MapReduce模型下数据隐私保护机制研究

杨绍禹 王世卿
(郑州大学信息工程学院 郑州 450052)

摘 要 在对海量数据进行分析和处理的过程中,敏感信息的隐私保护显得尤为重要。针对统计类型数据分析服务的效率和安全问题,在Map-Reduce计算模型的基础上引入差别隐私保护机制。在该模型上提出一种带有隐私保护的决策树生成算法,并证明其满足ε-差别隐私。实验表明,该算法具有良好的分类精度和满意的计算效率。

关键词 Map-Reduce,差别隐私,统计数据,并行计算,决策树生成算法
中图法分类号 TP309.2 **文献标识码** A

Research of Data Privacy & Security in Map-Reduce Model

YANG Shao-yu WANG Shi-qing
(College of Information Engineering, Zhengzhou University, Zhengzhou 450052, China)

Abstract In analyzing and processing of the large-scale database, it is crucial for protecting the privacy of sensitive data. In order to analyze the service efficient and security for the mass data with statistical character, the mechanism with differential privacy was implemented in the Map-Reduce computation model. A decision tree's generation algorism was proposed at the computation model, and proved to satisfy the ε- differential privacy. The experiment indicates that the algorism has well classification accuracy and computation efficiency.

Keywords Map-Reduce, Differential privacy, Statistical data, Parallel computing, Decision tree generation algorism

1 引言

随着网格计算和云计算等高性能计算模型的发展,针对海量统计类型数据的分析和处理效率得到很大程度的提高。相比传统的相对闭合的数据处理环境,新的计算模型大多建立在开放的系统环境下。这样的数据处理环境为分布式数据存储和并行计算提供了良好支持,但同时也存在着很多的数据敏感信息泄漏隐患。数据拥有者对于数据保密性的控制权在很大程度上被削弱了。原始数据可能在数据采集、传输和存储的任何过程被敌手获取或者篡改。数据隐私无法得到发布者有效的安全确认,数据敏感信息甚至可能被数据库管理员通过非法手段获取[1]。如何在保证数据隐私的前提下尽可能地提高海量数据的计算效率,成为海量统计类型数据处理需要解决的新问题。

本文从统计数据分析的特点出发,在保证统计数据分析准确性的前提下,应用Map-Reduce计算模型处理数据,提高数据处理效率;并把差别隐私保护技术应用于分布式统计数据库中,使得敌手在获取统计数据信息后,无法得到敏感数据;进而,在非交互式的数据访问环境中,达到数据隐私保护的目的。本文第2节分析了数据隐私保护的定义和目前的研究方法;第3节构建了基于差别隐私的Map-Reduce模型;第4节在这个模型基础上设计验证其性能的分类算法;第5节比较了典型的几种分类算法在安全性和高效性上与此算法的差异。

2 数据隐私与问题描述

数据隐私问题的研究有着很长的历史,其中包括加密技术、匿名化技术、去可识别性技术等很多研究方法。根据统计数据库的特点,关于数据隐私性的保护主要是在不影响数据统计趋势的前提下,提供对某些敏感信息的隐私保护,比如医疗统计信息中的个人病历信息、人口普查信息中的收入状况信息等。

2.1 数据隐私定义

定义隐私是一个困难的工作,主要挑战来自于如何对敌手背景知识进行描述。敌手对于打算攻击的数据可以是零知识背景的,也可以在合法的数据请求附加信息查询和连接查询等方法,间接地获取敏感数据。比如,敌手通过合法查询获取某人的全名信息和出生日期,可以用一些手段连接查询某医疗病历数据库,进而获取更多的个人隐私信息。

统计数据分析研究者Dalenius在早期的研究中对数据隐私保护目标做了非形式化的说明,他认为任何能够从统计数据库中获取的信息,都可以不通过访问数据库来实现。文献[6,7]从敌手对数据认识的角度来定义隐私,要求敌手之前和之后对于个体认识的观点不应有很大的不同(比如在访问统计数据库之前和之后),或者说访问统计数据库不应该很大程度地改变敌手之前对于数据库中个体的认识。Dwork等人在

