**哈尔滨工业大学**

**硕士学位论文开题报告**

**题 目：面向大数据隐私保护中的数据库动态脱敏算法研究**

**院 （系） 计算机科学与技术**

**学 科 计算机科学与技术**

**导 师 何慧**

**研 究 生 郭睿**

**学 号 17S103173**

**开题报告日期 2018.09**

**研究生院制**

目录

[1. 课题来源及研究的背景和意义 3](#_Toc524100467)

[1.1 课题的来源 3](#_Toc524100468)

[1.2 课题研究的背景和意义 3](#_Toc524100469)

[2. 国内外在该方向的研究现状及分析 4](#_Toc524100470)

[2.1 国内外研究现状 4](#_Toc524100471)

[2.2 国内外文献综述的简析 6](#_Toc524100472)

[3. 主要研究内容 8](#_Toc524100473)

[4. 已完成的研究工作 10](#_Toc524100474)

[5. 研究方案及进度安排，预期达到的目标和取得的研究成果 10](#_Toc524100475)

[5.1 研究方案（不少于500字） 10](#_Toc524100476)

[5.1.1 敏感数据识别的研究 10](#_Toc524100477)

[5.1.2 动态脱敏算法及原则的研究 11](#_Toc524100478)

[5.2 预期达到的目标和取得的研究成果 11](#_Toc524100479)

[5.3 进度安排（建议从进入研究课题时间开始） 12](#_Toc524100480)

[6. 为完成课题已具备和所需的条件和经费 12](#_Toc524100481)

[7. 预计研究过程中可能遇到的困难和问题，以及解决的措施 12](#_Toc524100482)

[8. 主要参考文献 12](#_Toc524100483)

# 课题来源及研究的背景和意义

## 课题的来源

本课题来源于云安宝项目，项目基于DBShield，实现对数据库敏感数据的加密与脱敏。通过对数据库协议分析，解析相关字段，实现对数据库的访问控制，以及对敏感字段进行加密或脱敏。

## 课题研究的背景和意义

随着人类社会进入大数据的时代，人们的各类信息被政府部门、企业组织甚至个人收集[1]，如医疗档案、社会调查、人口普查和商业数据等。这些海量的数据信息被用于进行数据挖掘和机器学习等研究[2]，来帮助政府制定相关政策或者企业创造商业价值，同时也为人们提供更加丰富、智能和便捷的生活方式。但随着企业收集的数据越来越多，数据外泄事件一再发生，用户敏感数据遭到泄露，企业信息受到严重威胁，信誉遭受损失。为此，企业积极投资于数据隐私和数据安全技术，将不可预见的安全成本最小化并减少风险损失[3]。

在大数据时代的背景下，由于数据信息的内容形式以及收集、发布和传播的方式变得多样化，信息安全问题涉及的范围更加广泛，并且研究的内容也更加复杂和多样[4]。其中，人们的隐私安全问题一直是信息安全问题的研究核心之一[5]。在大数据技术的不断推动之下，各个行业和政府部门之间的信息壁障逐渐被打破。数据的交换和共享成为了信息交流中越来越重要的活动，但是，这些交流的数据中包含了大量个人隐私和敏感信息。如果这些数据在没有经过隐私保护处理之前就对外进行发布或者交换，会非常容易造成用户的隐私泄露[6]。例如，在2006年，著名的因特网服务提供商美国在线（ American Online）发布了一份 2GB的数据文件，其中包括了关于 65 万用户的约 2 千万个查询词条信息。尽管这份数据使用了随机数代替数据中用户的 ID，但是当数据发布之后，两名纽约时报的记者仍然通过数据中的信息准确地找到并且采访了其中的某位用户[7]。

为了保护发布数据中的用户隐私信息，学者们提出了一种被称为隐私保护数据发布（Privacy-Preserving Data Publishing）的技术[8]。该技术主要通过匿名或者添加噪音的方式对数据表中的信息进行一定程度的加密，将用户隐私暴露的概率降低在一定阈值以内，它可以帮助如社交网站、保险公司、医院和政府部门等在对外发布数据时，既不会泄露其中用户的隐私信息，还可以保证匿名数据的信息可利用性。

