着对抓取的数据进行预处理，主要是进行数据清洗；再接着对预处理完成后的数据进行数据特征提取，包括时间特征提取、基于 *K-means* 的位置特征提取和基于 *KNN* 的操作特征提取，为后续基于用户行为的隐私保护机制的研究提供数据支撑。最后，通过实验验证了特征提取算法的准确性和有效性。

第3章 移动互联网中基于用户行为的访问控制模型

当前的访问控制模型并不能很好的对移动互联网环境下的用户隐私数据进行保护，本章结合移动互联网环境下的用户行为特征以及访问控制模型特点，给出移动互联网环境中基于用户行为的访问控制模型 *UA-BACM*，同时给出模型的训练方法以及模型的更新机制。为后续基于用户行为访问控制模型的 *k*-匿名隐私保护方法的研究提供基础。

3.1 基于用户行为的访问控制模型的表示

根据 2.1 小节中对于移动互联网环境下的用户行为信息结构的分析与定义，结合移动互联网环境下用户行为的特点，综合用户、好友、行为、权限的概念，给出移动互联网环境下基于用户行为的访问控制模型。

定义 3.1: 移动互联网环境下基于用户行为的访问控制模型 *UA-BACM*（User Action-Based Access Control Model），其中，$UA-BACM \preceq R, UB, P$），*UA-BACM* 是一个多元组，综合考虑用户当前行为、用户累积信誉、用户好友信誉、权限等，各个元组的具体定义如下：

$UB=(s_t，s_l，s_w，s_e)$，具体内容见定义 2.2。

$R=(U,F)$，R 代表的是当前角色，是指在特定环境下，实现某种特定操作的权限的集合的描述，用于将大量的用户归一化到特定的角色，减少运算的维度。$F_{ij}=(U_i,U_j,W)$ 表示用户的好友关系的集合，$U=\{u_1,u_2,...,u_m\}$ 表示移动社交网络用户的集合。

$P=\{p_1,p_2,...,p_m\}$ 表示一系列权限的的集合，其中 $p_i(1\le i\le m)$ 代表针对计算机系统中一个或多个对象的一种访问许可。

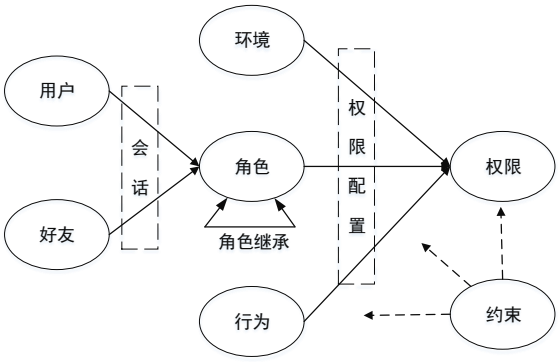


图 3.1 *UA-BACM* 基本结构

图 3.1 中展示了 *UA-BACM* 的基本结构以及各个元组（包括角色、权限、行为、环境）之间的相互关系。

定义 3.2：会话 S，会话是从用户信任等级到角色的映射，由于 $R=(U,F)$，而对于用户 u_i 其信任等级是变化的，其好友 F_{ij} 也是变化的，因此，在用户 u_i 提出访问请求时，需要通过函数 S 实现用户及其好友到角色的映射。

定义 3.3：约束，约束是指整个模型中的一系列条件限制，当用户行为产生时，用于控制不用的操作访问不同的资源，避免冲突。

定义 3.4：角色继承（role inheritance, RI），角色继承指角色在满足一定条件下可以继承令外一个角色的部分或者全部权限，它表示角色 R 与 R 之间的二元关系，$RI \subseteq R \times R$。$(r_1, r_2) \subseteq RI$ 表示角色 r_2 拥有 r_1 的部分或全部权限。其中 r_1 称为低级角色或父角色；r_2 称为高级角色或子角色。如图 3.2 所示。

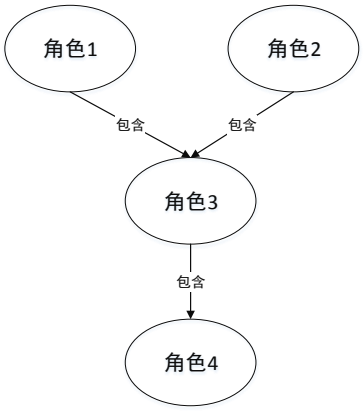


图 3.2 角色继承

定义 3.5：用户信任值 $D=\{d_1,d_2,\ldots,d_m\}, 0 \le d_m < 1, d_1 < \cdots < d_m$，表示用户信任度的 m 个等级。用户的初始信任值是一个默认的值，以后随着用户的操作信任值会随着动态的更新，每次计算用户新的信任值时其决策函数如下：

$$D(u_i)=\begin{cases} 0 & 0 \le D(u_i) < d_1 \\ d_1 & d_1 \le D(u_i) < d_2 \\ \vdots & \cdots \\ d_m & d_m \le D(u_i) < 1 \end{cases} \tag{3.1}$$

3.2 访问控制模型构建与训练

模型的构建主要是在第二章中已经预处理过的大量历史数据的基础上，通过用户信任度初始化、稀疏数据填充、行为特征库构建等操作建立基于用户行为的访问控制初始模型，具体流程如图 3.3 所示。

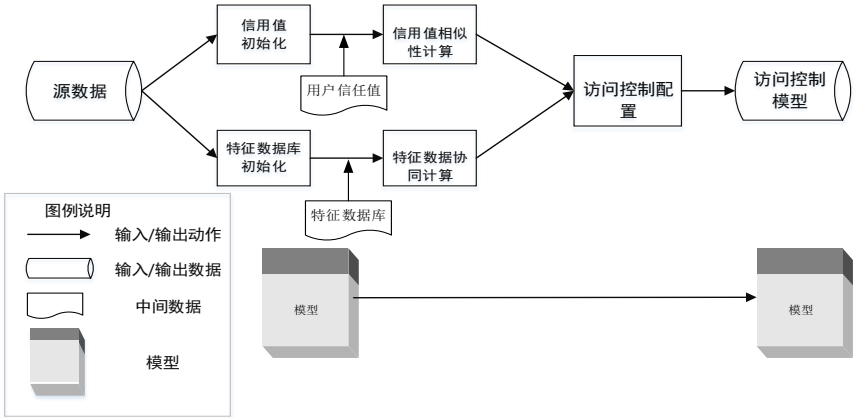


图 3.3 模型构建流程

其中信用值初始化和特征数据库初始化是可以并行执行的。首先，从已有的数据中初始化环境因素，设定各个环境的影响因子。其次，从用户当前的操作记录中，获取各个用户的初始信任值，对于数据稀疏的用户，通过其好友关系的信任值，通过相似性计算，计算出用户的信任值。然后，根据用户的行为数据，构建行为特征数据库，用于识别不同行为的安全等级。最后，从之前的计算结果中，设置相应的访问控制规则，从而构建出完整的访问控制模型。

3.2.1 信任值初始化

用户的初始信任值主要是从已经采集的用户信息中评定用户的初始信用值，假设用户信息中有 m 个属性分别为$\{e_1,e_2,\cdots,e_m\}$，设置每个属性的权重为$\{q_1,q_2,\cdots,q_m\}$，其中，$0\le q_i<1$，并且$\sum_{i=0}^{m} q_i=1$。对于每个属性的评判标准根据如下公式 3.2 进行计算。

$$q(e_i)=\begin{cases}0 & e_i=0 \ \square \ e_i=null\\ 1 & \mathrm{e_i}>0\end{cases} \tag{3.2}$$

然后，每个用户的初始信用度计算如公式 3.3 所示，最后进行归一化处理，获得具体的信任值。

$$D(u_i)=10*\sum_{i=0}^{m} q_i*q(e_i) \tag{3.3}$$

基于上述信用值初始化方法，给出信用值初始化算法。

算法 3.1 用户信任值初始化

输入： $UD=\{u_1,u_2,\cdots,u_n\}$，微博用户信息

$EM=\{e_1,e_2,\cdots,e_m\}$，所有需要考虑的属性

$QM=\{q_1,q_2,\cdots,q_m\}$，每个属性对应的权重

输出： UD，包含每个用户对应的初始信任值用户信息

1. UD word segmentation
2. UD remove stop words
3. **For** $i=1$ to n **do**
4. u_i = detail user information from table
5. Initial $D(u_i)$ = 0
6. **For** $j=0$ to m **do**
7. $D(u_i)$ = $10*\sum_{i=0}^{m} q_i * q(e_i)$
8. **End For**
9. update UD with $D(u_i)$
10. **End For**
11. **return** UD

3.2.2 信任值相似性计算

经过信用值初始化计算，得到了用户的初始信用值，但是有些用户可能由于信息的不完整性导致结果的不准确性。但是，在社交网络中，每个用户都或多或少的属于某一社区，同一社区里面的用户具有更加相似的特性，包括更加相似的信用度，如图 3.4 所示，具有相同属性用户更容易成为好友。在微博中，就类似于如果有好几个大 V 用户同时关注了你，那也能说明你拥有更好的信用值。因此，可以利用用户的好友的信用值，通过相似性计算，更新用户的初始信用值。

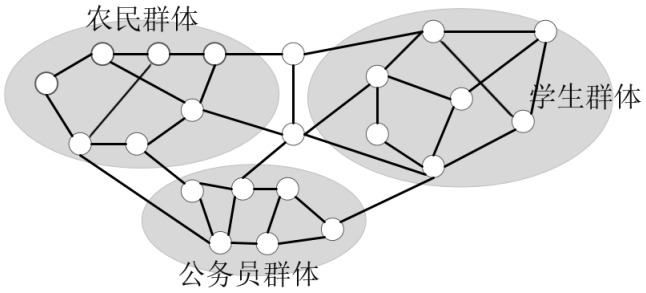


图 3.4 相同属性的用户群体

定义 3.6：用户信任值计算阈值δ，$\delta \in N+$，当用户初始信任值小于δ时，需要通过相似性计算来重置其信任值。

通过相似性计算用户的初始信任值，主要考虑用户关注的用户的信任值，记为 fd_1和用户粉丝的信任值，记为 fd_2。因此，用户的相似性信任值记为：

$$FD(u_i) = \sin^2 a * fd_1 + \cos^2 a * fd_2 \qquad 0 \le a \le 30 \tag{3.4}$$

其中，$0 \le a \le 30$主要是防止攻击者通过恶意的关注大量的高 V 用户来提高自己的信用值，因此限制用户关注的用户信用值所占比例，提高用户粉丝信用值所占比例。

用户关注用户信用值 fd_1 计算如下：

$$fd_1 = \sum_{i=0}^{m} fd(u_i) / m \tag{3.5}$$

其中，$fd(u_i)$ 为关注的每个用户的初始信任值，m 为总的关注用户数量。

用户粉丝信任值则采用令一种方法，由于用户的粉丝可能上千万，所以对于用户粉丝信任值，综合考虑粉丝的数量和质量。对于大量的粉丝，从中选取信任值最高的 100 个用户作为特征数。因此，$fd2$ 的计算如下：

$$fd_2 = a * f(uc) + b * \sum_{i=0}^{m} top(fd(u_i)) / m \qquad 0<\mathrm{m} \le 100, a+b=1 \tag{3.6}$$

其中，$f(uc)$ 的定义如下：

$$f(uc) = \begin{cases} 0 & \text{count} \le 10 \\ 0.2 & 10<\text{count} \le 100 \\ 0.4 & 100<\text{count} \le 500 \\ 0.6 & 500<\text{count} \le 2000 \\ 0.8 & 2000<\text{count} \le 100000 \\ 1 & \text{count}>100000 \end{cases} \tag{3.7}$$

$top(fd(u_i))$ 指根据用户的信任值进行从大到小排序，获取前面最大的 m 个用户的信任值。

算法 3.2 用户信任值相似性计算

输入： $UD = \{u_1, u_2, \cdots, u_n\}$，包含有用户初始信任值的微博用户信息

$UF = \{f_1, f_2, \cdots, f_m\}$，用户好友关系

α，需要进行相似性计算的用户信任值阈值

输出： UD，包含每个用户经过相似性计算的用户信息

12. **For** $i = 1$ to n **do**

13. u_i = detail user information from table
14. $D(u_i)$ = user trust data from u_i
15. **If** $D(u_i) < \delta$ **then**
16. **For** $j = 0$ to m **do**
17. u_j is friend for u_i
18. **Initial** $fd1 = 0$ **,** $fd2 = 0$
19. **Initial** $friendCount = 0$ **,** $followCount = 0$
20. **If** u_i follow u_j **then**
21. $fd_1 = fd_1 + fd(u_j)$
22. $friendCount + friendCount + 1$
23. **Else**
24. $fd_2 = fd_2 + fd(u_j)$
25. **End For**
26. $fd_1 = fd_1 / friendCount$
27. $fd_2 = a * f(uc) + b * fd_2 / folowCount$
28. $FD(u_i) = \sin^2 a * fd_1 + \cos^2 a * fd_2$
29. update UD with $FD(u_i)$
30. **End For**
31. **return** UD

3.2.3 特征数据库构建

在入侵检测系统中，通常会有一个正常行为模式库，对于每一次的行为，都与模式库进行匹配，从而判断行为的合法性，如果行为不合法，则立即启动入侵响应系统。本文参考入侵检测系统的设计，设计用户行为特征数据库，针对每种特征设计不同的安全策略，并采用遗传算法来进行特征数据库的初始化。

遗传算法[56]（Genetic Algorithm）是一种模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型，遗传算法的运行过程为一个典型的迭代算法，他在每一次迭代时产生一组解答，这组解答最初是随机生成的，在每一次迭代时又产生一组新的解答。然后计算出每一组解答的适应度，重复执行该步骤，直到函数符合一定的规则或者达到某种形式上的收敛。遗传算法的基本流程如图 3.5 所示。

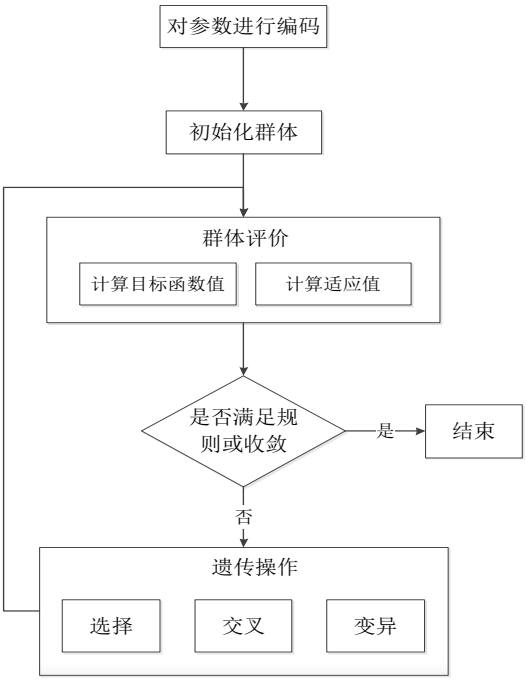


图 3.5 遗传算法基本流程

1. 特征编码

在编码时采用二进制编码，每一个八位的二进制数代表着一种操作，其中前两位代表操作类型，后 6 位代表着操作方法。具体如表 3.1 和 3.2 所示，具体可表示为一个集合 $X=\{x_1,x_2,\cdots,x_m\}$,每一个元素 $x_i(0\le i\le m)$ 对应着一种特定的编码。