到稿日期:2012-02-22 返修日期:2012-06-10 本文受国家"十一五"科技支撑计划项目(2006BAF01A00)资助。
杨绍禹(1980—),男,博士生,CCF会员,主要研究方向为信息安全、云资源访问控制,E-mail:ysymagnet@gmail.com;王世卿(1951—),男,博士,教授,主要研究方向为电子商务、信息安全。

文献[1]中给出了针对于统计数据库的隐私保护目标：无论是数据库在线还是离线状态，确保数据库访问前后的数据威胁最小化。敌手在之前和之后对个体认识的比较转换成比较个体在数据库中的威胁的变化。

2.2 隐私攻击

在数据库交互式数据访问环境中，敌手可以通过数据侦听、伪访问操作等方式来对数据库敏感信息进行攻击。为了保证这些敏感数据在交互过程中不被敌手获取，可以通过访问操作上下文来确定访问操作是否安全。如果有暴露隐私的危险，则拒绝。但是这种方法也存在着缺陷，即可能拒绝访问某些操作本身就会暴露数据的隐私性[2]。非交互式数据访问形式是在统计数据库未进行数据发布之前对数据进行相应的处理以保证敏感信息不被暴露。这些处理有很多的技术和方法，比如替换可识别性敏感信息[3]、k-匿名隐私保护[4,5]、添加噪音数据[6]等。但是，它们也存在着一些问题，敏感数据属性替换方法容易受到敌手的连接攻击；k-匿名保护方法可能会覆盖一些特征数据；而噪音数据的大小可能会很大，从而影响数据的发布。

3 基于差别隐私的 Map-Reduce 模式

统计类型数据分析的研究对象并非针对某一个或者几个数据单元，而是研究数据的群体性现象。对于此类数据的敏感信息隐私保护，可以采用数据失真隐私保护方法。这种方法的特点是计算开销低、隐私保护强度较高等，但是个体信息的数据损害比较大，可能很难保持原始数据的真实值。对于海量的统计类型数据，由于其研究目的和数据规模，采用失真的隐私保护机制和并行计算模型可以很好地保证数据的安全性和计算的高效性。下面介绍的差别隐私保护机制就是一种基于数据失真处理的隐私保护机制。

3.1 差别隐私保护

差别隐私是对原始数据集进行的一种数据处理方法，经过这种方法得到结果输出的概率与真实数是否存在于输入数据集中无关。这个输出的概率不依赖于这个数据在数据集中的存在与否。或者说，对于一个隐私保护机制处理过的输出数据，敌手无法分辨其中的某个来自真实数据的输入，因为这个输入的存在性对于产生输出数据没有任何影响。差别隐私的定义是这样的。

定义 1 一个隐私保护机制 f 满足(ε,δ)-差别隐私[7]（其中，ε,δ 是隐私参数），当对于两个数据集 D 和 D' 来说，它们的差别最多包含一个记录（$|D\Delta D'|\leqslant 1$），则对于所有的输出 $S\subseteq Rang(f)$，有

$$\Pr[f(D)\in S]\leqslant e^{\varepsilon}\times\Pr[f(D')\in S]+\delta \tag{1}$$

式中，隐私参数 δ 主要是针对一些特殊计算的松弛参数，δ 参数的引入保证了数据集中可以包含小概率差别数据，在本文中，δ 设置为 0。对于函数 F 来说，为了能够衡量其在原始数据处理过程中的敏感程度，也就是当原始数据发生差别性变化后，输出结果的最大变化量，引入函数敏感度概念。

定义 2（敏感度，Sensitivity） 对于函数 $f:D\to\mathcal{R}^d$，f 敏感度是

$$\Delta f=\max_{D,D'}\|f(D)-f(D')\|_1 \tag{2}$$

对于所有的 D 和 D' 来说，$|D\Delta D'|\leqslant 1$。

为了实现差别隐私，需要对个体数据信息添加噪音数据。关于噪音数据的添加，差别隐私提供了很多机制[9,10]来保证其满足安全参数 ε 的要求。本文采用比较常用的拉普拉斯机制来添加噪音数据。