因此，对于产生和存储大规模数据的企业和组织而言，如何在提供正常服务的同时保证用户隐私信息不被泄露，在生产环境中结合隐私保护数据发布技术，动态地对用户敏感数据进行脱敏，在数据交换、共享及使用等过程中实现对敏感数据的定向、精准和彻底脱敏，达到数据安全、可信、受控使用的目标，是数据产生者和管理者亟待解决的技术问题。

# 国内外在该方向的研究现状及分析

## 国内外研究现状

目前，数据脱敏的主要实践者包括 IBM、ORACLE 和 Informatica。他们凭借在传统数据库行业较早的进入时间、较深厚的实践经验和技术积累，占据了多数市场份额。在国内，数据脱敏的研究和应用刚刚起步，银行、通信运营商根据自身需求制订了一些数据脱敏解决方案，多以静态脱敏为主，设计流程固定，工具能力有限，专用性较强，配置规则复杂，维护困难，不能满足数据交互流量的不断增长和复杂多变的安全处理需求。

学者提出了相应的脱敏原则，来保证脱敏后数据表的隐私安全。其中，匿名原则是指对数据表提出的匿名要求，当一个数据表满足了某一匿名原则的条件时，数据表就会具备相应的隐私保护能力。

k-anonymity 匿名原则[20]可以有效地防止数据表中用户身份的泄露；在介绍该原则前，需要先明确数据表中可能会出现的属性类型。

一般情况下，记录型数据表包含三种属性类型 [17]：（ 1）标识符属性（ ExplicitIdentifier），能够唯一或者在很大程度上识别用户身份的属性，并且在发布的数据中需要被移除；（ 2）准标识符属性（ Quasi-Identifier，简称 QI），作为用户的一般属性记录用户的非敏感信息，并且尽管单一项属性无法准确地识别用户的身份，但是当使用多项 QI 值作为条件在数据表中进行匹配时会有极大的概率辨认用户的真实身份；（ 3）敏感属性（ Sensitive Attribute），记录了用户的敏感信息。例如，姓名属于标识符属性， 年龄、性别和邮政编码属于QI 属性， 疾病属于敏感属性。

l-diversity 匿名原则要求数据表，对数据表中的任意个体及与其 QI 信息相同的所有个体，它们所携带的敏感属性值有l种不同的表现形式。

对称/非对称加密、同态加密、访问控制、安全审计和备份恢复等数据安全防护手段在用于敏感数据防护方面仍有欠缺，无法在不妨碍已有的数据处理、操作及分析过程的同时，实现对敏感数据的针对性保护。例如：传统加密技术由对称、非对称和散列算法构成，具有极高的安全强度，能够保证数据在传输过程中的机密性和完整性。但是，由于数据在使用时必须完全解密，对最终用户而言，敏感数据依然是明文，因而无法同时满足敏感数据安全性和可用性的需求。

数据脱敏可以划分为静态数据脱敏（ Static Data Masking，SDM）和动态数据脱敏（Dynamic Data Masking，DDM）。随着数据脱敏的应用领域从非生产系统拓展到生产系统，业界的技术需求也逐步从 SDM 过渡到 SDM/DDM 并重。DDM通常用于生产环境，在敏感数据被低权限个体访问时才对其进行脱敏，并能够根据策略执行相应的脱敏方法。SDM 与 DDM的区别在于，是否在使用敏感数据时才进行脱敏。

动态数据脱敏目前具有两类实现机制：基于视图的实现机制和基于代理的实现机制。项目采用的基于代理的实现机制，脱敏代理部署在数据容器的出口处以网关方式运行，检测并处理所有用户与服务器间的数据请求及响应。这种实现机制的好处是，无需对数据存储方式及应用程序代码做出任何更改。代理实现数据脱敏的具体方法是查询语句或响应语句替换。代理能自动识别目标为敏感数据的查询语句，并将语句改写为不包含敏感字段，或对敏感字段进行变换处理的查询语句。查询结果返回代理时，会被重新计算、修改并包装为与原请求一致的格式交付用户，从而完成一次敏感信息的查询过程，其原理图如图所示：