表 3.1 操作类型编码

操作类型	编码
读（read）	01
写（write）	10
执行(execute)	11

表 3.2 操作方法编码

操作方法	编码
浏览个人微博	000000
浏览大 V 微博	000001
浏览私人微博	000010
发布个人微博	000011
…	…

2. 适应度函数

在遗传算法中，适应度(fitness)用于度量某个物种对于生存环境的适应程度。在定义适应度函数之前，先定义几个参数：

定义 3.7：参考模式库T，T为提前定义好的一系列操作的安全等级。用于作为后续选择的参考。

定义 3.8：$C(T)$为参考模式库T中所有记录的数量。

定义 3.9：相似性$S(X_i,T_j)$表示X 中的某条记录x_i与T中的某条记录t_j 的相似性。其中，相似性采用的是短系列匹配法。

定义 3.10：适应度函数：

$$F(x)=\frac{\max\{S(x,t_j),0\le j<C(T)\}*C(T)}{\sum_{j=0}^{C(T)}S(x,t_j)} \tag{3.8}$$

3. 遗传操作

遗传操作主要有选择（Selection）、交叉（Crossover）和变异（Mutation）三种。其中，选择是一种基于适应度的优胜劣汰的过程，即从当前群体中选择适应度高的个体形成新的群体。具体的选择策略有很多，本文采用了轮盘赌选择，其基本思想是各个个体被选中的概率与其适应度大小成正比，其基本思路如下：

(1) 计算各个个体的适应度$F(x_i),0\le i<m,m$为群体大小。

(2) 计算每一个个体被遗传到下一代的概率

$$P(x_i)=\frac{F(x_i)}{\sum_{j=0}^{M}F(x_j)} \tag{3.9}$$

(3) 计算出每个个体的累计概率，计算公式为：

$$Q(x_i)=\sum_{j=0}^{i}P(x_j) \tag{3.10}$$

(4) 在[0,1]之间产生一个随机数r，假设$Q(x_i)\le r<Q(x_{i+1})$，则选择个体x_i。

(5) 重复执行步骤(3)和(4)m次。

对于交叉，在遗传学上也称基因重组或杂交，是指两个染色体的某一个相同位置处 DNA 被切断，前后两串分别交叉组合形成新的两个新的染色体。本文采用类似的方法来进行种群的遗传。本文中随机选择双亲，然后在第二位后面切断，进行单点交叉，交叉后得到两个新的特征编码，如图 3.6 所示。

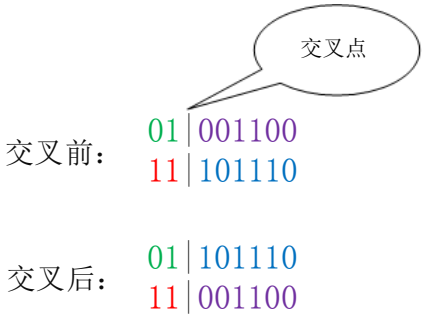


图 3.6 交叉前后对比

变异在遗传学上是指基因在复制时，其基因组 DNA 在结构上发生了碱基对组成或者排列顺序的改变，并且这种改变是可以遗传的。在本文中，对于 x_i 中的每一位二进制数，其均有一定的概率（记为 p_m ）发生变化，变成其允许范围内的另一个数。具体的，定 $p_m = 0.02$，每次突变时，如果原本为 0 则变为 1，原本为 1 则变成 0。每条记录 x_i 发生突变的概率为：其中 $length(x_i)$ 为 x_i 中二进制的位数。

$$PM(x_i) = 1-(1-p_m)^{length(x_i)} \tag{3.11}$$

对于变异的前后变化，如图 3.7 所示。

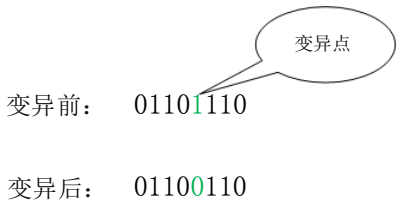


图 3.7 变异前后对比

通过重复的执行遗传算法，不断的淘汰与更新，使得整个群体的适应值明显提高，初步得到一个满足要求的特征数据库。

3.3 访问控制模型更新

用户的行为是一个动态变化的过程，用户的使用习惯会随着用户需求或者一些外界因素的改变而发生变化，因此访问控制模型需要实时更新以满足用户行为的变化。

访问控制模型的更新主要包括两部分，如图 3.8 所示。一方面，随着时间的推移，用户的信任值是会类似于人的记忆一样不断的衰减，同时，用户的操作结果又会导致用户的信任值动态的发生变化；另一方面，由于外界环境不断变化，

特征库也需要不断的更新以满足外界的变化，本文中采用遗传算法对特征数据库进行更新。

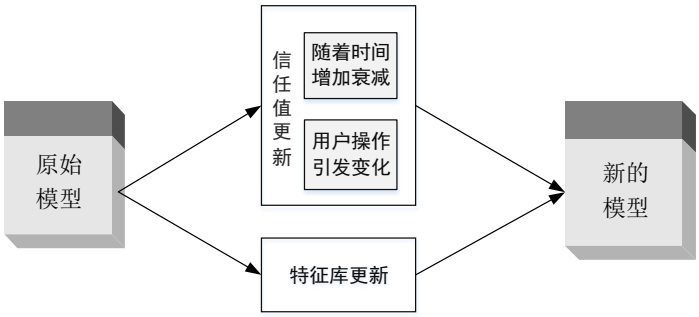


图 3.8 访问控制模型更新流程

3.3.1 基于记忆原理的用户信用值更新

在 3.2 节中根据历史数据利用公式 3.12 和公式 3.13 对用户的初始信用值进行了初始化，主要解决了数据的冷启动和数据稀疏问题。但是，用户的信任值是一个动态变化的过程，随着用户可信行为的累计，用户的信任值会不断增加，而大量的不安全行为则会导致信任值值减少甚至被禁止访问。同时，用户信任值具有时间相关性，用户的信任值会随着时间的流逝而衰减。总体来说，用户的信任值变化就是原始的信任值加上用户变化的信任值，而用户变化的信任值主要由三部分组成，分别是瞬时记忆，短期记忆和长期记忆，三部分的转换如图 3.9 所示。用户每一次的操作都会引起信任值的变化，但是只有记忆长久的操作才会产生更大的影响，瞬时记忆引起的变化则更小。

$$D(u_i) = \sum_{i=0}^{m} e_i * q(e_i) \tag{3.12}$$

$$FD(u_i) = \sin^2 a * fd_1 + \cos^2 a * fd_2 \qquad 0 \le a \le 30 \tag{3.13}$$

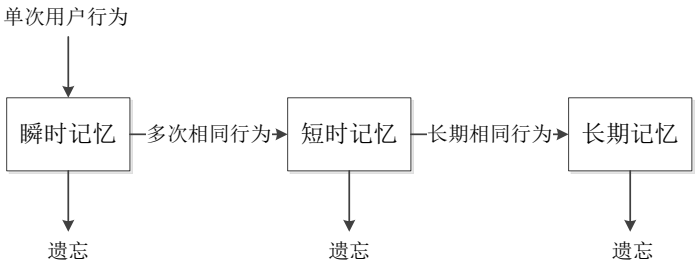


图 3.9 记忆与遗忘关系

用户的信任值更新有时间驱动和事件驱动两种方式，在本文中，同时具备了时间驱动和事件驱动，时间驱动是指每隔一段时间都对用户的信任值进行一次重

新计算，假设间隔时间为t'，记为步长。事件驱动是指每一次的用户操作都产生一次信任值更新，可有效避免t'长短带来的精确度与计算复杂度之间的矛盾，区别如图 3.10 所示，每个黑点为信任值更新点。t'为一个步长，e'为一次行为。

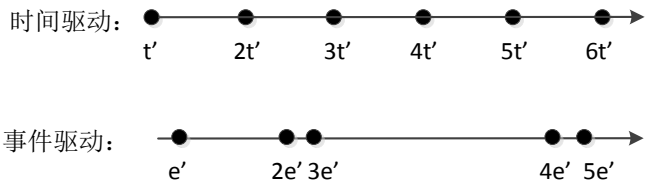
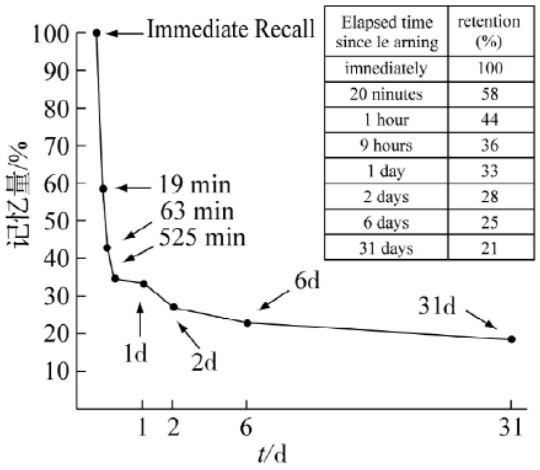


图 3.10 时间驱动与事件驱动的区别

用户的信任值遗忘则采用了符合艾宾浩斯的遗忘曲线，即遗忘的速度是不均匀的，起初遗忘的速度最快，然后会逐渐变慢，慢慢地趋于一个稳定值，随后遗忘速度停止。如图 3.11 所示[57]，竖坐标轴表示记忆量，横轴表示时间。



Elapsed time since le arning	retention (%)
imnediately	100
20 ninutes	58
1 hour	44
9 hours	36
1 day	33
2 days	28
6 days	25
31 days	21

图 3.11 艾宾浩斯遗忘曲线

在本文中，我们根据记忆原理，提出了用户信任值的更新流程，更新流程如图 3.12 所示。

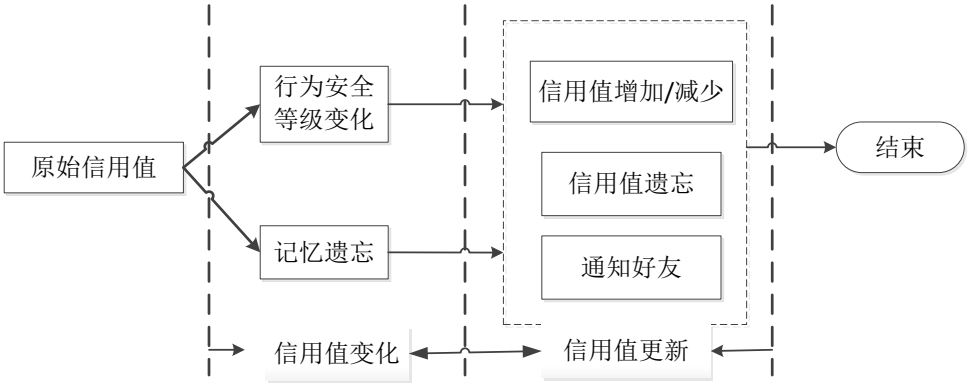


图 3.12 信用值更新流程

其中，信任值流程主要分成两部分，一部分是用户行为产生的信任值更新驱动，具体流程如图 3.13 所示。另一部分为时间驱动，主要是参考艾宾浩斯遗忘曲线，提出信任值遗忘曲线，并利用信任值遗忘曲线进行更新。

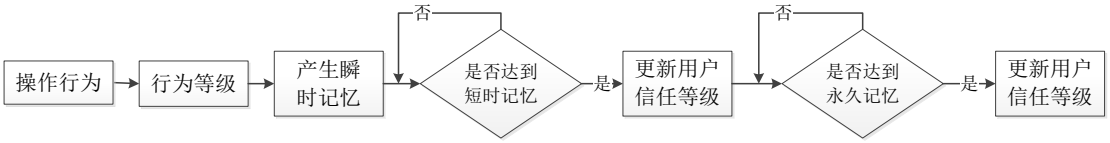


图 3.13 用户信任值事件驱动更新流程

3.3.1.1 记忆遗忘触发更新

对于记忆遗忘的更新流程，本文在算法中保留了用户近一段时间的信任值，每隔一个周期保留一个值，假设更新周期为T。

定义艾宾浩斯遗忘曲线公式，其中k为调节参数，反应的是遗忘程度的快慢，k值越大，表示遗忘的速度越快，相反则表示遗忘的速度越慢。

$$Y=\frac{1}{k*x^{0.2}+1} \tag{3.14}$$

下图列举出当k=1 和k=1/4 是对应的艾宾浩斯遗忘曲线，从图 3.14 中可以看出，一开始的遗忘速度较快，越往后面遗忘速度越慢，最终稳定在某个值。

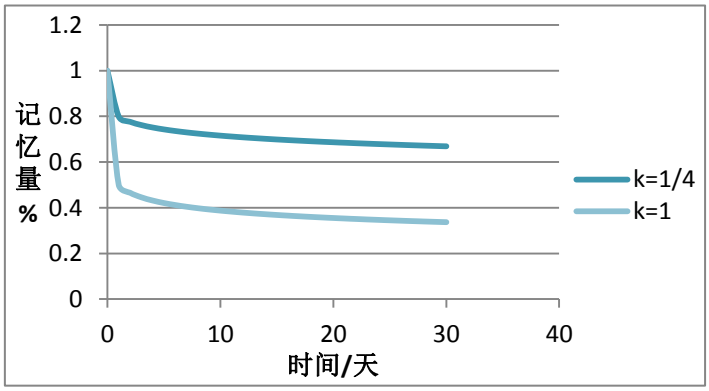


图 3.14 不同k值对应的艾宾浩斯遗忘曲线

根据已经定义的艾宾浩斯遗忘曲线，我们给出了用户信任值遗忘曲线，如公式 3.15 所示：

$$D'(\Delta t)=\frac{1}{k*(\frac{\Delta t}{T})^{0.2}+1}*D(\Delta t) \tag{3.15}$$

其中，$D(\Delta t)$为用户上一次信用值，$D'(\Delta t)$为经过一段时间遗忘后的所代表的值，Δt表示更新D值时到现在所经过的时间。

根据用户信用值遗忘曲线以及近段时间的用户历史信用值，其中历史信用值取值时间点Δt分别为$\{T,2T,\cdots,mT\}$， m为历史记录因子，表示记录的用户历史信任值数目，定义用户信用值计算方法为：

$$\begin{aligned} D' &= \lambda * \frac{\sum_{1}^{m} D'(\Delta t)}{m} + (1-\lambda)*D(u) \qquad \Delta t=\{T,2T,\cdots,mT\} \\ &= \lambda * \frac{\sum_{1}^{m} \frac{1}{k*tw^{0.2}+1} * D(T)}{m} + (1-\lambda)*D(u) \quad \Delta t=\{T,2T,\cdots,mT\} \end{aligned} \tag{3.16}$$

其中$tw=\frac{\Delta t}{T}=\{1,2,\cdots,m\}$， $0\le\lambda\le1$， $D(u)$为经过信任值初始化和信任值相似性计算后得到的用户初始信任值。

3.3.1.2 用户行为触发更新

用户每一次的操作行为都会有一个反馈，良好的操作行为会形成正反馈，而不安全的行为则会形成负反馈。根据记忆原理，每一次反馈都会对用户信用值产生一次瞬时记忆，当瞬时记忆多次出现累计成短时记忆时就会触发用户信任值更新机制，从而上调、下调用户的信任值。