定理 1[11] 对于任何函数 $f:D\to\mathcal{R}^d$，隐私保护机制 f 满足 ε-差别隐私，通过为每条记录的输出添加拉普拉斯分布 $Lap(\Delta f/\varepsilon)$ 独立产生噪音数据。

$$f(d)+Lap(\Delta f/\varepsilon) \tag{3}$$

式中，$Lap(\Delta f/\varepsilon)$ 是一个标准差是 $\sqrt{2}\Delta f/\varepsilon$ 的对称拉普拉斯分布。

差别隐私保护通过对原始数据添加噪音的方式来实现对个体数据的敏感信息保护。这些噪音数据与函数敏感度有关，函数敏感度越高，说明数据存在性越易暴露隐私信息，因而需要添加更多数量的噪音数据信息来屏蔽输入数据和输出数据的差别。这样，对于敌手来说，输出的数据与其背景知识无关。

定理 2[16] 对于任意函数 $f:(D\times T)\to\mathcal{R}^d$，算法 Ag 以概率 Pr 选取输出 t 来满足 ε-差别隐私，其中

$$\Pr=\exp\left(\frac{\varepsilon\times u(D,t)}{2\times\Delta u}\right) \tag{4}$$

在实际问题中使用差别隐私为数据提供安全保障，有很多满足 ε-差别隐私保护的安全机制[16]，指数机制是其中应用比较广泛的一种。这个机制中证明了通过计算 $u(D,t)$ 函数，算法以指数分布的概率选取某一输出来满足 ε-差别隐私，这为 4.2 节中证明候选参数的选取过程是否满足 ε-差别隐私提供了理论支持。

3.2 改进的 Map-Reduce 模式

Map-Reduce 是 Google 研究人员面向分布式并行计算环境所设计的高效的计算模型[12]。它与 GFS(Google File System)文件系统和 BigTable 表结构共同组成了可实现分布式数据存储和并行计算的框架。

Map-Reduce 计算框架主要包含 3 个部分，如图 1 所示。1)输入：将处理数据分配给一个分布式的文件系统，这个文件系统可以将数据按照要求分解成若干个切片。2)Maper：编程人员需要对每一个分解的切片按照需求设计 Maper 函数，这些 Maper 函数的输入是每一个切片中的数据，输出是一个键/值对列表数据；3)Reducer：这个阶段由 Reducer 函数对 Maper 阶段产生的键/值对列表进行相应操作（如求和、计数、最值等），并将得到的结果输出。

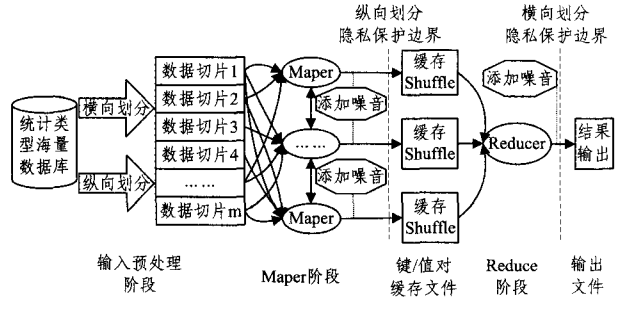

图 1 差别隐私保护的 Map-Reduce 模型

Map-Reduce 计算框架采用并行计算的模式，很大程度上提高了计算效率，降低了分析海量数据的时间复杂度。对于统计数据库来说，其数据量庞大，数据维度一般比较高，利用

并行化计算模型可以极大地提高具有同步性特点计算的运行效率。为了更好地对统计数据进行处理，采用横向和纵向两种划分形式对统计数据进行输入预处理[13,14]。横向划分是将大小为 S 的统计数据根据 maper 单元的数量 M 水平分割记录集，每一个 maper 单元的记录集的大小为 S/M；纵向划分是将统计数据集 S 根据 maper 单元数量和容量把属性集 A 划分成($\bigcup A_{i.1\leqslant i\leqslant k}, A_{cls}$)的形式。水平划分数据集方式可以使 maper 函数计算过程具有良好的并行性，可降低 $O(1/M)$ 的时间复杂度。但是，由于在每个 maper 过程中破坏了数据的完整性，因此需要增加各并行进程的通信开销；纵向的属性划分由于保持了属性上的数据完整性，减少了通信开销，因此适合对统计数据进行整体分析，但是会有敏感性属性信息泄漏的危险。