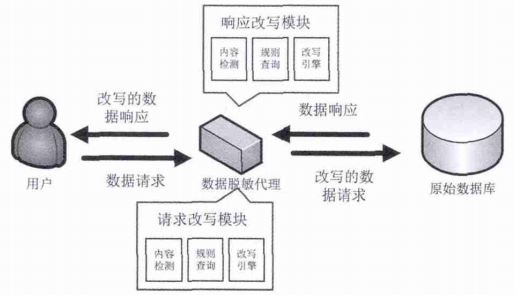


Figure 1 基于代理的动态脱敏实现机制

生产环境中数据的生成速度极快，脱敏过程必须能够在规则的引导下自动化进行，才能达到可用性要求。脱敏策略是在脱敏过程中贯彻的规则、规范、方法和限制的统称。脱敏规则是根据数据及用户的特点制订的全局和个别配置，用以指导脱敏过程的实现；脱敏规范是数据在处理中必须遵循的安全法规及行业标准；脱敏方法是对敏感数据进行具体变换操作的算法及流程；脱敏限制是应用脱敏方法时受到的条件和制约，如时空复杂度要求、时效性要求、接口要求等。

## 国内外文献综述的简析

不同原则的适用性简析：

1.k-Anonymity 可以避免身份泄露，这是因为被k匿名化的数据集中的记录无法被映射回与原始数据集相对应的记录。

缺点：k-Anonymity不能保护敏感属性的泄露。比如：具有相同QI属性的k个病人，都患有相同的疾病，那只要确定了某个具有类似QI值的病人身份，就能推断出患有该病。

2.p-Sensitive k-Anonymity 用于防止敏感属性的泄露，

定义：k>1 并且 p<=k,在满足k-Anonymity的前提下，具有相同QI值的记录至少有p种不同的敏感属性值。

缺点：该原则适用于不同敏感属性值出现频率相似的情况，如果不是这样，则会造成大量的数据丢失。比如：1000个病人中只有5个AIDS属性值为yes，其余为no，现在需要满足2-Sensitive k-Anonymity原则，那么最多只能划分为QI值相同的5组记录，这就意味着QI值需要进行粗粒度的泛化，从而造成信息的丢失。

3.l-Diversity 解决k-Anonymity中存在的敏感属性泄露问题。

定义：每组具有相同QI值的记录，在每个敏感属性上至少有l种“well-represented”。对于“well-represented”，有以下几种定义方式：

1）不同值l-diversity。这种方式与l-sensitive k-anonymity相同。

2）熵l-diversity。对于一个组内特定敏感属性的熵的定义如下所示：

其中，p(G,c)是c在G中出现的概率，C是敏感属性值域，满足entropy l-diversity的数据集每个组G都有，。

3）Recursive (c,l)-diversity。确保让最频繁的值出现的不要太多，让最不频繁的值出现的不要太少。m表示组G中敏感属性值的个数 表示G中第i个最频繁出现值得次数。如果，就说组G满足递归(c,l)-diversity，如果数据集中每个组都满足递归(c,l)-diversity，那么整个数据集就满足递归(c,l)-diversity。

缺点：难以且没必要实现，同2中例子。有以下两种攻击其无法防御：

——偏斜攻击Skewness attack。满足上述原则的组中，患有AIDS和未患有AIDS的病人数目相同，那么如果攻击者成功确认一位特殊病人属于这个组，那么就可以得到在这个5/1000的数据集中，该病人患有AIDS的概率有50%。

——相似性攻击Similarity attack。如果敏感属性值存在语义上的相似，隐私泄露还是会发生。比如在一个组中存在{lung cancer，liver cancer，stomach cancer}三种属性值，那攻击者可以推断出该组中的人患有癌症。