假设用户自上一次用户信任值更新后一些列操作的安全等级分别为$\{d_1,d_2,\cdots,d_m\}$，那么用户当前的瞬时记忆累计为

$$d=\sum_{1}^{m} d_i \tag{3.17}$$

如果d大于某一个设定的阈值，那么就启动信任值更新程序，同时，为了防止用户蓄意通过不断的合法进行等级较低的服务来提高自己自身的信任值，以便于信任值提高到某一等级后就可以对安全性较高的服务进行破坏，本文还设定了每次更新的信任值上限m，信任值上限主要和用户之前信任等级以及本次操作的平均信任等级有关。因此用户信任值的更新如下。由于d是信用值更新的触发条件，所以d必然大于0或者小于0，不存在等于0的情况。

$$d_{new}=\begin{cases} d_{old}+1 & d>0 \ \&\& \ (d/m)>d_{old} \\ d_{old} & d>0 \ \&\& \ (d/m)\le d_{old} \\ d_{old}-1 & d<0 \end{cases} \tag{3.18}$$

算法 3.3 用户信任值更新

输入： $UHD=\{D_1'(\Delta t),D_2'(\Delta t),\cdots,D_n'(\Delta t)\}$，用户近一段时间的信任值

t，固定的时钟触发

$D=\{d_1,d_2,\cdots,d_m\}$，近段时间操作安全等级

$D(u)$ 用户的初始信任值

输出： *UHD*，更新后的用户历史信任值

1. **If** $t>0$ && $t \bmod T=0$ **then**
2. **Initial** $D'=0$
3. **For** $i=1$ to n **do**
4. $\Delta t=i*T$
5. $D'(\Delta t)=\lambda*\dfrac{1}{k*(i)^{0.2}+1}*D(T)+(1-\lambda)*D(u)$
6. $D'=D'+D'(\Delta t)$
7. **End For**
8. $D'=D'/m$
9. update *UHD* with new D'
10. **Else**
11. $D'=D_1'(0)$
12. Initial $d-summary=0$
13. **For** $i=1$ to m **do**
14. $d-summary=d-summary+d_i$
15. **End For**
16. **If** $d-summary>0 \&\& d-summary>D'$ **then**
17. **If**
18. $D'=D'+1$
19. **Else If** $d-summary<0$ **then**
20. $D'=D'-1$
21. **End If**
22. update *UHD* with new D'
23. **End If**
24. **return** *UHD*

3.3.2 基于遗传算法的特征数据库更新

在 3.2.3 节中，本文通过遗传算法对用户的历史行为数据进行处理，得到初始的特征数据库。同时，由于各种攻击手段不断更新，各种行为模式也在与日激增，因此，特征数据库并不是一成不变的，而是在实际的使用过程中不断的迭代更新，以适应外界的环境变化。特征数据库的更新主要来源于两部分，一方面是模型根据已有的数据通过遗传算法自动更新，另一方面则是根据反馈信息，人工干预下的更新。具体的更新流程如图 3.15 所示。

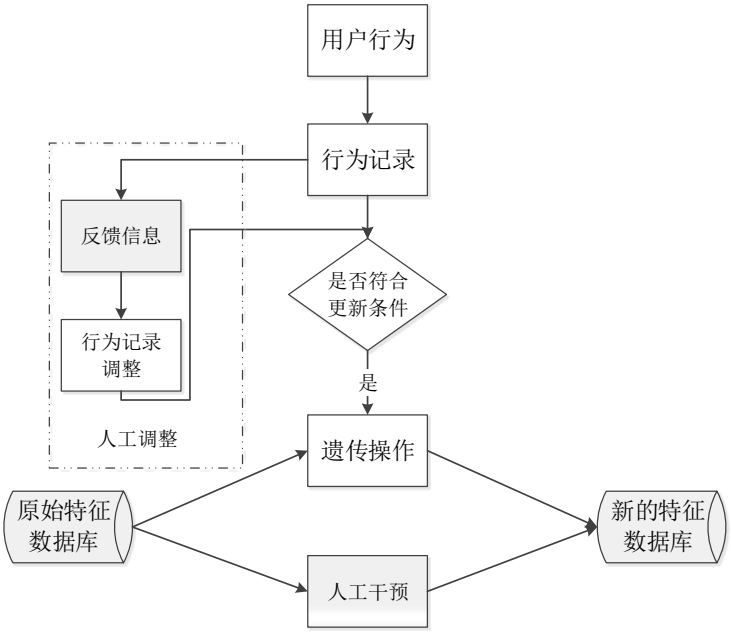


图 3.15 特征数据库更新流程

用户的行为记录可以用函数映射 $ub \to f(x)$ 来表示，其中 ub 为用户操作行为，用户的每一次操作，最终都会得到一个用户的行为等级评定， $f(x)$ 代表着用户当前的行为评定。针对每一条行为记录，可以根据使用者或者管理员的反馈信息，调整 $ub \to f(x)$ 对应的 $f(x)$ 的值。

定义 3.11：确定因子 $P(ub \to f(x))$，表示用户当前行为 ub 最终评定为 $f(x)$ 的确定程度。 $P(ub \to f(x)) \in [0,1]$，数值越高，表示越确定。

定义更新阈值 δ，当 $P(ub \to f(x)) > \delta$ 时就会通过一次遗传操作，将该次行为记录通过遗传操作加入到特征数据库中。对于每一次的人工调整，调整过程需同时给出 $P(ub \to f(x))$ 的值。给出的值可能大于 δ，也可能小于 δ。

人工干预则是利用专家知识，根据最新的研究成果，人为的去对对特征数据库中的记录进行修改，包括删除过时或者错误的特征数据，添加最新的特征数据

等等。

3.4 仿真实验与结果分析

3.4.1 实验环境与实验方法

实验采用 Java 实现，在 Intel(R) Core(TM) i5-3470 CPU @ 3.2GHz 3.20GHz、8G 运行内存的 Windows 8.1Pro 平台上运行，编译器为 Eclipse Juno(4.2)+JDK1.7。后台数据库为 MySQL5.0。

实验中首先通过三组用户数据利用算法 3.1 和算法 3.2 进行 *UA-BACM* 的构建和训练，统计得出用户的信任值概率分布。其中，需要考虑的用户信息属性以及各个属性所在比例具体设置如表 3.3 所示。其中，粉丝数、关注数、微博数和用户互粉数均按照一定的等级划分给分。

表 3.3 各属性所占比例设置

属性	所在比例
用户昵称	2%
用户所在地	6%
用户个人描述	4%
粉丝数	10%
关注数	4%
微博数	5%
用户互粉数	6%
创建时间	1%
是否允许定位	7%
是否实名认证	26%
用户头像	3%
手机号是否认证	13%
邮箱是否认证	13%

然后再利用算法 3.3 进行用户信任值的更新，以个别用户为例，获得用户的信任值变化曲线。在实验过程中，分别设置T 为 24 小时、 $m=30$、 $\lambda=0.6$，也就是说保存下用户最近的 30 个信任值，对于初始情况设置这 m 个$D'(\square t)$ 均为用户的初始信任值。同时，实验中还信任值波动的矫正情况进行了测试，测试时取

$m=10$，观察用户的信任值遗忘曲线，用户的历史信任值如表 3.4 所示。

表 3.4 用户历史信任值

分组	历史信任值	用户初始信任值
实验组 1	6，6，6，6，6，6，6，6，6，10	6
实验组 2	6，5，6，7，6，6，5，6，9，10	6
实验组 3	6，6，4，6，6，7，6，6，10，9	6

最后，将 *UA-BACM* 与当前较为经典的访问控制进行对比，包括自主访问控制模型（*DAC Model*）、强制访问控制模型（*MAC Model*）、基于角色的访问控制模型（*RBAC Model*）、基于行为的访问控制模型（*ABAC Model*）、基于任务的访问控制模型（*TBAC Model*）。用于判断当前 *UA-BACM* 的实际效果。

3.4.2 实验结果及分析

3.4.2.1 模型构建实验结果及分析

图 3.16 显示的是针对三组实验数据经过信任值初始化后得到的用户信任值概率分布图。

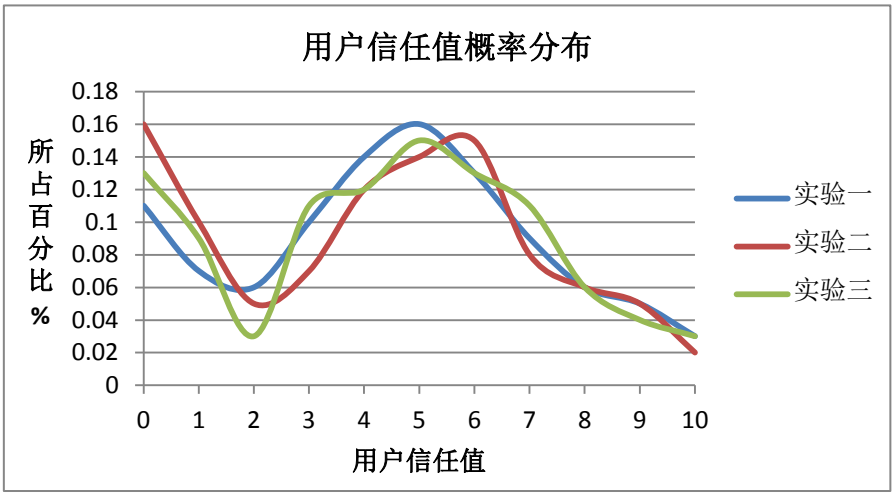


图 3.16 初始用户信任值概率分布

从图 3.16 中可以看出，概率分布整体上符合一定的高斯分布，说明用户信任值的初始化过程中所要求的属性设置以及各个属性所在比例是合理的。而对于信任值为 0 或 1 所在用户较多导致不符合高斯分布的情况，主要是因为爬取的用户中存在着较多的注册后没有使用或者只关注别人不发微博的用户。对于这部分用户无法直接通过信任值初始化获取其对应的信任值。

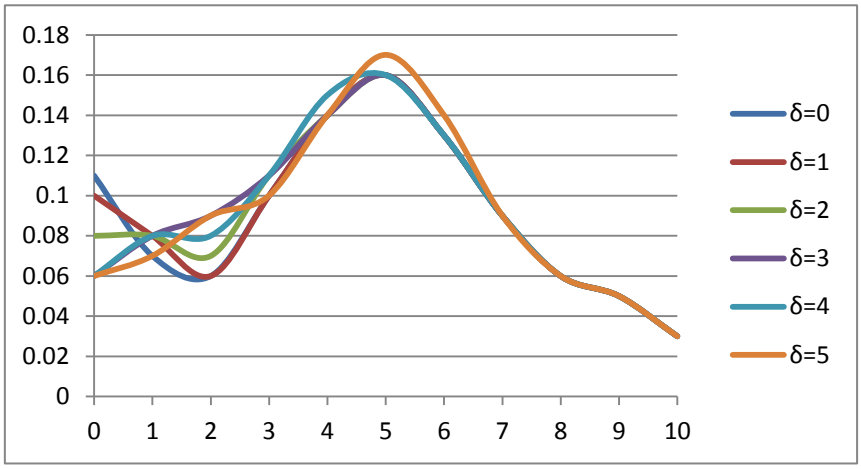


图 3.17 δ 对用户信任值概率分布的影响

计算完用户的初始信任值之后，需要对用户信任值进行相似性计算。其中，设置不同的δ会对用户信任值分布曲线产生不同的影响，图 3.17 显示的是不同δ下针对实验数据一进行信任值相似性计算得到的用户信任值概率分布图，从图中可以看到，随着δ的增大，曲线更倾向于高斯分布，但δ=3 及之后曲线变化很小，而随着δ的增大，计算量却快速增加，所以最终选择δ=3 应用于后续的研究中。

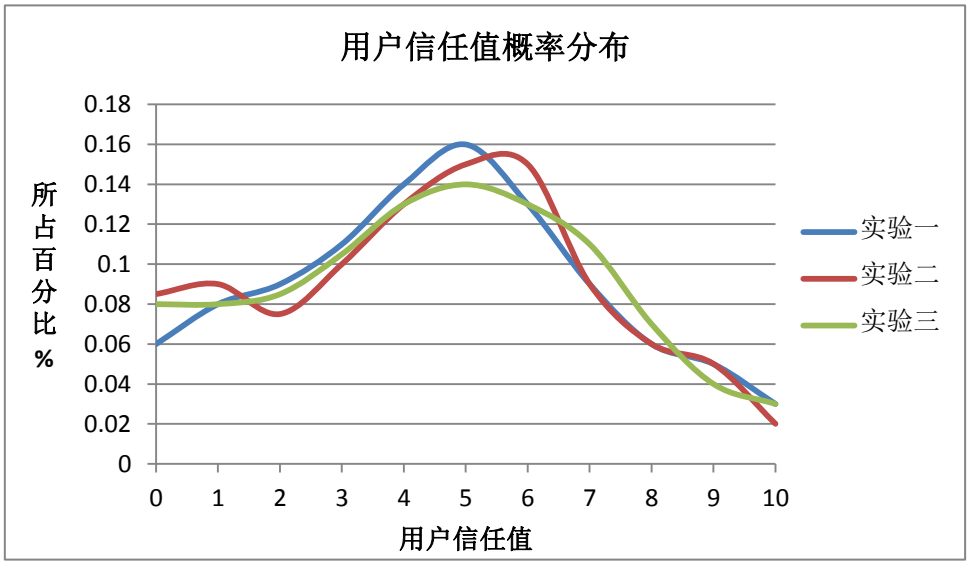


图 3.18 经过相似性计算后的用户信任值概率分布

图 3.18 显示的则是对于实验中的三组数据，在经过用户信任值初始化之后，又通过δ=3 的相似性计算后的用户信任值概率分布，从图 3.18 中可以看出，相比于图 3.16，图 3.18 所示的曲线更加的接近于高斯分布函数，图 3.16 中所阐述的用户初始信任值为 0 或者 1 过多的问题通过相似性计算得到了较好的解决。进一步验证了用户信任值初始化算法的可行性和准确性，也验证了各属性所占比例

设置的合法性。

3.4.2.2 模型更新实验结果及分析

图 3.19 给出当用户在 30 天内无任何操作时的用户信任值变化曲线，当用户一直无操作的时候，用户的信任值会随着时间的流逝而逐渐减小，最终倾向于零。图 36 中展示的是不同的 k 值条件下信任值遗忘的速度。k 值越大，信任值遗忘就越快，通过对实验结果的研究以及后续研究的需要，本文最终选择 $k=0.2$ 作为后续研究的基础。

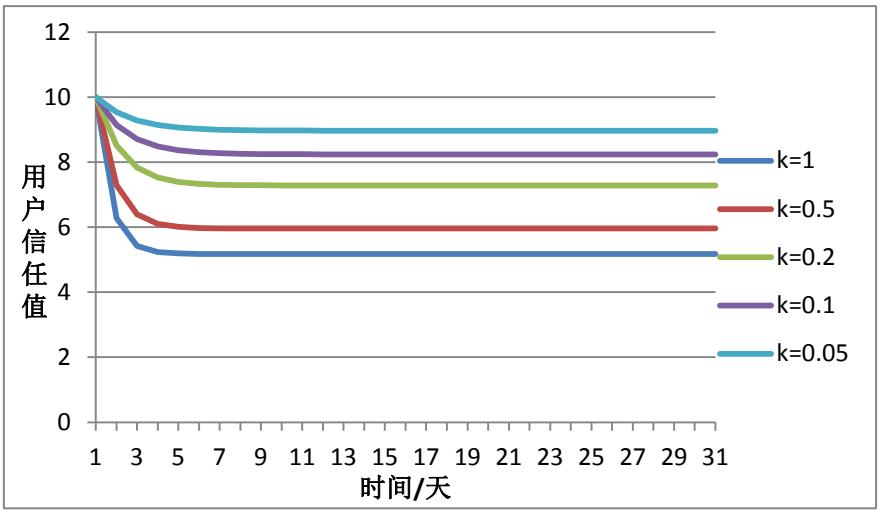


图 3.19　k 值与用户信任值遗忘曲线关系

图 3.20 给出的则是表 3.4 中用户信任值发生波动时对应的信任值遗忘曲线，在图 3.20 中，第 1 天到第 11 天是用户的历史信任值，信任值分别在第 10、11 天产生了较大的波动（可能是攻击者通过一系列方法恶意提高自己的信任值），然后是用户的遗忘曲线分布，从图中可以看出，遗忘曲线能够很好的规避这种波动带来的影响，快速降回到正常水平，后续逐渐趋于平稳。验证了信任值遗忘曲线能够较好的防止短时的信任值波动。