在输入预处理阶段，用户通过 Map-Reduce 框架 maper 函数对每个 maper 单元的数据进行读取，通过 reducer 函数完成结果汇总。根据不同的预处理输入方式可以选择不同的差别隐私保护机制。对于水平预处理输入，由于得到的是不完整的数据集，可以在 Reducer 阶段添加噪音数据以屏蔽结果数据；对于垂直预处理输入，则需要在 Maper 阶段后添加噪音数据，防止敌手从 Shuffle 中获取敏感性信息。

4 带隐私保护的决策树生成算法 DP-MR

基于差别隐私保护的 Map-Reduce 模型将数据处理过程分为了 3 个阶段和 2 种划分的形式。为了有效地验证这个模型在分析统计类型数据过程中的可行性，本节针对统计数据库提出一种基于 Map-Reduce 模型带有差别隐私保护的决策树生成算法，并给出了算法满足 ε-差别隐私的证明，最后讨论了这个算法实现的时间复杂度。

4.1 概述

对统计数据进行数据分析有很多方法，为了更好地验证差别隐私在保证数据一致性上的效果，本文提出一种决策树生成算法 DP-MR。其通过数据的水平分解对统计数据库进行切片划分，采用 Map-Reduce 模型对这些数据切片进行处理，采用信息增益度作为节点候选计算函数，以指数机制作为控制输出的概率分布，保证决策树节点生成满足 ε-差别隐私。最后，在拉普拉斯分布内完成对原始数据的噪音添加，确定隐私保护边界。算法 DP-MR 描述如下：

输入：统计数据库 D，隐私保护参数 ε，迭代次数 n；

输出：统计数据决策树 T。

步骤 1　输入预处理

初始化数据切片 $split_i$，$1\leqslant i\leqslant M$，且 $split_i$ 的属性向量为($A_1,\cdots,A_k,A_{cls}$)，其中 A_{cls} 是分类属性，有 t 个不同取值 aq_t。

步骤 2　执行 maper 函数

maper(〈key,value〉,〈key′,value′〉)，其中 key 为数据切片中的每一个属性 $A_i\in\{A_1,\cdots,A_k\}$；value 为数据切片中的记录 $r_j\in split_i, j=|Rows_{split_i}|$；key′为输出的属性键值列表，value′为输出的键值列表对应的值。具体实现如下：

```
1. for each r_j:
2.   if A_i. r_j〈〉aq_t then:
3.   〈key′,value′〉=〈key′∪A_i,value′+1〉;
4. next r_j;
5. return 〈key′,value′〉;
```

步骤 3　执行 reduce 函数，选择决策树属性扩展节点

reducr(〈key′,value′〉)，其中 u(S,A_i)函数是确定生成节点选择函数的一个描述；infoGain(value′,A_i)表示计算 key′列表中 A_i 个属性的信息增益度；通过递归方法获取决策树节点并返回阶段属性值和添加噪音数据的计数。具体实现如下：

1. 初始化〈key′,value′〉列表中具有相同属性的键/值对，形成新的键/值列表〈∪Key′,Value′〉；
2. 判断迭代次数是否达到用户定义最大值；
3. 计算候选节点信息增益度 infoGain(value′,A_i)；
4. 以定理 2 中的概率 Pr 在 A∈∪key′选取信息增益度最大的节点输出；
5. 为 A. value′ 添加噪音数据 Lap($\Delta f/\varepsilon$)；
6. 输出键值对〈A,A. value′〉，记录决策树节点；
7. 在原始键值对中去除已输出项，并更新；
8. 递归调用 reduce 函数。

步骤 4　生产决策树

利用 reduce 递归过程获取的〈A,A. value′〉键/值对依次生成决策树节点 v，并将当前节点 r 位置指向其子树节点 r→child(r)，返回统计数据分类决策树 T。