4.t-Closeness 解决了l-diversity中固有的敏感属性泄露问题。

定义：每个组中敏感属性的分布和整个数据集属性的分布，之间距离小于阈值t。

因为两者分布接近，所以可以规避Skewness attack，此外，因为组中的属性分布模拟了整个数据集中的属性分布，所以只要数据集中没有语义相似，那么组中就不会存在语义相似。

缺点：破坏了数据集的信息完整性，1.限制了有用信息的发布数量，2.破坏了QI和敏感信息值的相关性，不利于数据挖掘的进行，可以通过放宽t值来减小危害。

常见的脱敏方法：

泛化算法：它通过将数据表中的 QI 属性值进行一定程度的泛化，即将具体的数据值转化为概括和抽象的形式，从而防止攻击者使用目标用户的 QI 值获取用户在数据表中的个体标识。对于数字类型的属性，数值将被转化为值域的类型，如将数值 24 转化为[10-30] 。而对于分类类型的属性，数值将根据用户自定义的泛化层次树进行泛化。

替换：以虚构的数据代替真值。例如，建立一个较大的虚拟值数据表，对每一真实值记录产生随机种子，对原始数据内容进行哈希映射替换。这种方法得到的数据与真实数据非常相似。

无效化：以 NULL 或 \*\*\*\*\* 代替真值或真值的一部分，如遮盖信用卡号的后 12 位。

置乱：对敏感数据列的值进行重新随机分布，混淆原有值和其他字段的联系。这种方法不影响原有数据的统计特性，如最大 / 最小 / 方差等均与原数据无异。

均值化：针对数值型数据，首先计算它们的均值，然后使脱敏后的值在均值附近随机分布，从而保持数据的总和不变。通常用于产品成本表、工资表等场合。

反推断：查找可能由某些字段推断出另一敏感字段的映射，并对这些字段进行脱敏，如从出生日期可推断出身份证号、性别、地区的场景。

偏移：通过随机移位改变数字数据。

FPE：Format Preserving Encryption， 即 格 式 维持的加密是一种特殊的可逆脱敏方法。通过加密密钥和算法对原始数据进行加密，密文格式与原始数据在逻辑规则上一致，如都为日期、卡号、结构化值等。通过解密密钥可以恢复原始数据。

基于其他参考信息进行屏蔽：根据预定义规则仅改变部分回应内容（例如，屏蔽 VIP 客户姓名，但显示其他客户等）。

限制返回行数：仅提供响应数据的子集，防止用户访问到全部符合要求的数据。

常用脱敏效果评测方法：

1.敏感数据保护性评估

首先，判断脱敏后数据集合是否满足相应的脱敏原则，然后模拟相应的攻击模式，评价用户敏感数据泄露的程度。

2.信息可利用性评估

目前，应用比较广泛的信息损失量评估方法包括：

1. 最小的失真（ Minimal Distortion）： 最小的失真[104]是被用于计算单个属性的信息损失量。对于一个属性，每当某个用户在该属性中的一个具体值被泛化时，惩罚值就会加 1。 文献[14] 和[18] 等使用最小的失真惩罚值作为算法中的启发式控制算法的分组策略。
2. 损失度量（ Loss Metric）：损失度量[105]被用于计算将一个具体的数据值进行泛化时产生的信息损失量。它的表达公式为：

其中，表示数据值在泛化树中包含子孙的数量，表示属性 A 中值域包含值的数量。

目前基于文本的敏感数据识别算法主要有三种：一种是基于概率和信息理论的分类算法，如朴素贝叶斯算法，最大熵算法。另一种是基于TFIDF权值计算方法，这类算法主要包括K邻近算法等。第三类是基于知识学习的算法，如支持向量机VSM算法，人工神经网络算法ANN等。