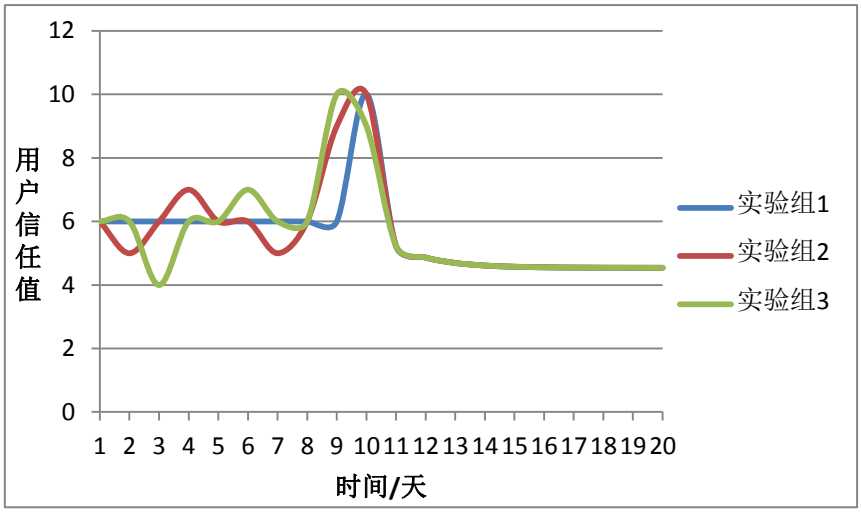


图 3.20 用户信任值波动时对应的遗忘曲线

3.4.2.3 *UA-BACM* 与其他访问控制模型对比分析

UA-BACM 综合了基于角色的访问控制模型以及基于行为的访问控制模型的优点，与当前几种访问控制模型的特性比较如表 3.5 所示。

表 3.5 与已有访问控制模型特性比较

	角色	时态	行为	移动网络	计算量	环境	自我学习
DAC Model	不支持	支持	不支持	一般	小	不支持	不支持
MAC Model	不支持	支持	不支持	一般	小	不支持	不支持
RBAC Model	支持	支持	不支持	一般	小	一般	不支持
ABAC Model	支持	支持	支持	一般	一般	一般	不支持
TBAC Model	支持	支持	支持	一般	一般	支持	不支持
UA-BACM	支持	支持	支持	支持	大	支持	支持

从表 3.5 中的对比中可以看出，*UA-BACM* 更加适合移动互联网环境下访问控制，同时还支持自我学习方式，能够不断的在实际使用中调整自己以适应外界环境的变化，但相对于其他的访问控制模型，最大的弊端就是计算量较大，需要消耗较多的计算机计算能力。

3.5 本章小结

本章提出了移动互联网环境下基于用户行为的访问控制模型 *UA-BACM*，首先给出了 *UA-BACM* 的数学表示；然后详细介绍了模型的构建与训练过程，包括

信任值初始化、信任值相似性计算、特征数据库初始化等；紧接着，给出了模型的更新方法，即基于记忆原理的用户信任值更新和基于遗传算法的特征数据库更新；最后，通过仿真实验验证了 *UA-BACM* 在构建、训练、更新过程的正确性，并通过将 *UA-BACM* 与其他访问控制模型进行对比验证了 *UA-BACM* 的先进性与准确性，为后续基于用户行为访问控制模型的动态 k-匿名隐私保护方法的研究提供基础。

第4章 基于用户行为访问控制模型的 K-匿名隐私保护方法

在第三章的研究基础上，通过基于用户行为的访问控制模型获得访问安全等级，并利用映射函数获得用户隐私需求，根据隐私需求给出移动互联网环境下基于用户行为访问控制模型的动态 *k*-匿名隐私保护方法，包括动态 *k*-匿名隐私保护方法和基于四叉树的动态位置 *k*-匿名方法，在时空域上对用户在移动互联网环境下的行为数据进行泛化，为不同安全等级的操作提供不同泛化程度的隐私数据。

4.1 问题提出

移动网络环境下产生的数据相对于传统网络更加的复杂多样，各式各样的网络环境，数目众多的操作平台，精确到米的定位技术等。这些特性使得服务商对于用户隐私数据的保护显得更加的困难。当前，针对用户的隐私数据保护，服务商会在数据发布时采用 *k* -匿名等方案对数据进行一定的泛化处理。而在当前的基于位置的服务（LBS）中，对于用户的位置信息，主要的研究思路则是用户在将数据传给 LBS 服务提供商前先进行 *k*-匿名处理，从而来保护自己的位置信息，具体的解决方案主要有二大类：集中式模式和分布式模式。

对于集中式的位置匿名方法，其主要思路是在移动用户和 LBS 服务提供商之间有一个可信的第三方位置匿名服务器，所有的位置请求服务都经过这个第三方服务器匿名处理后再发给 LBS 服务提供商，其系统结构如图 4.1 所示。

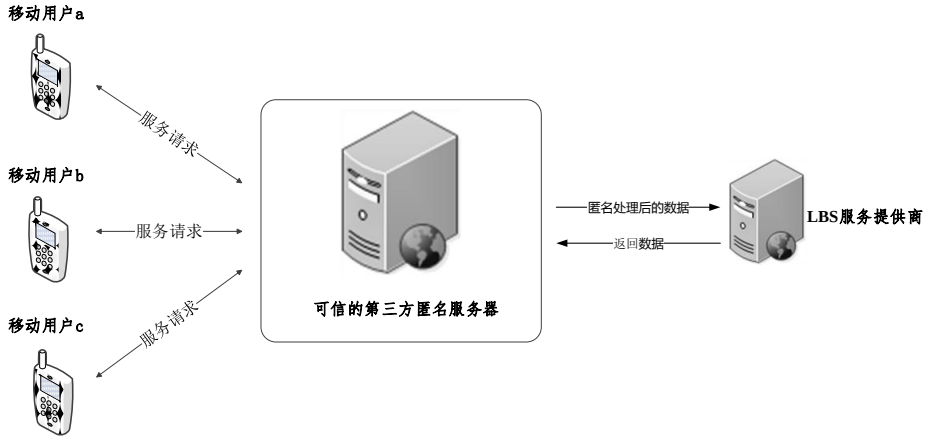


图 4.1 集中式位置匿名系统

与集中式位置匿名方法相反，分布式的位置匿名方法则是一种 p2p 模式下的

用户自组织位置匿名方法。该方法利用用户间组成的自组织网络，通过多跳通信手段，查找周围的 k 个用户，生成一个包含 k 个用户的匿名区，然后将该匿名区发送给 LBS 服务提供商。其系统结构图如图 4.2 所示。

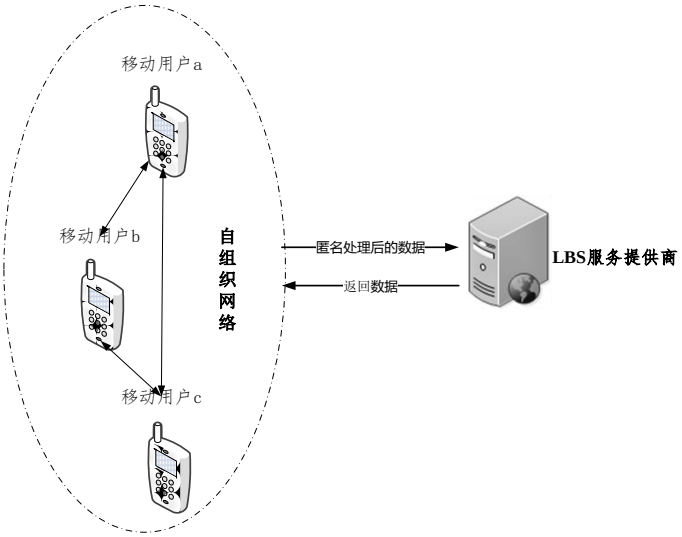


图 4.2 分布式位置匿名系统

但是，当前的位置匿名方案使得即使是可信的 LBS 服务提供商，也无法知道用户的确切位置，当然也就无法为用户提供精确的服务。图 4.3 所示为用户请求获取周边微博时的一种服务场景，该该场景下，用户获得的服务效果可能与预想中的效果存在很大的区别。

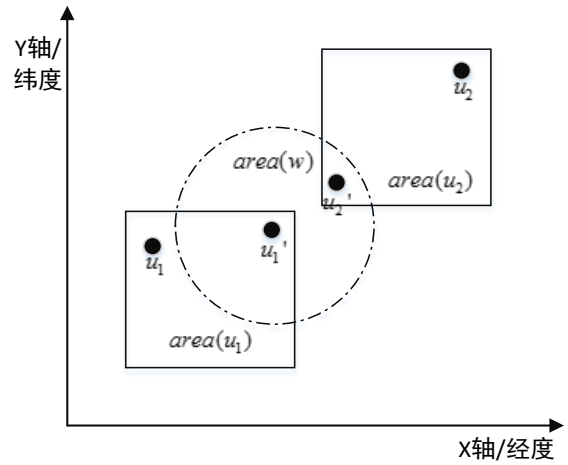


图 4.3 位置匿名带来的数据失真

具体的，在图 4.3 中，现在用户 u_1 发起查看周边微博的服务请求，其所在匿名区为 $area(u_1)$，LBS 服务提供商根据 u_1 的请求以及 u_1 传给 LBS 服务提供商的位置返回某一区域的微博信息。记用户实际发给 LBS 服务提供商的位置为 u_1'。同时记 LBS 提供商返回的区域为 $area(w)$，然后 u_1 看到了用户 u_2 发布的微博信息，

由于 u_2 在发布微博的时候也对自己的位置进行了匿名处理，其匿名区域为 $area(u_2)$，记 LBS 服务提供商收到的 u_2 的位置 u_2'。

定义 4.1：位置偏差 $\delta(x,x')$，表示用户请求时理想中的位置 x 与实际的位置 x' 之间的距离。

定义 4.2：偏差叠加，偏差叠加是指当好几个偏差叠加在一块的时候，可能会产生了一个超出预期的偏差。

在图 4.3 所示的情况中，用户理想中的位置就是用户 u_1 本身所在的位置，而实际的位置则是用户 u_2 所在的位置。所以图 4.3 所示情况的位置偏差就应该为 $\delta(u_1,u_2)$。记 $\delta(u_1,u_2)$ 在最坏情况下的值为 $\delta'(u_1,u_2)$，则 $\delta'(u_1,u_2)$ 应该为：

$$\begin{aligned}\delta'(u_1,u_2)=&\sqrt{width(area(u_1))^2+height(area(u_1))^2}+width(area(w))/2\\&+\sqrt{width(area(u_2))^2+height(area(u_2))^2}\end{aligned}\tag{4.1}$$

其中，*width* 表示该区域对应的宽度，*height* 表示该区域对应的高度。那么，当用户较为稀疏的时候，也就是匿名区 $area(u_1)$ 和匿名区 $area(u_2)$ 均处在一个用户可容忍的最大位置偏差的时候，这时候用户实际得到的值却可能是一个超出了其容忍范围的服务。这就是偏差叠加导致的数据失真问题。

同时，当前的数据泛化方法基本是在数据发布时进行统一的匿名处理，处理完成后的数据写入发布数据库，以后所有用户，不管是合法用户还是非法用户得到的均是同样的数据。具体的系统架构如图 4.4 所示，由于所有用户获得的数据都是相同的，那么如果数据泛化程度较大，将导致合法用户获取的数据失真较为严重，服务质量较差。如果数据泛化程度较少，那么攻击者就可以从查询到的数据中通过一系列攻击手段获得用户的部分获取全部隐私信息。

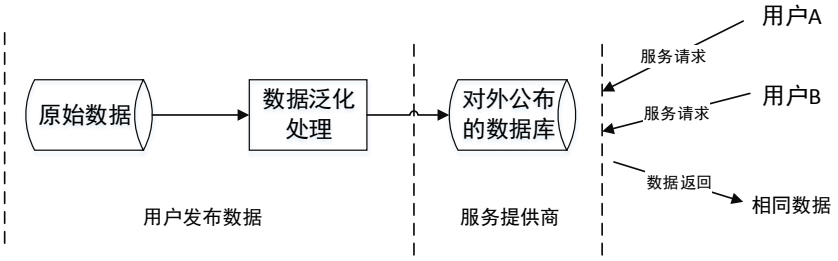


图 4.4 当前数据隐私保护系统架构

4.2 支持数据保真的 *k*-匿名隐私保护方案

针对 4.1 节所描述的 *k* -匿名隐私保护方案的数据失真问题，本文在第三章基于用户行为的访问控制模型的研究基础上，提出了移动互联网环境下基于 *k*-匿名的数据保真隐私保护方法。具体的，根据基于用户行为的访问控制模型所得到的

安全等级，为每一次访问行为提供不同泛化程度的数据，从而保证高安全的行为可以获取高保真的数据，而攻击者只能得到严重失真的数据，甚至被拒绝访问。从而保证服务提供商特别是 LBS 服务提供商在提供高质量服务（QOS）的同时又能有效的保护用户的隐私数据。具体的系统架构如图 4.5 所示，系统为不同等级的行为提供不同的数据，并通过缓存方式减少计算的次数。

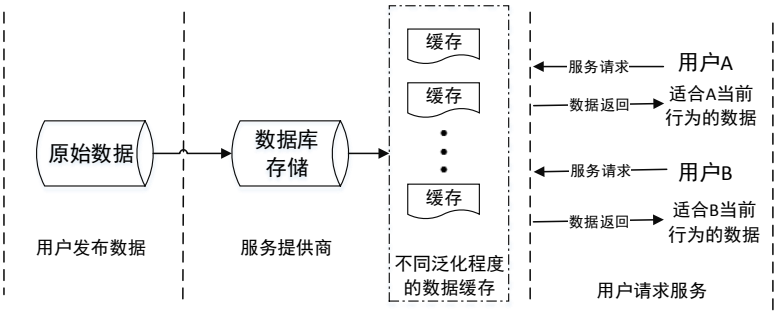


图 4.5 隐私保护系统架构

对于存储在服务提供商数据库中的数据，如用户的医疗数据、微博数据、运动数据等，一般为数据表形式，表中的每一条记录（或每一行）对应一个人，包含有多个属性值，这些属性主要可以分成五类，具体如下：

定义 4.3：属性 A，$A=\{A_1,A_2,\cdots,A_m\}$，$A_i(1\le i\le m)$ 对应于表中的一个字段。

定义 4.4：标识符 I（Identifier），$I\subseteq A$。标识符表示可以唯一标志某一个个体身份的属性，如身份证号码等。

定义 4.5：准标识符 QI（Quasi Identifier）。准标志符表示潜在的可以标识某一个个体身份的一组属性，属性的个数大于等于 2，$QI=\{qi_1,qi_2\cdots,qi_m\},m\ge 2$，其中 $\{qi_1,qi_2\cdots,qi_m\}\subseteq\{A_1,A_2,\cdots,A_m\}$。比如种族、出生日期、性别和邮编组成的属性组。