4.2 算法隐私性分析

算法 DP-MR 的核心部分在于 reduce 阶段在键/值对〈$\bigcup Key', Value'$〉中选择生成节点的步骤。节点选择的依据是该节点的信息增益度 $infoGain(value', A_i)$。在这里用 S 表示键/值对列表〈$\bigcup Key', Value'$〉，A_i 表示候选生成节点，$A_i.child$ 表示 A_i 节点的划分，则属性 A_i 的信息增益度可以表示为：

$$infoGain(S,A_i)=H_{A_i}(S)-H_{A_i|A_i.child}(S) \tag{5}$$

式中，$H_{A_i}(S)$ 是属性 A_i 相对于数据集 S 的信息熵，$H_{A_i|A_i.child}(S)$是取值介于 1 到 $H_{A_i}(S)$之间属性 A_i 的条件熵。信息增益度函数的敏感度 $\Delta f=\log_2^{|D(A_{cls})|}$。算法 DP-MR 中的 $|D(A_{cls})|$表示分类属性取值范围，对于 A_{cls} 的结果只有 Yes/No，所以其取值范围为 2，Δf 的最大取值是 1。

引理 1　从键/值对列表〈$\bigcup Key', Value'$〉中选取生成节点的算法 Ag 满足 ε-差别隐私，当选取生产节点的概率是 Pr(见定理 2)。

证明：在候选节点选取过程中，选择过程满足指数机制，选择概率是：$\Pr/\sum_{A_i\in\bigcup Key'}\Pr$，其中函数 u 的敏感度为 $\log_2^{|D(A_{cls})|}$，即 $\Delta u=1$。根据定理 2 可知，在对键/值对列表〈$\bigcup Key', Value'$〉进行选择输出时，选取概率满足指数机制，所以节点的选取满足 ε-差别隐私。

引理 2　DP-MR 算法满足 ε-差别隐私。

证明：算法首先对统计数据进行横向划分，生成数据切片，根据并行组合定理[15]，对于正交数据集上的算法序列满足 ε-差别隐私。其次，对于算法中决策树节点的选取，由引理 1 可知其满足 ε-差别隐私。最后，对每一个组中带噪音数据的统计计数输出采用拉布拉斯机制，它能够保证 $\varepsilon/\Delta f$-差别隐私，$\Delta f=1$，所以也满足 ε-差别隐私。综上所述，DP-MR 算法的各阶段均满足 ε-差别隐私，根据顺序组合定理，可知 DP-MR 算法满足 ε-差别隐私。

4.3 算法复杂度

相比于传统的决策树生成算法，由于 MapReduce 计算模型的并行性特点，算法 DP-MR 的运算效率得到很大提高。m

个 maper 单元可以将扫描一次统计数据库的时间减少到原来的 $1/m$，每一个 maper 阶段由一次数据切片遍历构成，其时间复杂度为 $O(|D|/m)$。统计数据库经过 maper 阶段处理会产生 m 个缓存 Shuffle，在每一个 Shuffle 中包含了〈key，$value$〉列表，这个列表用来统计这个数据切片中的属性及其计数。在 reduce 阶段，首先初始化〈key'，$value'$〉列表，时间复杂度为 $O(m)$；其次，采用指数机制选择候选节点输出和添加噪音数据均为 $O(|\cup Key'|)$。利用迭代方式选取生成节点，则复杂度为 $O(n\times|\cup Key'|)$。

综合两个阶段可知，总的时间复杂度由 $O(|D|/m)$、$O(m)$ 和 $O(n\times|\cup Key'|)$ 决定，这与 Map-Reduce 计算模型并行程度，即数据集 D 的细化程度以及数据集 D 的属性分类有关。通常情况下，由于 maper 数量相比较统计数据集 D 中记录个数要小得多，生成的决策树迭代次数 n 也在指数范围内，因此可以用 $O(|D|/m)$ 来描述算法的时间复杂度。