# 主要研究内容

智能大数据脱敏系统的功能按数据类型划分，主要包括数据库脱敏、文件脱敏、图片及视频脱敏几个主要部分，其组成如图所示：

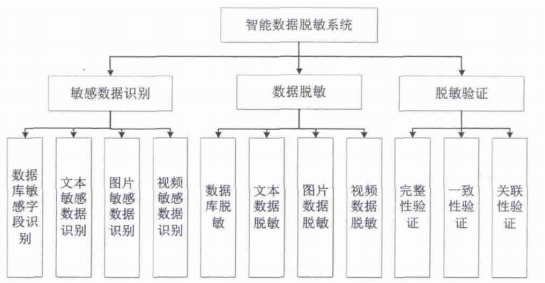


Figure 2 数据脱敏系统功能组成

结合项目需求，本文主要研究内容为数据库动态脱敏。在敏感数据识别阶段，主要对数据库敏感字段进行识别，若数据库有关字段值为文本信息，也可进行相应的文本敏感数据识别；在数据脱敏阶段，连接相应的数据源，针对数据库中不同类型的数据形态，实现不破坏其数据格式和可用性的数据脱敏处理；在脱敏验证阶段，对脱敏后的数据进行完整性、一致性和关联性三个方面的验证。

在紧密联系项目需求的基础上，设计出可用的数据脱敏系统架构，采用适当的脱敏算法，满足一定的脱敏原则，定义精确、灵活的脱敏策略，按照用户的权限等级，针对不同类别的数据以不同方式脱敏，实现跨工具、应用程序和环境的迅速、一致性的访问限制，解决具体情景下企业用户在数据共享、交换、使用过程中所面临的敏感数据安全问题。在脱敏系统处理的各个流程中，分别研究以下几方面的问题，其中敏感数据识别和脱敏策略是研究重点。

1.确定脱敏目标

数据敏感程度的分级和确认，包括确认原始数据的主观敏感性、使用场景下数据的关联性以及脱敏之后数据在系统开发和测试中的可用性。

2.敏感数据的识别

对目标系统的数据进行智能识别，获取用户数据源中数据元信息、数据结构等信息。对数据字段的内容进行分析，对格式和语义进行识别，对主键/ 外键进行处理，识别出系统中存在的敏感数据。核心技术采用数据特征学习以及自然语言处理技术，可能采用的方式有规则、知识库以及自然语言处理中的命名实体识别、特征词提取，特征密度计算等方式进行智能识别，大致识别流程如下图所示：

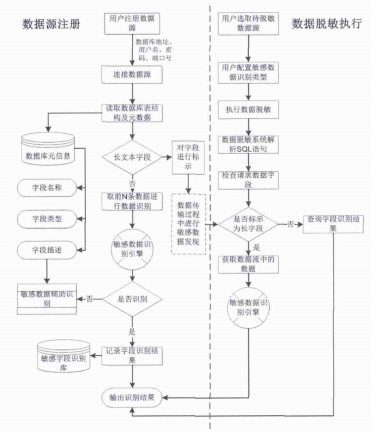


Figure 3 敏感数据识别方法

3.制定脱敏策略

对于不同类型的敏感数据字段，使用不同的脱敏算法，支持用户自定义的脱敏策略，可以灵活地更改脱敏算法，满足不同的脱敏原则。

4.设计脱敏系统架构



参照如图所示的系统架构，结合需要提供的功能，来设计系统的架构。

5.评测脱敏效果

根据相应的评测标准，从不同方面评价脱敏后数据的信息可利用性和脱敏后数据集合是否满足相应的脱敏原则。

6.结合脱敏与加密

探索如何将传统加密技术与数据脱敏技术结合起来的方式方法，对于没有数据挖掘需求的个人用户，采用传统加密技术保护其个人数据；对于需要大量数据并且有数据挖掘等相关需求的企业和组织，采用数据脱敏技术来保护用户的敏感信息；此外，考虑在脱敏数据基础上采用传统加密技术的必要性和可行性。