定义 4.6：敏感属性 S，$S\subseteq A$。表示个体不希望让其他人知道的属性，如工资、疾病、年龄等。

定义 4.7：时间 T，$T\subseteq A$。表示对应的时间操作时间，比如发布微博时的时间。

定义 4.8：位置 L，$L\subseteq A$。表示该记录对应的地点，比如发布微博时所在的位置。

根据前面定义的数据表中的属性，我们可以将数据表中的每一条记录表示为一个五元组，表中的每条记录可以拥有全部的属性，也可以只具有一部分属性。

定义 4.9：数据表 DB（Database）。$DB=(I,QI,S,T,L)$，则一个数据库中的全部数据可以表示为 $\{db_1,db_2,\cdots,db_m\}$。其中 m 为数据表中记录总数。同时用 $db_i[A_j]$ 表示第 i 条记录的第 j 个属性，并用 $db_i[QI]$ 表示第 i 条记录对应的准标志

符属性组。

定义 4.10：泛化格。如果同时对多个属性的准标志符进行 泛化操作，就会形成不同等级的泛化系列，这些准标志符形成的泛化等级系列，就是泛化格。

4.2.1 动态 *k*-匿名隐私保护方法

对于数据表中每条记录 $db_i(0 \le i \le m)$ ，如果该条记录含有准标识符，那么就对该准标志符组采用 k -匿名规则进行泛化处理。同时，根据访问控制模型得到的安全等级，记为 d ，对于每一个 d 值，通过映射函数 $F: d \to k$ 都能唯一的对应于某一个 k 值。使得用户可以动态的获取不同匿名程度（不同 k 值）的隐私数据。

动态 k -匿名又可以称作个性化 k -匿名。为了支持动态 k -匿名方法，本文在传统 k -匿名的基础上引入聚类相似块的概念，类似于动态规划算法中的最优子结构，在不同 k 值的计算过程中，如果他们具有相同的子集，则直接采用子集的计算结果，从而为动态 k -匿名的实现提供支撑。例如对于不同的 k 值下的泛化，记两个 k 值分别为为 $k_1, k_2(k_1 = k_2 + 1)$，对于准标志符 $\{S, Z, B\}$，其中 S, Z, B 满足 $S \in QI \&\& Z \in QI \&\& B \in QI$，如果准标志符在 k_2 对应的泛化后的表示为 $\{S^*, Z^*, B^*\}$，那么在计算 k_1 对应的泛化结果时，不需要从头开始遍历层次树，直接从 $\{S^*, Z^*, B^*\}$ 对应的层次开始遍历即可。

本文在 Incognito 算法的基础上[58]，泛化算法的主要步骤如下：

1、定义各个准标志符 QI 的泛化层次结构。例如对于数据表 4.1，其层次结构定义如图 4.6 所示：

表 4.1 原始数据

编号	出生日期	性别	邮编	疾病
1	1976-01-21	男	53715	Flu
2	1986-04-17	女	53715	Hepatitis
3	1976-01-21	男	53710	Brochitis
4	1976-01-15	男	53703	Broken Arm
5	1986-04-13	女	53706	Sprained Ankle
6	1976-04-19	女	53706	Hang Nail

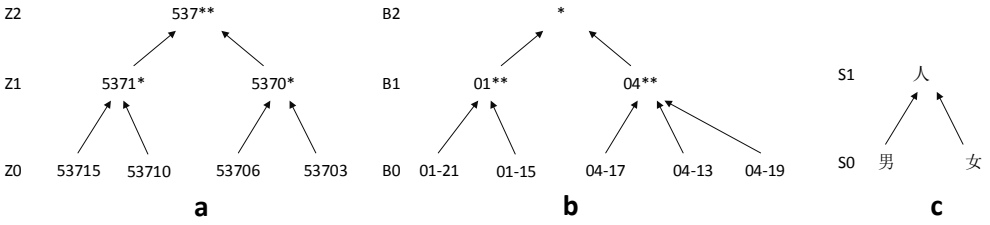


图 4.6 泛化等级

2、对于需要泛化的属性组定义泛化格，泛化格中的每个节点都是一个数据表，代表着各个属性的泛化。例如对于图 4.6 中的$<Zip, Birth, Sex>$组成的泛化格如图 4.7 所示。例如对于节点$<Z1, B1, S0>$对应着泛化数据表如表 4.1（a）所示，而$<Z1, B2, S0>$则对应着数据表如表 4.2（b）所示。

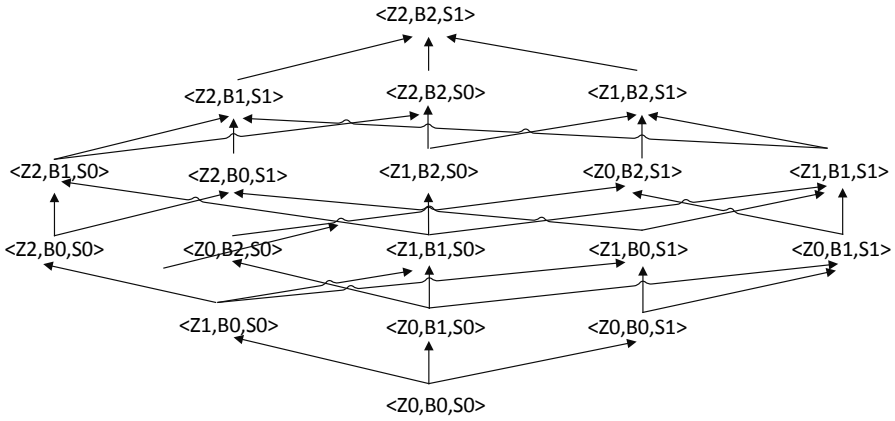


图 4.7 $<Zip, Birth, Sex>$对应的泛化格

表 4.2 泛化数据表举例

日期	性别	邮编	疾病
01**	男	5371*	Flu
04**	女	5371*	Hepatitis
01**	男	5371*	Brochitis
01**	男	5370*	Broken Arm
04**	女	5370*	Sprained Ankle
04**	女	5370*	Hang Nail

a

日期	性别	邮编	疾病
01**	男	537**	Flu
04**	女	537**	Hepatitis
01**	男	537**	Brochitis
01**	男	537**	Broken Arm
04**	女	537**	Sprained Ankle
04**	女	537**	Hang Nail

b

3、遍历各个属性各个泛化的组合，找出所有可以是数据表满足k-匿名性质的；

4、从找出的所有泛化步骤中，选出最有的一组泛化步骤作为最终的k-匿名解决方案。

5、对于不同的k值，重复执行步骤 3、4，并在重复过程中充分利用之前计

算得到的最优子结构，减少重复计算。

算法 4.1 动态 k -匿名隐私保护算法

输入：数据表 DB，含有 n 个属性的准标志符 Q，维度表（泛化层次结构，Q 中的每个属性都对应着一个维度表），k （k -匿名系数）

输出： 符合 k -匿名的全域泛化表 DB

1. C_1 = {Q 中所有属性泛化树的所有节点}
2. E_1 = {Q 中所有属性泛化树的所有边}
3. Initial *queue as empty*
4. **For** $i = 1$ to n **do**
5. $S_i = copy\ of\ C_i$
6. **If** $k-1$ is completed **then**
7. 删除 C_i 中所有可以到达 $k-1$ 对应泛化结果的泛化格中的点
8. Initial Tree as *roots*, {*roots*}={所有节点属于 C_i 且没有任何属于 E_i 的边指向这些节点 }
9. **else**
10. Initial Tree as *roots*, {*roots*}={所有节点属于 C_i 且没有任何属于 E_i 的边指向这些节点 }
11. **End if**
12. 将{*roots*}中的所有节点插入 *queue*，并保持树按照高度排列
13. **while** *queue* is not empty **do**
14. *node* = first node from *queue*
15. **If** *node* is not marked **then**
16. **If** $node \in \{roots\}$ **then**
17. *frequencySet*=根据各属性计算 *DB* 的频率集
18. **else**
19. *frequencySet*=根据父频率集计算 *DB* 的频率集
20. **End if**
21. //**根据** *frequencySet 判断当前的计算是否满足 k-匿名性质*
22. **If** *DB* is k-anonymous with respect to attributes of *node* **then**

23. 标记 *node* 之上的所有泛化操作
24. **else**
25. 从 S_i 中删除节点 *node*
26. 将 *node* 之上的所有泛化操作节点加入 *queue*，并保持树按照高度排列
27. **End if**
28. **End if**
29. **End while**
30. $C_{i+1}, E_{i+1} = GraphGeneration(S_i, E_i)$
31. **End for**
32. **Return** *DB* 中满足 *k*-匿名的所有泛化过程

4.2.2 基于四叉树的动态位置 *k*-匿名方法

对于数据表中的位置信息 L 和时间 T，在传统的 *k*-匿名隐私保护方案后，攻击者仍可以轻松的通过用户当前的一些信息获取用户的敏感信息。例如图 4.8 所示，用户 u 在一段时间内有连续的活动，虽然经过 k 匿名处理后，攻击者无法确切的知道用户的具体位置，但是却可以通过一系列查询，得到用户的一系列模糊位置 $(t_i, L_i)\{0 \le i \le m\}$。攻击者以这些位置为点，不但可以画出用户的一个大概活动轨迹，如图 4.8 张的红色线条所示，还可以结合背景知识攻击（例如图中的小区、办公区等基本信息），得到具体的个体以及对应的敏感信息。

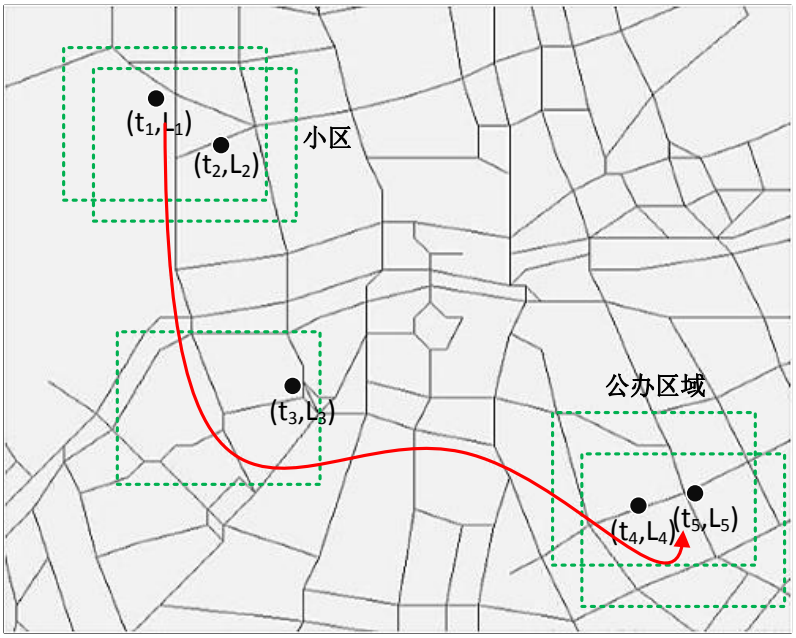


图 4.8 多次位置发布形成位置轨迹

定义 4.11：位置 k -匿名。当一个用户的位置信息即将被 LBS 提供商提供给其他使用者时，必须确保其位置信息与其他至少 $k-1$ 个用户的位置信息是不可区分的，使得攻击者无法确定发出该请求的用户的个体身份。

定义 4.12：动态位置 k -匿名。又称个性化 k -匿名，指在定义 4.11 的基础上，可以在不同的 k 值条件下，动态的为用户返回不同泛化程度的数据。

针对动态位置 k -匿名，本文在第三章的研究基础上，利用访问控制模型获得的访问等级 d，通过二级缓存原理，在保证计算复杂度的情况下，动态的为用户提供不同泛化程度的位置信息。在保证用户隐私数据的同时为用户提供更好的服务。用定义 4.1 中的位置偏差 $\delta(x,x')$ 来说，也就是使得合法用户在合法的行为下获得的平均位置偏差 $average(\delta)$ 最小。其中 $average(\delta)$ 的值如公式 4.2 所示，m 为请求的总次数。

$$average(\delta)=\sqrt{\sum_{i=0}^{m}(\delta(x,x'))^2} \tag{4.2}$$

本节在定义 4.12 的基础上，提出了基于四叉树的动态位置 k -匿名方法，通过在原有的时空匿名方法基础上，引入用户隐私需求，解决用户稀疏情况下匿名区域过大、用户稠密情况下位置信息泄露等问题，并通过隐私度量标准来判断位置隐私保护效果。

定义 4.13：用户隐私需求 pr （*privacy requirements*）。表示用户的位置信息被服务商公布出去时应该得到的隐私保护力度。可以用一个五元组表示，具体如公式所示。

$$pr=\{k,A_{min},A_{max},T_{min},T_{max}\} \tag{4.3}$$

k 为匿名度，即计算所得的时空匿名区中至少包含 k 个用户。显然，k 值越大，匿名区中的用户就越多，被攻击者识别的概率也就越低，隐私保护的力度也就越强。

A_{min} 为时空匿名区中位置平面的最小面积。当用户较为稠密，甚至极端情况下有大量用户处于同一个点，就需要通过 A_{min} 来对位置匿名区域进行限制。保证基本的匿名区不会暴露用户的位置信息。

A_{max} 时空匿名区中位置平面的最大面积。当四叉树的空间划分中，如果一直没有找到符合要求的设定，算法将不断扩大位置范围以查找满足用户设定参数的匿名区，当超过用户设定的匿名区最大面积仍未找到符合用户隐私要求的结果时，表明位置匿名失败。

T_{min} 为时空匿名区中时间的最小跨度。当有大量数据表记录发生在同一时间点时，就需要通过 T_{min} 来对时间范围进行限制。

T_{max} 为时空匿名区中时间的最大跨度。在算法在没有找到符合要求的匿名区域情况下将会一直扩大匿名区范围，如果达到 T_{max} 时仍没有找到匿名区，则匿名失败。

通过用户的隐私需求 pr，可以得到最大的匿名空间为 $A_{max} * T_{max}$，最小的匿名空间为 $A_{min} * T_{min}$。

定义 4.14：访问等级到隐私需求映射函数 $F : d \rightarrow pr$。用于将访问控制模型给出的当前安全访问等级映射到具体的用户隐私需求上。

定义 4.15：匿名评价模型 *aem*（anonymous evaluation model），根据文献[59]给出的 k-匿名方法评价标准，给出基于四叉树的动态位置 k-匿名方法的评价准则，综合考虑了服务质量（QOS，包括位置准确率、时间偏差）、性能消耗、匿名成功率等。具体如图所示。

$$aem = \{T, A, t, s\} \tag{4.4}$$

其中 T 为时间跨度，A 匿名区的位置平面面积，t 为计算时间，s 为匿名成功率。

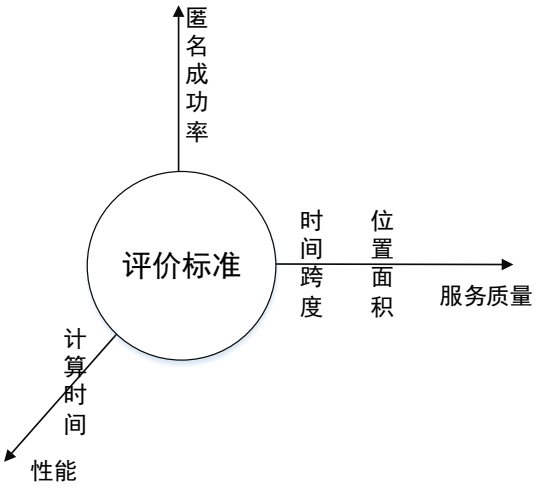


图 4.9 匿名评价模型

本章的动态位置 k-匿名方法基于四叉树模型来进行空间的划分，每次匿名操作时，不断的对匿名空间用四叉树递归划分，直至某一空间不满足用户的隐私需求 pr，则返回其上一级的子空间作为用户的匿名区。同时，为了支持动态的位置 k-匿名，需要记下每个用户隐私需求对应的四叉树节点，当计算一个范围更大的隐私需求的时候，直接从当前记录的节点开始搜索，而不需要从最底层开始搜索。四叉树空间划分如图 4.10 所示。