5 实验与评价

为了更好地说明算法 DP-MR 在保证数据分析准确性的基础上，能提供敏感信息的差别隐私保护，将本文算法与典型的决策树生成算法 ID3 和 C4.5 算法的分类精度进行对比。实验数据采用开源的人口普查数据集 Adult，其中包含 45468 个普查数据，14 个不同类型的属性。为了便于数据分析，选择其中的 8 个分类性属性数据作为分析对象。数据集中 50%作为训练集，用来生成决策树，其余为测试集，利用生产的决策树测试其分类精度。硬件平台采用 8 台 PC，其中 3 台配置为四核 3.6G，4G 内存，其余机器为双核 2.8G，2G 内存。构建 Hadoop 集群使用 Hadoop 1.0 版本，在每个节点上部署 R 统计数据分析工具。集群中各节点通过千兆以太网连接。

通过调节隐私保护参数 ε 来控制隐私保护强度，ε 取值越大，说明隐私保护能力越弱，反之越强。在不同的隐私保护强度下，利用不同的算法生成决策树，在测试集上得到的测试精度数据如图 2 所示。

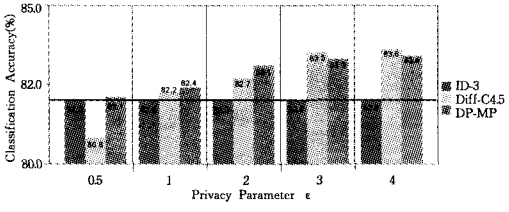

图 2 隐私保护强度与分类精度

3 种算法的分类精度基本上在 80%到 85%之间，ID3 算法没有隐私保护机制，参数 ε 对其没有影响，分类效率稳定在 82%。两种带有隐私保护的算法 Diff-C4.5 和 DP-MR 在不同的隐私保护强度下表现各不相同，但是趋势比较类似，都是随着隐私保护强度的减弱，分类精度呈现升高的趋势。相比于 Diff-C4.5 算法，DP-MR 算法的隐私保护参数 $\varepsilon<3$ 时，其分类效果要优于前者，随后两者比较接近。因为对于统计类型数据集来说，为了很好地保护数据隐私，ε 一般取值小于 1，所以 DP-MR 算法相比于其他两个算法，在限定的隐私保护强度内有一定的优势。

为了更好地分析 DP-MR 算法的计算效率，对上面 3 种算法在不同迭代次数情况下在测试集上的分类效率进行分析，如图 3(a)所示。分类精度开始时随着细化次数的增加而增加，但是，在一个确定的细化次数后，分类精度开始随着细化次数的增加而减小。这是因为随着迭代次数的增加，$\cup Key'$ 中的数量也在增加，这就造成每一个节点的统计计数值很小。虽然迭代次数越多，得到的分类属性越细，但是拉普拉斯噪音数据对于统计值的影响也越来越大。当迭代次数超过某一阈值时，可能出现噪音数据与计数值接近。这样得到的决策树会影响到分类精度。图 3(b)用时间开销来衡量 DP-MR 算法在时间上相比于其他算法具有显著的优势。不过，在实验过程中，发现当 maper 节点数量 m 在增加的过程中，m 小于 5 时，时间开销接近串行计算开销的 $1/m$；但是，之后时间开销基本相同，稳定在一个时刻上。这是由于 DP-MR 算法的主要时间开销在 Reducer 阶段的迭代环节，如果数据切片划分得越多，Maper 阶段的时间开销就会与 Reducer 阶段越接近。

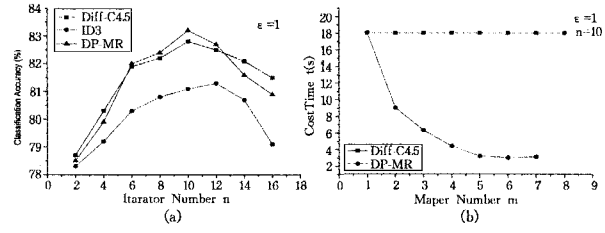

图 3 分类精度与 M-R 模型效率

结束语 本文对海量统计数据库中敏感信息的隐私保护机制进行了研究，利用 Map-Reduce 计算模型在一定程度上提高了海量数据的处理效率，同时又能够提供用户控制的隐私保护机制。针对分类决策树生成问题，引入差别隐私保护机制，在 Map-Reduce 计算模型下提出高效、安全的 DP-MR 算法。文章分析了这个算法的复杂度，并证明了其满足 ε-差别隐私，最后对这个算法的分类精度、隐私保护强度和运行效率进行了验证。本文提供的算法 DP-MR 能够很好地适应并行化的计算环境，提高了运算效率，并对数据的敏感信息提供了有效的隐私保护。