# 已完成的研究工作

已经完成了基本的脱敏算法，以及有关的脱敏原则的调研工作，了解了敏感数据识别所采用的机器学习方法，分析了美国人口普查等基本数据集；在分析数据库协议的基础上，了解了基于代理机制实现动态脱敏的大致流程。

# 研究方案及进度安排，预期达到的目标和取得的研究成果

## 研究方案

### 敏感数据识别的研究

对于文本内容的数据库记录，将大致采用如下流程来使用相应的模型：首先建立数据集，包含训练集和测试集。然后建立文本表示模型并进行文本特征选择。然后在训练集上进行机器学习，建立分类器。最后进行测试和性能的评价。

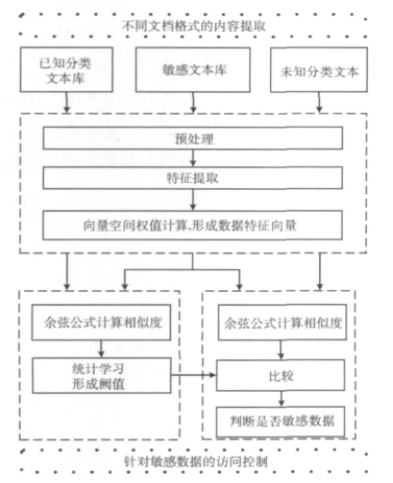


Figure 4 基于内容的敏感数据识别方法架构

对于数据库敏感字段，因为其语义和结构较为固定，将来拟采用规则库的方式来进行敏感数据的动态识别。

### 动态脱敏算法及原则的研究

动态脱敏采用基于代理的机制实现，对用户每次查询的不同结果集，动态的进行脱敏，然后返回脱敏后的查询结果。代理能自动识别目标为敏感数据的查询语句，并将语句改写为不包含敏感字段，或对敏感字段进行变换处理的查询语句。查询结果返回代理时，会被重新计算、修改并包装为与原请求一致的格式交付用户，从而完成一次敏感信息的查询过程。

在脱敏算法上，对于不同类型不同格式的敏感字段，测试不同脱敏算法在其上的效果；对于用户查询的不同规模的数据集，采用适当的脱敏原则，测试其在整个数据集上的脱敏效果。

## 预期达到的目标和取得的研究成果

完成整个基于代理的动态脱敏流程，通过机器学习方法和辅以人工标注的方式，对于数据库中的敏感字段和文本中的敏感信息进行有效的识别，并可以对识别的敏感字段推荐合适的脱敏算法，对整个查询的数据集推荐适当的脱敏原则，在保护隐私数据的基础上，力求数据信息的完整性，同时支持用户自定义相应的脱敏策略。

## 进度安排（建议从进入研究课题时间开始）

# 为完成课题已具备和所需的条件和经费

# 预计研究过程中可能遇到的困难和问题，以及解决的措施

难点1：敏感数据识别模型的实现

难点2：分析识别的不同敏感属性，推荐使用相应的脱敏算法

难点3：动态脱敏方式以及整个系统架构的设计和实现，如何脱敏的各个环节中保证隐私数据的安全性

难点4：在保护敏感信息的同时，注重维护数据信息的完整性，保证数据挖掘等的效果

# 主要参考文献

1. 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(06): 647-657.
2. 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述[C]. 中国计算机学会人工智能会议, 2013.
3. 陈天莹, 陈剑锋. 大数据环境下的智能数据脱敏系统[J]. 通信技术, 2016, 49(7): 915-922.
4. 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.
5. Walker S J, Viktor M-S, Kenneth C. Big Data: A Revolution That Will Transform How We Live, Work, and Think[J]. International Journal of Advertising, 2014.
6. 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014(4): 927-949.
7. Barbaro M, Zeller T. A face is exposed for AOL Searcher no.4417749[J]. New York Times, 2006.
8. Cox L. Suppression Methodology and Statistical Disclosure Control[J]. Publications of the American Statistical Association, 1980, 75(370): 377-385.