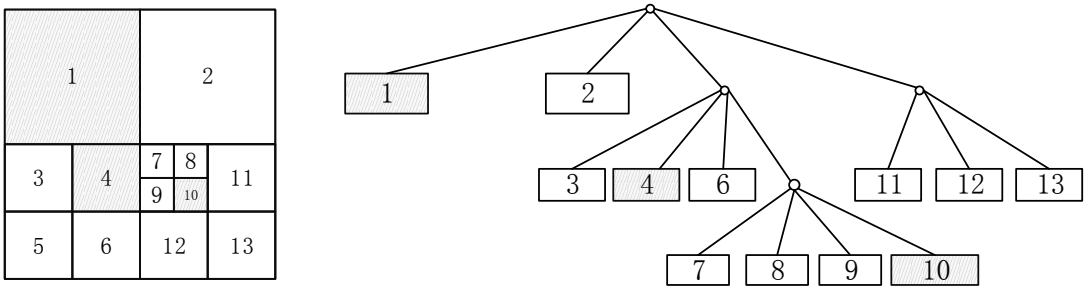


图 4.10 四叉树划分空间模型

定义 4.16：四叉树节点{*Node*}，*node* 中保存有该节点的节点 *id*、位置范围 $A_{position}$、位置面积 A、时间跨度 T 以及用户数量 N 。同时{*Node*}还包含有五个指针，分别是 *parent*，*firstChild*，*secondChild*，*thirdChild*，*forthChild*。

$$Node=(id,A_{position},A,T,N,firstChild,\text{sec}ondChild,thirdChild,forthChild) \quad (4.5)$$

其中，$A_{position}$ 可由四个变量来表示，分别是起始经度 *startLongitude*、起始纬度 *startLatitude*、截止经度 *endLongitude* 和截止纬度 *endLatitude*。

$$A_{position}=(startLongitude,startLatitude,endLongitude,endLatitude) \quad (4.6)$$

定义 4.17：最低满足隐私需求 $f(node\supset pr)$，表示当前节点表示的匿名区大于 *pr* 的最低匿名区要求。

$$f(node\supset pr)=node.A\geq pr.A_{min} \;\&\&\; node.T\geq pr.T_{min} \quad (4.7)$$

定义 4.18：最高满足隐私需求 $f(node\subset pr)$，表示当前节点表示的匿名区大于 *pr* 的最低匿名区要求。

$$f(node\subset pr)=node.A\leq pr.A_{max} \;\&\&\; node.T\leq pr.T_{max} \quad (4.8)$$

当节点 node 同时满足 $f(node\supset pr)$ 和 $f(node\subset pr)$ 时，*node* 对应的节点就是满足隐私需求 *pr* 的一个匿名区，但不一定成为最终的匿名区。

具体的求解步骤如下所示：

(1) 根据定义 4.14 计算当前用户操作对应的用户隐私需求 $pr(ub_i)$。同时设定系统允许最小的匿名面积 A^* 和匿名时间跨度 T^*。$pr(ub_i)$ 需要满足以下几点。

$$pr(ub_i)=\begin{cases}A_{min}\geq A^* \;\&\&\; A_{min}\bmod A^*=0\\ A_{max}\geq A_{min} \;\&\&\; A_{max}\bmod A^*=0\\ T_{min}\geq T^* \;\&\&\; T_{min}\bmod T^*=0\\ T_{max}\geq T_{min} \;\&\&\; T_{max}\bmod T^*=0\\ k\geq 2\end{cases} \quad (4.9)$$

(2) 以 $A*$ 和 $T*$ 为基本单位，初始化四叉树空间划分 $\{roots\}$ ，其中，通过 $A = width* height$（本文中设定 $width = height$，也就是可以用 A 逆向得到 $width$ 和 $height$）对平面空间进行降维，将由宽高代表的矩形区域由矩形面积进行表示。

(3) 遍历表 DB 中的各个位置记录 db_i，通过四叉树遍历，将每个位置划分到对应的节点上。

(4) 针对每一个 db_i 中的位置通过遍历四叉树来寻找匿名区，首先获取该位置对应的匿名节点，判断当前节点是否满足隐私需求，如果不满足，则寻找其父节点，直到找到满足隐私需求条件的节点，返回该节点。如果当前节点已经超出隐私需求的最高限制，则匿名失败。

算法 4.2 动态位置 k -匿名算法

输入： 数据表 DB，隐私需求 pr
输出： 符合 k -匿名的泛化表 DB

1. $\{roots\}$=Initial *quad tree*
2. Initial *stack* as empty
3. **Foreach** db_i in DB **do**
4. 将树的第一个节点$\{roots\}$插入 *stack*
5. *//将每个位置对应到四叉树中的节点*
6. **While** (stack is not empty)
7. *node = 获取 stack 的第一个节点*
8. **If** $db_i[L]$ in $node.A_{position}$ **then**
9. **If** $node.firstChild = null$ **then**
10. 更新当前 node 节点号为 $db_i[L]$ 对应的位置匿名区号
11. **Else**
12. put $node.firstChild, node.secondChild, node.thirdChild, node.forthChild$ into *stack*
13. **End if**
14. **End if**
15. **End while**
16. **End Foreach**

17. //获取每个位置对应的匿名区

18. **Foreach** db_i in DB **do**

19. $node = db_i[L]$对应的节点

20. **While** true **do**

21. **If** $f(node \supset pr)$ && $f(node \supset pr)$ **then**

22. 更新 $db_i[L]$对应的匿名区为该 $node$。

23. **Break**

24. **Else if** $node.A \geq pr.A_{max}$ || $node.T \geq pr.T_{max}$

25. 更新 $db_i[L]$对应的匿名区为 $fail$。

26. **Break**

27. Else

28. $node = node.parent$

29. **End if**

30. **End while**

31. **End Foreach**

32. **Return** 匿名化处理后的 DB

4.3 仿真实验与结果分析

4.3.1 实验环境与实验方法

实验采用 Java 实现，在 Intel(R) Core(TM) i5-3470 CPU @ 3.2GHz 3.20GHz、8G 运行内存的 Windows 8.1Pro 平台上运行，编译器为 Eclipse Juno(4.2)+JDK1.7。后台数据库为 MySQL5.0。

实验中采用[60]中的美国人口普查数据作为实验对象，数据中含有用户的个人敏感信息，实验时以 $\{age, Sex, occupation\}$ 为 QI 属性，$marital-status$ 为敏感属性，然后利用改进的 Incognito 算法（算法 4.1）来对数据实现动态的 k-匿名处理。然后增加 QI 的数量，研究 QI 数量与计算时间的关系。

然后利用在第二章处理后的微博数据以及社区物联网服务平台中的路况信

息，在地图上选取一个$2560m\times2560m$大小的区域，且设定单元格的宽度为$10m$，形成一个包含有256×256单元格的平面区域。然后获取区域四个顶点的经纬度值，遍历微博数据和路况信息数据，获取处于该区域的微博信息和路况信息，然后将每条信息的经纬度值转成该平面上对应的点。最后根据算法 4.2 对该区域中的位置信息进行位置动态k-匿名化处理。将经纬度转成的二维平面上的点分布如图 4.11 所示，实验中获取了两组数据，第一组中约含有 600 个位置点，第二组约含有 400 个位置点，既有一些看似无序的点，也有一些点聚集在一起，这些聚集点往往就是人群较为集中的地方。

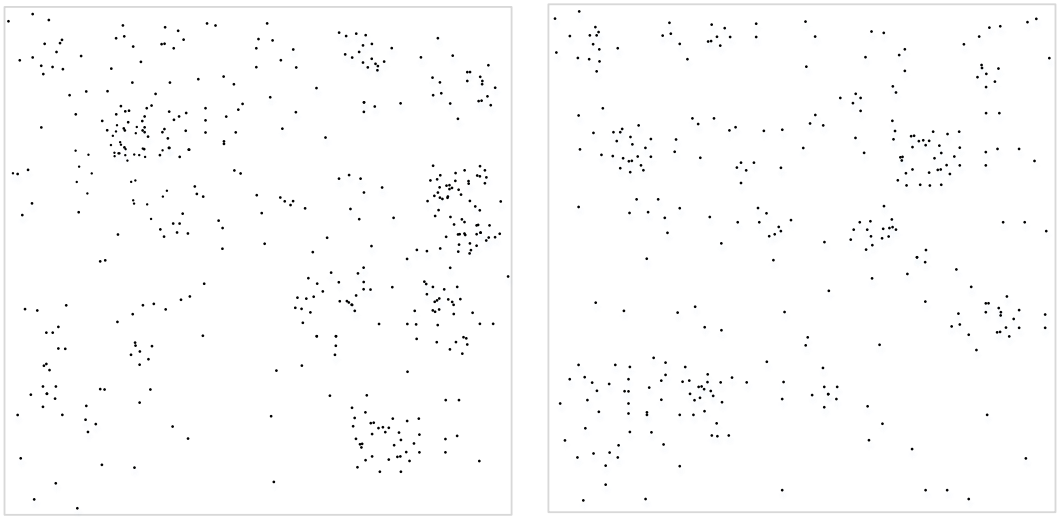
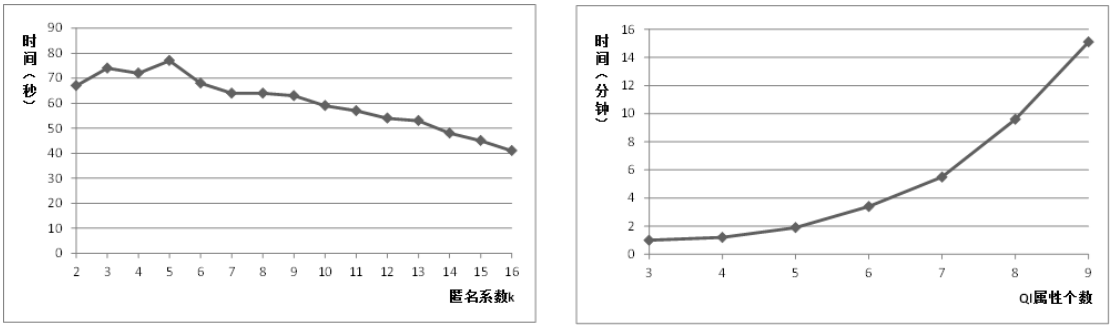


图 4.11 二维平面上的位置分布图

实验时，首先研究了匿名区域面积与匿名系数的关系，然后研究了匿名时间跨度与匿名系数的关系，其中，为了控制单一变量，在实验时设置$\{A_{min},A_{max},T_{min},T_{max}\}$为允许的最大或者最小值。最后研究了不同安全等级$d$与匿名失败之间的关系。

4.3.2 实验结果及分析

图 4.12（a）显示的是匿名系数k与计算时间的关系，其中QI属性的个数为 3。从图中可以看出，随着匿名系数k的增大，计算时间基本不会变化，甚至有一些下降，主要是因为随着最优子结构的加入，后面的计算可以依赖于前面的计算，从而使得算法的复杂度降低。图 49（b）显示的则是随着QI属性的增加，计算时间呈指数方式快速增加，其中$k=3$。因为随着QI属性的增加，其对应的泛化层次树的复杂度快速增加，遍历时所需要的时间也就越来越高。



a b

图 4.12 k、QI 与计算时间关系

针对移动互联网下的动态位置 k -匿名方案，不同的安全等级会映射到不同的隐私需求上，图 4.13（a）显示的是不同的匿名度 k 与匿名面积的关系，其中，匿名面积为一个平均值。从图中可以看出，随着 k 值的增加，不管是对于 600 个位置点的情况还是 400 个位置点的情况，所需的匿名面积不断增加，总体上呈现出线性的关系。图 4.13（b）显示的是不同的匿名度 k 与匿名时间跨度之间的关系，整体上也是呈现出线性的关系。另外，从图 4.13 中我们可以看出，不管是所占的面积比还是所占的时间比，在同样大小的区域内，拥有 600 个点的区域在相同的匿名系数下拥有更小的比例，这是因为在位置相对平均的环境中，拥有更多位置点的区域更容易在更小的匿名区内找到满足匿名要求的匿名区，而位置相对更小的区域则需要寻找更大的匿名区才能达到同样的需求。

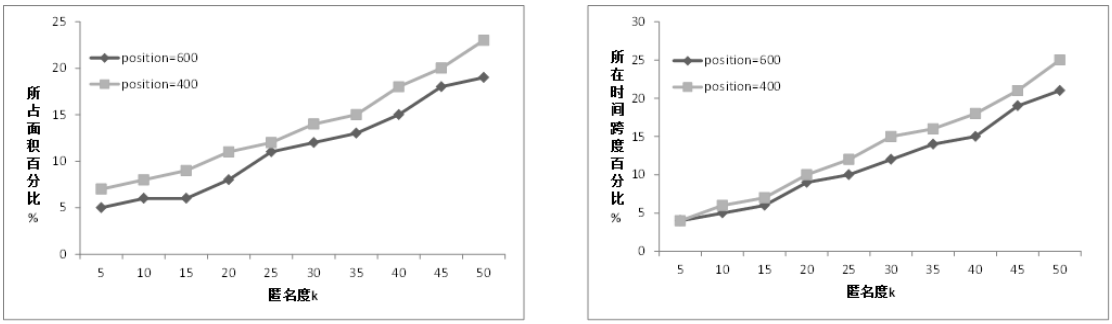


图 4.13 匿名系数 k 与匿名面积和匿名时间的关系

由于不同的安全等级会映射到不同的隐私需求，安全等级越高，对应的隐私需求就越小，而安全等级越低，对应的隐私需求就越高，就会导致匿名失败的概率变大，主要是随着时间与面积同步的增长，很容易出现算法中匿名失败情况。如图 4.14 所示的是安全等级 d 与匿名失败率之间的关系。

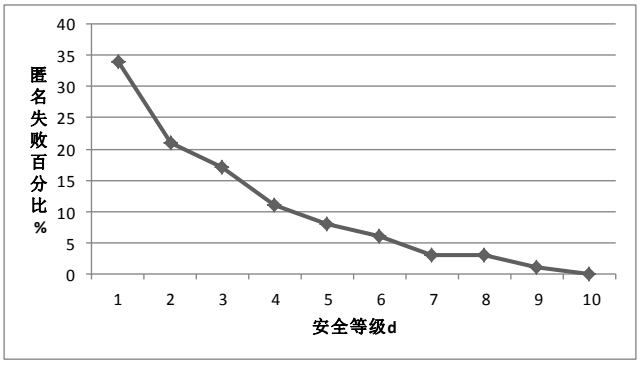


图 4.14 安全等级与匿名失败的关系

4.4 本章小结

本章在第三章的研究基础上，给出了移动互联网环境下基于用户行为访问控制模型的动态 k -匿名隐私保护方法，包括针对常规数据的动态 k -匿名隐私保护方法和针对位置数据的基于四叉树的动态位置 k -匿名方法。弥补了当前基于 k -匿名的隐私保护方法在移动环境下隐私保护力度不足以及数据失真等问题，并通过实验分析了各隐私参数对于匿名结果的影响与意义，验证了隐私保护方法的正确性。