参 考 文 献

[1] Dwork C. A Firm Foundation for Private Data Dnalysis[J]. Communications of the ACM, 2011, 54(1): 86-95

[2] Kenthapadi K, Mishra N, Nissim K. Simulatable auditing[C]// 24th ACM Symposium on Principles of Database Systems. New York, NY, USA: ACM, 2005: 118-127

[3] Mohammed N, Chen Rui, et al. Differentially Private Data Release for Data Mining[C]// KDD '11. New York, NY, USA: ACM, 2011: 493-501

[4] Samarati P. Protecting respondents' identities in microdata release[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 6(13): 1010-1027

[5] Sweeney L. k-anonymity: A model for protecting privacy[J]. IEEE Security And Privacy, 2002, 5(10): 557-570

[6] Dinur I, Nissim K. Revealing information while preserving privacy [C]// the 22nd ACM Symposium on Principles of Database Systems. New York, NY, USA: ACM, 2003: 202-210

[7] Dwork C. Differential Privacy[C]// ICALP'06. Venice, Italy: Springer Verlag, 2006: 1-12

[8] 金华,刘善成,鞠时光. 面向多敏感属性医疗数据发布的隐私保护技术[J]. 计算机科学,2011,38(12):171-177

[9] Barak B, Chaudhuri K, Dwork C, et al. Privacy, accuracy, and consistency too: a holistic solution to contingency table release [C]// PODS'07. Beijing, China: Association for Computing Machinery, Inc, 2007: 273-282

[10] Dwork C. Differential privacy: A survey of Results [C]// TAMC'08. Xi'an, China: Pringer-Verlag, 2008: 1-19

[11] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]// TCC'06. New York, NY, USA: Springer, 2006: 265-284

[12] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters[J]. Communications of the ACM, 2008, 1(51): 107-113

[13] 刘欣阳,王国仁,乔百友,等. 决策树的并行训练策略[J]. 计算机科学,2004,31:129-130,135

[14] Satuluri V. A survey of parallel algorithm for classification[J]. the International Parallel Processing Symposium, 2007, 2(21): 1233-1245

[15] McSherry F. Privacy integrated queries [J]. Communications of the ACM, 2010, 9(53): 89-97

[16] McSherry F, Talwar K. Mechanism design via differential privacy[C]// FOCS'07. Washington, DC, USA: IEEE Computer Society, 2007: 94-103

(上接第 120 页)

知识库规则:

条件:在 TBox 中存在 C⊑D 操作:

操作:L(x)=L(x){¬C⊔D}

2. 推理算法优化:由于算法采用深度优先的策略进行生成,算法执行时的最大深度为知识库规模的指数,因此推理算法的执行耗费很大,需要进行一系列的优化。主要的优化包括:a)设计存储比较结果的数据结构,因为在算法中不需要关心具体的数值,仅需要关心数值间的大小比较,所以在算法初始化阶段就将所有出现的数值比较一次,在后续的算法过程中不再重复比较;b)算法中已经遍历过的标记集 $L(x)$可能在后续被再次检查,因此在算法中保留全局的标记集寄存集合,若该标记集 $L(x)$被检查过就将该集合放入标记集寄存集合。

3. 推理分割机制:由于算法的复杂度较高,执行一次全局知识库的推理耗费很大,因此可以将全局的大知识库 $\mathcal{K}=\langle\mathcal{T},\mathcal{A}\rangle$分成小的局部知识库:

$\mathcal{K}_1=\langle\mathcal{T}_1,\mathcal{A}_1\rangle;\mathcal{K}_2=\langle\mathcal{T}_2,\mathcal{A}_2\rangle;\cdots;\mathcal{K}_n=\langle\mathcal{T}_n,\mathcal{A}_n\rangle$

式中,$\mathcal{T}_i\subseteq\mathcal{T},\mathcal{A}_i\subseteq\mathcal{A},1\leqslant i\leqslant n$。通过对局部知识库的检查获得全局知识库的近似结果,如果局部知识库不可满足,全局知识库也不可满足;但反之不成立,因此该方法是查错的不保真的方法。但由于原算法的复杂很高,局部推理的代价将远远小于全局推理,因此该分割机制能极大地降低推理的耗费。

结束语 本文分析当前基于描述逻辑的软件建模中的现状和问题:对于复杂的数值间关系,描述逻辑直接表示会导致极高的推理复杂性甚至推理问题不可判定。为解决该问题,本文提出一种模糊近似化软件数值域,然后将扩展模糊描述逻辑作为软件模型的形式化基础的软件模型建模框架,讨论了框架的 3 个核心问题:软件数值域模糊化、软件数值知识库构造和软件数值模型推理;研究解决了 3 个核心问题的方法和步骤。未来工作包括对现有的描述逻辑体系增加支持动态特性的逻辑组件,构建更具表示能力和推理能力的软件模型。

参 考 文 献

[1] Berners-Lee T, et al. The Semantic Web [J]. Scientific American, 2001, 284(5): 34-43

[2] Horrocks I, et al. Reducing OWL entailment to description logic satisfiability [J]. Journal of Web Semantics, 2004, 1(4): 345-357

[3] Baader F, et al. Description logics as ontology languages for the semantic Web[J]. Lecture Notes in Artificial Intelligence, 2005, 2605: 228-248

[4] http://www.w3.org/TR/owl-features/

[5] 董明楷,等. 基于动态描述逻辑的主体模型[J]. 计算机研究与发展,2004,41(5):780-786

[6] 蒋运承,等. 基于描述逻辑的模糊 ER 模型[J]. 软件学报,2006,17(1):20-30

[7] 何坚,等. 面向电子商务的知识描述语言[J]. 中文信息学报,2004,18(6):37-42

[8] 梁晟,等. 语义 Web 规则标记语言 OWLRule+的设计与实现[J]. 计算机研究与发展,2004,41(7):1088-1096

[9] 左志宏,等. 描述逻辑 SHIQ 与网络资源自动分类[J]. 通信学报,2004,25(7):200-206

[10] 常亮,等. 可判定的时序动态描述逻辑[J]. 软件学报,2011,22(7):1524-1537

[11] 贾育,等. 基于领域特征空间的构件语义表示方法[J]. 软件学报,2002,13(2):311-316

[12] 张大鹏,等. 一种基于主体的可信网构软件设计方法[J]. 电子学报,2010,38 (11):2523-2528

[13] 刘磊,等. 一种基于描述逻辑的 Web 服务动态组合算法[J]. 吉林大学学报:理学版,2008,46(2):259-264

[14] 牟向伟,等. 基于模糊描述逻辑的个性化推荐系统建模[J]. 计算机应用研究,2011,28(4):1429-1433

[15] Baader F, et al. GCIs Make Reasoning in Fuzzy DL with the Product T-norm Undecidable[C]// Rosati R, Rudolph S, Zak-haryaschev M, eds. Proceedings of the 24th International Workshop on Description Logics (DL 2011). volume 745 of CEUR-WS, Barcelona, Spain, 2011

[16] Borgwardt S, et al. Fuzzy ontologies over lattices with t-norms [C]// Rosati R, Rudolph S, Zakharyaschev M, eds. Proceedings of the 24th International Workshop on Description Logics (DL 2011). volume 745 of CEUR-WS, Barcelona, Spain, 2011

[17] 康大周,等. 支持模糊隶属度比较的扩展模糊描述逻辑[J]. 软件学报,2008,19(10):2498-2507

[18] Horrocks I, et al. A tableaux decision procedure for SHOIQ[C]// Proceedings of Nineteenth International Joint Conference on Artificial Intelligence. Edinburgh, Scotland, UK, 2005