第5章 基于用户行为的隐私保护机制的应用

本章将基于用户行为的隐私保护机制应用于社区物联网创新服务平台，首先介绍了社区互联网创新服务平台以及系统设计，接着给出了隐私保护机制在社区物联网创新服务平台中的隐私保护效果，通过应用效果来进一步验证隐私保护机制的有效性和准确性。

5.1 社区物联网创新服务平台概述

国家科技支撑项目“社区物联网创新技术与服务平台与应用示范（项目号：2012BAH15F03）”是以便民出行为背景，为社区用户提供一系列实时资讯和互动服务，包括基于位置的实时查询服务、基于位置的语音和视频服务等，集成周边医疗、餐馆、银行、停车场、商店等社区服务，实现及时便捷的预约、预订、信息咨询和通知等服务，从而减少排队等候时间，营造和谐生活综合应用，从而有机将虚拟世界和现实社会结合起来，创造良好舒适的用户体验，加速信息的传递，促进用户间的交流，为人们的生活带来了便利。系统的主要功能主要包括三个部分。

1. 基于地理位置的路况分享

基于地理位置的路况分享主要是每个用户都可以将自己当前看到的路况信息通过文本、图片、视频等形式并附上地理位置信息分享出去。该功能主要可以分为两大组成部分：客户端与服务器端。其中，客户端负责获取用户位置信息，照片拍摄以及与服务器端的数据通信，同时负责界面显示，并与用户进行交互。服务器端则负责接收用户提交的路况分享信息并存储到数据库中，同时向请求用户发送合适的路况信息。其基本模块关系如图 5.1 所示。

图 5.1 模块整体关系图

2. 基于位置的问答服务

当用户需要查询某一特定区域的的信息时，可以将自己的需求以及自己当前

的位置上传到服务器，服务器收到问题后，对问题进行简单的分析，然后将问题推送给符合要求、指定位置的用户。其他用户收到问题后就可进行回答并回传给服务器。这样提问用户就可以看到回答者的回答以及回答者的位置信息。

3. 基于位置的微博客发布平台

用户可以在该平台上像新浪微博一样，发布短博客、浏览他人发布的短博客。同时可以构建社交关系，查看其他人的信息，关注其他人等。

社区物联网创新服务平台基于服务器/客户端架构，系统分为两个逻辑模块：服务器端和客户端模块。其中服务器主要分成三层结构，分别是最里层的数据库服务器（方便做到数据隔离）、中间的计算集群、与客户端交互的 Web 服务器，包含的功能有用户管理，数据管理，资源管理，推荐系统，位置服务系统等，并提供服务接口。客户端有用户注册登陆，用户信息管理，好友关系管理，用户位置获取，用户位置更新，信息发布，信息浏览等功能。系统架构如图 5.2 所示。

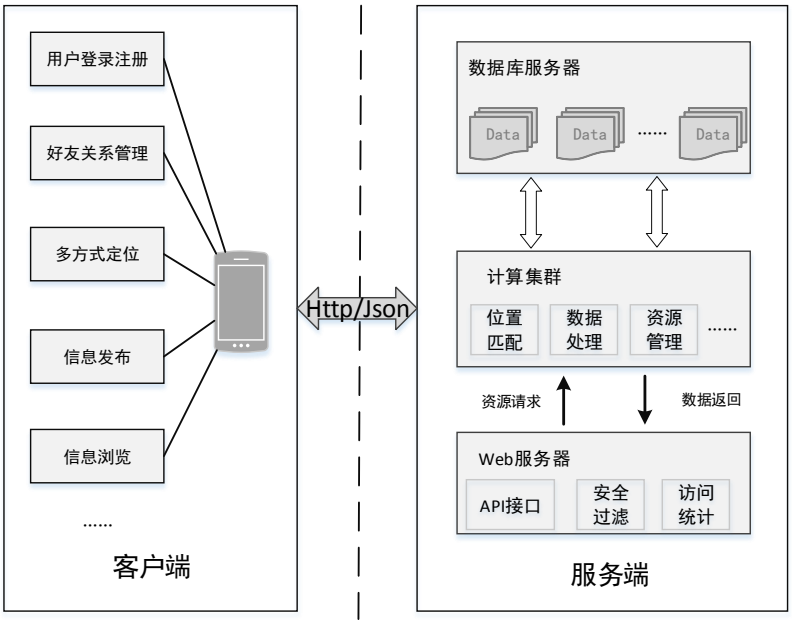


图 5.2 系统结构

5.2 隐私保护系统设计与实现

根据移动互联网环境的特点，并在前面章节的研究基础上，本章设计了一套隐私保护系统，并将该隐私保护系统应用于物联网创新服务平台。本隐私保护系统主要分成两部分，一部分是手机客户端，主要是与用户进行交互，获取用户相关的信息，并将对应的信息展现给用户。另一部分则是服务端，主要是进行隐私保护系统的部署，包括用户行为分析、 k -匿名的实现等等。具体的系统结构如

图 5.3 所示。

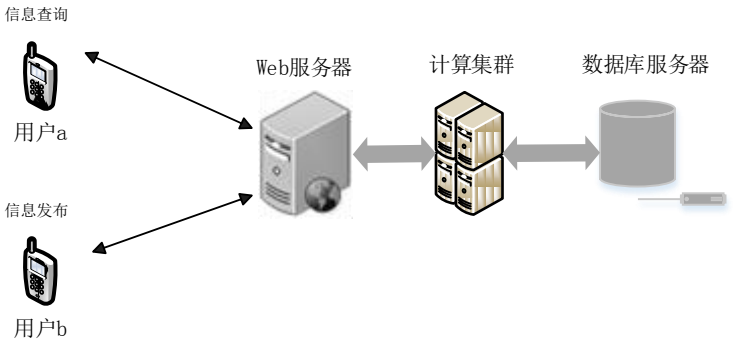


图 5.3 隐私保护系统

5.2.1 服务端设计与实现

隐私保护系统的隐私保护模块部署于服务端，也就是由服务提供商提供，系统结构如图 5.4 所示。和物联网创新服务平台一样，该套隐私保护系统包含三部分，分别是 Web 服务器、计算机群和数据库服务器。其中，Web 服务器主要负责用户操作行为获取，并进行简单的行为特征提取，另外在作为与客户端唯一交互的服务器，同时具备有访问控制的功能，对于已发现的攻击者行为，可以直接给予拒绝访问。计算集群主要是负责各种复杂的计算，包括模型的构建与训练、模型更新、用户当前操作安全等级计算、动态 k-匿名隐私数据泛化、k-匿名位置匿名区计算等。数据库服务器则用于存储各类数据，包括用于操作行为特征库、用户信任值、匿名数据缓存等。

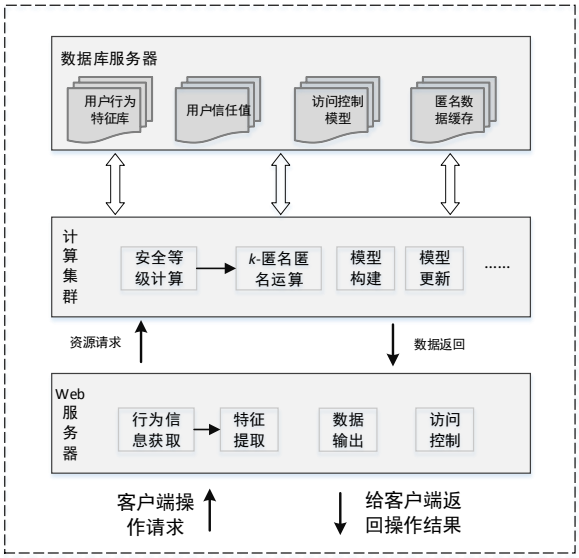


图 5.4 隐私保护系统结构

对于用户的一个操作请求，详细流程如图 5.5 所示，首先是 Web 服务器接收到用户的操作行为，进行简单的行为特征提取；然后将提取到的行为特征以及对应请求传给计算集群，计算集群通过已有的访问控制模型计算出当前操作行为的安全等级。并通过映射函数 $F:d \rightarrow pr$ 将当前的安全等级映射为用户隐私需求。如果对应隐私需求的数据已经存在且时间较新，则直接回来对应数据，否则就通过该隐私需求去计算对应的泛化数据。最后将对应数据反馈给 Web 服务器，Web 服务器将数据输出给用户。

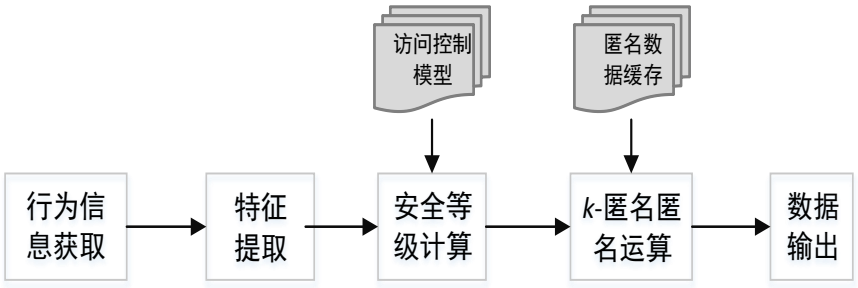


图 5.5 隐私保护系统流程

5.2.2 手机客户端的设计与实现

客户端则是对于整个隐私保护系统效果的呈现，客户端基于安卓框架开发，主要包括用户登录注册、好友关系、路况信息发布与分享、语音问答模块、周边信息查询、微博客等。客户端系统结构如图 5.6 所示。

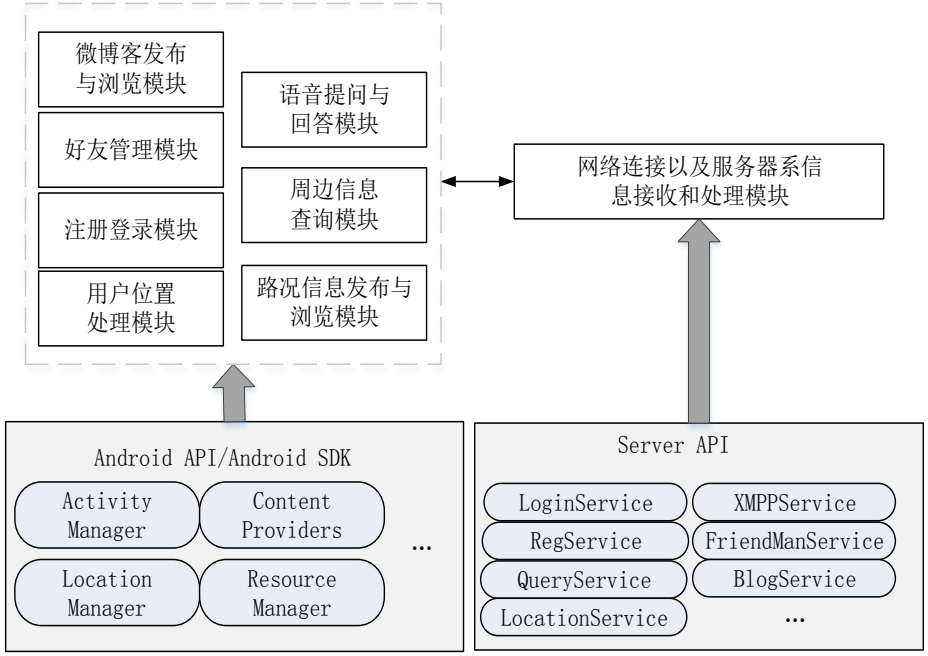


图 5.6 客户端系统结构

客户端基于安卓 SDK-17（安卓 4.4）开发，采用经典的 MVC 框架，通过对原有的应用进行升级，客户端每次对服务器发起请求时都会同步上传自己当前的终端 IMEI（International Mobile Equipment Identity，移动设备国际识别码，又称为国际移动设备标识，手机的唯一识别号码）以及自己当前的位置，具体的功能升级点如图 5.7 所示。

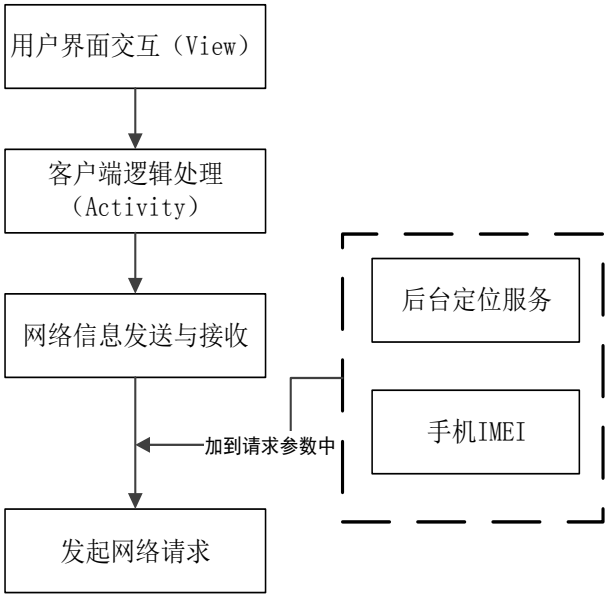


图 5.7 客户端升级功能点

5.3 应用效果展示与分析

社区物联网创新服务平台通过长期的投入已拥有了一定的用户数据。通过对这些数据进行处理、训练，足以构建了移动互联网下基于用户行为的访问控制模型。并通过在服务端的部署，实现了在数据向用户发布时根据用户当前的安全等级动态提供不同泛化程度的数据。具体的一些应用效果如下所示。

1. 用户登录

图 5.8 和图 5.9 分别显示的是使用常用终端进行登陆和使用新终端进行登陆的区别，图中显示，用户在常用终端（图 5.8）上连续密码错误 5 次时将会被暂时禁止登陆，而如果在新的终端（图 5.9）上连续输错密码 3 次就会被暂时禁止登陆。

图 5.8 常用终端登陆

图 5.9 非常用终端登陆

每个用户终端都有唯一的标志 IMEI，用户在常用的终端上登陆经过长期的训练更新成为一种正常的操作行为，突然在新的终端上登陆，访问控制模型会认为这是一个较为危险的操作，甚至有可能是利用机器在进行暴力密码破解，所以输入错误的次数会比在正常的终端上少。

2. 路况信息查看

图 5.10 和图 5.11 显示的是两个用户 legolas 和 jame 在登陆系统后在同一地点看到的周边路况信息的区别，其中 legolas 是位资深用户，而且经过了实名认证，而 jame 则是一个普通用户，所以在同样的位置，legolas 看到的位置更加精确，而 jame 看到的则相对更宽泛。

点击某一个路况信息，就可以看到详细的路况信息，包括图片、文字描述、地点等。从图 5.12 和图 5.13 中可以看到，两人在看同一个含有视频信息的路况信息时，两人看到的路况描述、大体发生位置都是一样的，唯一不同的就是和自己的距离，legolas 看到的距离更加详细，精确到 1.7 公里，而 jame 看到的位置则比较笼统，只显示大约为 2 公里。

图 5.10 legolas 看到的路况信息

图 5.11 jame 看到的路况信息

图 5.12 legolas 看到的路况详情

图 5.13 jame 看到的路况详情

对于同样的操作，由于不同用户的用户信任值不一样，所以最终得到的当前行为安全等级也会不一样。这就是 legolas 和 jame 看到的内容不一样的原因，而对于不会暴露用户隐私的数据，系统则是毫无失真的呈现看用户，所以 legolas 和 jame 看到的路况详细介绍是一致的。

3. 查看他人信息

图 5.14 和图 5.15 显示的是同一个用户在常用地点查看陌生人信息和在非常用地点查看陌生人信息的区别，可以看到，在常用地点基本上能看到的其他人信息更加的准确，而在非常用地点，看到的陌生人信息则被进行了匿名化处理。

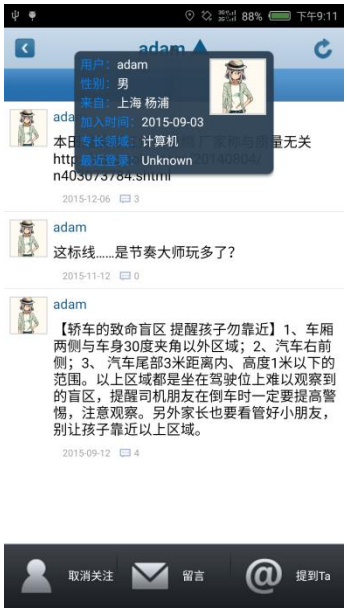


图 5.14 常用地点看到的用户信息

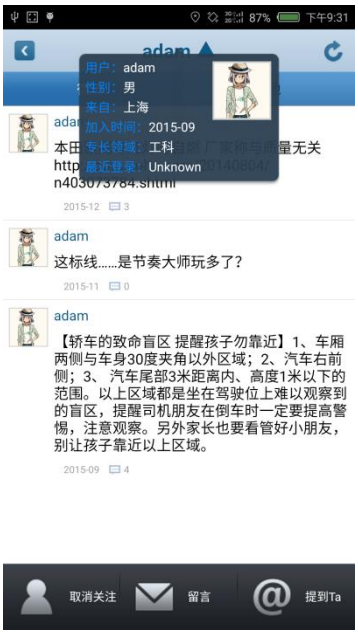


图 5.15 非常用地点看到的用户信息

相同用户在不同的位置所具备的安全等级是不一样的，所以看到的信息也会有区别，对于不会泄露隐私的信息可以随意查看，而对于可能泄露用户隐私的敏感数据，则可能会被经过一定的匿名处理，正如图 5.15 中所看到的。

5.4 本章小结

本章将基于用户行为的隐私保护机制应用于社区物联网创新服务平台。详细介绍了社区物联网服务平台以及系统设计，并展示了基于用户行为的隐私保护机制在该平台的应用效果。通过实际的效果展示，进一步验证了隐私保护机制的有效性和准确性，体现了隐私保护机制的实际应用价值。

第6章 总结与展望

6.1 总结

本文针对当前隐私保护方法不能很好的为移动互联网环境下的用户隐私数据提供隐私保护以及数据失真的问题，结合移动互联网环境下用户行为的特点，提出了移动互联网中基于用户行为的隐私保护机制。首先，通过对移动互联网环境下用户行为的分析、特征提取，给出了移动互联网环境中基于用户行为的访问控制模型，同时给出了模型的构建、训练和更新方法；然后，通过基于用户行为的访问控制模型获得访问安全等级，利用映射函数获得用户隐私需求，给出基于用户行为访问控制模型的动态 k -匿名隐私保护方法；最后，设计并实现应用于社区物联网创新服务平台的隐私保护系统，从而为用户提供更好的隐私保护，减少数据失真，提高移动互联网的服务质量，展现出较好的实际应用价值。也为后续移动互联网环境下的隐私保护机制研究提供了一个新思路。本文的主要工作具体包括：

(1) 研究了当前隐私保护的基本概念、研究思路、研究现状以及关键技术点，深入剖析了当前隐私保护方法在移动互联网环境下面临的问题。阐述了移动互联网环境下用户行为信息结构，主要包括用户信息、好友关系、行为信息等，给出了分布式的用户行为信息采集方案，并对采集到的数据进行预处理和特征提取，包括基于 *k-means* 的位置特征提取和基于 *KNN* 的操作特征提取。最后通过仿真实验验证了特征提取方法的准确性，从而为移动互联网环境下基于用户行为的隐私保护机制的研究提供数据保障。

(2) 提出了移动互联网环境下基于用户行为的访问控制模型，首先给出了模型的相关定义；然后通过已有的数据对模型进行构建与训练，包括用户初始信任值计算、相似性计算、特征数据库初始化等；再接着给出了模型的更新方法，包括基于记忆遗忘曲线的用户信任值更新以及基于遗传算法的特征数据库更新；最后通过实验验证了模型的有效性以及准确性。

(3) 通过基于用户行为的访问控制模型获得访问安全等级，利用映射函数获得用户隐私需求，给出基于用户行为访问控制模型的动态 k-匿名隐私保护方法，包括动态 k -匿名方法以及基于四叉树的动态位置 k -匿名方法，针对不同的隐私需求给出不同泛化程度的隐私数据。最后通过实验验证了算法的正确性。

(4) 依托于国家科技支撑计划项目，设计并实现应用于社区物联网创新服务

平台的隐私保护系统，从而进一步验证了基于用户行为的隐私保护机制的正确性和合理性。

6.2 展望

本文结合移动互联网下用户行为的移动性、位置性，针对传统隐私保护方法在移动环境下的不足，给出了移动互联网中基于用户行为的隐私保护机制，并通过实验以及实际应用验证了该隐私保护机制的有效性和准确性，为以后移动互联网环境下的隐私保护研究及应用起到了一定的积极作用。但随着移动互联网的快速发展，今后还将在以下几个方面开展后续研究：

(1) 随着分布式计算、大数据的到来，如何利用分布式计算快速的进行模型的训练与更新、利用更高效的数据挖掘方法进行数据的处理、并发的进行动态 k -匿名的计算将是接下去的研究重点。

(2) 当前的访问控制模型仍然需要一定的人工干预，如何利用更好的机器学习算法（比如深度学习）来实现全封闭的模型更新，使得模型可以在不需要任何人工干预的情况下不断适应外界环境的变化是下一步的研究任务。

参考文献

[1] Agrawal R, Srikant R. Privacy-preserving data mining[C]//ACM Sigmod Record. ACM, 2000, 29(2): 439-450.

[2] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias[J]. Journal of the American Statistical Association, 1965, 60(309): 63-69.

[3] Muralidhar K, Sarathy R. Security of random data perturbation methods[J]. ACM Transactions on Database Systems (TODS), 1999, 24(4): 487-493.

[4] Kargupta H, Datta S, Wang Q, et al. On the privacy preserving properties of random data perturbation techniques[C]//Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003: 99-106.

[5] Oliveira S R M, Zaiane O R. Privacy preserving clustering by data transformation[J]. Journal of Information and Data Management, 2010, 1(1): 37.

[6] Iyengar V S. Transforming data to satisfy privacy constraints[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 279-288.

[7] 华蓓, 钟诚. 数据挖掘中的隐私保护技术进展分析[J]. 微电子学与计算机, 2009, 26(8): 38-41.

[8] Goldreich O. Secure multi-party computation[J]. Manuscript. Preliminary version, 1998.

[9] Du W, Atallah M J. Secure multi-party computation problems and their applications: a review and open problems[C]//Proceedings of the 2001 workshop on New security paradigms. ACM, 2001: 13-22.

[10] Yang X C, Liu X Y, Wang B, et al. K-anonymization approaches for supporting multiple constraints[J]. Ruan Jian Xue Bao(Journal of Software), 2006, 17(5): 1222-1231.

[11] Fan S, Truong S. Access control for networks: U.S. Patent 6,219,706[P]. 2001-4-17.

[12] Escamilla T. Intrusion detection: network security beyond the firewall[M]. John Wiley & Sons, Inc., 1998.

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

[14] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity[C]//Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005: 49-60.

[15] 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.

[16] Li J, Tao Y, Xiao X. Preservation of proximity privacy in publishing numerical sensitive data[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008: 473-486.

[17] He X. Data Privacy Preservation for Dynamic Numerical Sensitive Attributes[J]. Jisuanji Kexue yu Tansuo, 2011, 5(8): 740-750.

[18] 吴英杰, 倪巍伟, 张柏礼, 等. k-APPRP: 一种基于划分的增量数据重发布隐私保护 k-匿名算法[J]. 小型微型计算机系统, 2009 (8): 1581-1587.

[19] Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking[C]//Proceedings of the 1st international conference on Mobile systems, applications and services. ACM, 2003: 31-42.

[20] Xiao X, Tao Y. Personalized privacy preservation[C]//Proceedings of the 2006 ACM SIGMOD international conference on Management of data. ACM, 2006: 229-240.

[21] LIU M, YE X. Personalized K-anonymity[J]. Computer Engineering and Design, 2008, 2: 007.

[22] 魏彦鹏. "基于移动社会网络的用户行为分析." (2014).

[23] 马力, 焦李成, and 董富强. "一种 Internet 的网络用户行为分析方法的研究." 微电子学与计算机 22.7 (2005): 124-126.

[24] Whang L S M, Lee S, Chang G. Internet over-users' psychological profiles: a behavior sampling analysis on internet addiction[J]. CyberPsychology & Behavior, 2003, 6(2): 143-150.

[25] 余慧佳, 刘奕群, 张敏, 等. 基于大规模日志分析的搜索引擎用户行为分析[J]. 中文信息学报, 2007, 21(1): 109-114.

[26] 王微微, 夏秀峰, 李晓明. 一种基于用户行为的兴趣度模型[J]. 计算机工程与应用, 2012, 8.

[27] Ji T G, Tian L Q, Hu Z X, et al. AHP-based user behavior evaluation method in trustworthy network[J]. Jisuanji Gongcheng yu Yingyong(Computer Engineering and Applications), 2007, 42(19): 123-126.

[28] 连一峰, 戴英侠. 基于模式挖掘的用户行为异常检测[J]. 计算机学报, 2002, 25(3): 325-330.

[29] Sandhu R S, Samarati P. Access control: principle and practice[J]. Communications Magazine, IEEE, 1994, 32(9): 40-48.

[30] Sandhu R. Rationale for the RBAC96 family of access control models[C]//Proceedings of the first ACM Workshop on Role-based access control. ACM, 1996: 9.

[31] Sandhu R S, Coyne E J, Feinstein H L, et al. Role-based access control models[J]. Computer, 1996, 29(2): 38-47.

[32] Sandhu R, Bhamidipati V, Munawer Q. The ARBAC97 model for role-based administration of roles[J]. ACM Transactions on Information and System Security (TISSEC), 1999, 2(1): 105-135.

[33] Shafiq B, Joshi J B D, Bertino E, et al. Secure interoperation in a multidomain environment employing RBAC policies[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(11): 1557-1577.

[34] Chang N Z, Yang C. An object-oriented RBAC model for distributed system[C]//Software Architecture, 2001. Proceedings. Working IEEE/IFIP Conference on. IEEE, 2001: 24-32.

[35] Takabi H, Joshi J B D, Ahn G J. Securecloud: Towards a comprehensive security framework for cloud computing environments[C]//Computer Software and Applications Conference Workshops (COMPSACW), 2010 IEEE 34th Annual. IEEE, 2010: 393-398.

[36] Pereira A L. RBAC for high performance computing systems integration in grid computing and cloud computing[C]//Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on. IEEE, 2011: 914-921.

[37] Joshi J B D, Bertino E, Latif U, et al. A generalized temporal role-based access control model[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(1): 4-23.

[38] Ray I, Kumar M, Yu L. LRBAC: a location-aware role-based access control model[M]//Information Systems Security. Springer Berlin Heidelberg, 2006: 147-161.

[39] Yang Y, Ding R, Min Y. Object-based access control model[J]. Automation of Electric Power Systems, 2003, 27(7): 36-40.

[40] Thomas R K, Sandhu R S. Task-based authorization controls (TBAC): A family of models for active and enterprise-oriented authorization management[J]. DBSec, 1997, 113: 166-181.

[41] Bertino E, Bonatti P A, Ferrari E. TRBAC: A temporal role-based access control model[J]. ACM Transactions on Information and System Security (TISSEC), 2001, 4(3): 191-233.

[42] Yuan E, Tong J. Attributed based access control (ABAC) for web services[C]//Web Services, 2005. ICWS 2005. Proceedings. 2005 IEEE International Conference on. IEEE, 2005.

[43] 李凤华, 王巍, 马建峰, 等. 基于行为的访问控制模型及其行为管理[J]. 电子学报, 2008, 36(10): 1881-1890.

[44] Downs D D, Rub J R, Kung K C, et al. Issues in discretionary access control[C]//Security and Privacy, 1985 IEEE Symposium on. IEEE, 1985: 208-208.

[45] Lindqvist H. Mandatory access control[J]. Master's Thesis in Computing Science, Umea University, Department of Computing Science, SE-901, 2006, 87.

[46] 邓集波, 洪帆. 基于任务的访问控制模型[J]. 软件学报, 2003, 14(1): 76-82.

[47] 李晓峰, 冯登国, 陈朝武, 等. 基于属性的访问控制模型[J]. 通信学报, 2008, 29(4): 90-98.

[48] Bhatti R, Bertino E, Ghafoor A. A trust-based context-aware access control model for web-services[J]. Distributed and Parallel Databases, 2005, 18(1): 83-105.

[49] Almenárez F, Marín A, Campo C, et al. TrustAC: Trust-based access control for pervasive devices[M]//Security in Pervasive Computing. Springer Berlin Heidelberg, 2005: 225-238.

[50] 廉捷, et al. "新浪微博数据挖掘方案." 清华大学学报: 自然科学版 51.10 (2011): 1300-1305.

[51] 刘明吉, 王秀峰, 黄亚楼. 数据挖掘中的数据预处理[J]. 计算机科学, 2000, 27(4): 54-57.

[52] 周奕辛. 数据清洗算法的研究与应用. Diss. 青岛: 青岛大学, 2005.

[53] 新浪微博商务部，企业微博助理，企业运营规律发博时间分析 http://wenku.baidu.com/link?url=8Dn_LplJR-f1NWtQ--4jP7W-DcZTqel0TR_jVKu8cKXDdX9N9FJvvZkaNfqSsnPH3wm0bcRXPNOJEbf6X3A1MXoOupzB_jQ92q75Vrtznuy

[54] Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge[C]//ICML. 2001, 1: 577-584.

[55] Peterson L E. K-nearest neighbor[J]. Scholarpedia, 2009, 4(2): 1883.

[56] 运筹学, 树栋, 遗传学. 遗传算法原理及应用[M]. 国防工业出版社, 1999.

[57] 皮佳明.基于用户兴趣变化的协同过滤推荐算法研究：[硕士学位论文].云南：云南财经大学，2014

[58] LeFevre K, DeWitt D J, Ramakrishnan R. Incognito: Efficient full-domain k-anonymity[C]//Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005: 49-60.

[59] 赵泽茂, 胡慧东, 张帆, 等. 圆形区域划分的 k. 匿名位置隐私保护方法[J]. 北京交通大学学报, 2013, 37(5): 13-18.

[60] Ronny Kohavi , Barry Becker. http://archive.ics.uci.edu/ml/datasets/Adult. 1996