大数据安全及其评估

陈文捷 蔡立志
(上海市计算机软件评测重点实验室 上海 201112)

摘 要 大数据的安全问题是影响大数据应用的关键因素之一,而评估大数据应用的安全性成为业界关注的课题。针对大数据应用安全性的评估问题,在梳理大数据安全研究现状的基础上,从数据和计算两个层面上分析大数据所面临的安全问题,综述目前主要的解决大数据安全问题的研究成果,包括分布式计算的安全技术、数据溯源技术、隐私保护的数据挖掘技术等。最后从数据的可信性、隐私保护程度等方面提出一些大数据安全性的评估指标。

关键词 大数据 安全 隐私保护 隐私保护的数据挖掘 安全评估

中图分类号 TP309 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2016.04.009

BIG DATA SECURITY AND ITS EVALUATION

Chen Wenjie Cai Lizhi
(*Shanghai Key Laboratory of Computer Software Testing and Evaluating, Shanghai 201112, China*)

Abstract Big data security is one of the key factors affecting big data applications, and the evaluation of the security of big data applications becomes the industry concern. In light of this issue, in this paper we analyse the security challenges encountered by big data from the aspects of data and computing based on sorting the status quo of big data security studies. Then we give a survey on the main research outcomes of solving these challenges, including the security technology of distributed computing, the data traceability technology, and the data mining technology for privacy protection. Finally, from the aspects of data creditability, privacy protection degree, etc., we also propose some evaluation indices for the big data security.

Keywords Big data Security Privacy protection Data mining for privacy protection Security evaluation

0 引 言

近几年,随着移动终端以及互联网的发展,数据呈现出爆发式增长,“大数据”成为 IT 领域关注的热点。2013 年英特尔公司的一组调查数据显示:一分钟之内全球每分钟传输的数据几乎可以达到 640 000 GB[1]。对于大数据的定义目前还不统一,不同的公司和机构有着不同角度的诠释,但基本都提到了大数据是一种无法通过人力和主流软件在短时间内处理的海量数据。随着大数据时代的到来,大数据的应用和技术已经开始逐渐渗透到社会的各个领域,大数据分析也成为一门新兴学科。

尽管大数据的涌现为人们提供了前所未有的宝贵机遇,但同时也提出了重大的挑战。其中的一个重大挑战是大数据的安全问题。随着各种数据挖掘手段的推进,人们可以从大数据中挖掘出大量有价值的信息,有些甚至涉及到企业机密甚至国家机密,因而吸引了黑客的各种攻击行为,例如数据窃取和篡改、隐私挖掘等。人们越来越觉得自己的隐私有被泄露的危险。近年来,关于大数据的安全事件不断发生,例如斯诺登“棱镜门”事件、MongoHQ 数据泄露事件等,使得人们越来越关注大数据的信息安全。

大数据安全是一个综合性的课题,涉及的技术包括密码学、数据挖掘等许多学科。产业界和学术界也积极关注大数据的安全问题。云安全联盟 CSA(Cloud Security Alliance)在 2012 年 4 月组建了大数据工作组 BDWG(Big Data Working Group),旨在寻找大数据面临的主要安全问题及其解决方案。国内外也有一些学术文献对大数据环境下的风险、安全问题进行过探讨[2,3]。本文在梳理大数据安全研究现状的基础上,分析了大数据所面临的安全问题,阐述了目前主要的解决大数据安全问题的研究成果。最后针对大数据安全性的评估提出了一些评估指标。

1 大数据的安全问题

大数据由于其分布式、数据量大、蕴含知识等特性,产生了很多新的安全问题,这些安全问题涉及大数据处理流程的各个环节。图 1 是大数据处理的典型流程,数据源中的数据进行抽取和集成后存入数据存储设备中。然后对存储的数据进行分布式计算或者数据挖掘等分析手段,最后将分析结果提交给具体的应用。

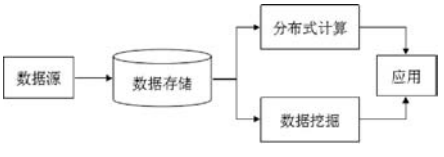

图 1 大数据处理流程

收稿日期:2014 - 11 - 08。上海市科委科技攻关项目(13511505303);上海市技术带头人项目(13XD1421800)。**陈文捷**,硕士生,主研领域:大数据,软件测试。**蔡立志**,研究员。

在这整个过程中,大数据的安全问题会出现在数据源、数据存储、数据分析以及数据传输的各个环节中。主要可以分为三类安全问题:数据安全、分布式计算安全和数据挖掘的安全。其中数据安全是指数据的来源、存储和传输过程中面临的安全问题,分布式计算安全和数据挖掘安全是指对大量数据进行计算和挖掘时产生的安全问题。

1.1　数据安全问题

(1) 数据来源安全。大数据处理的第一步是数据采集,对于采集得到的数据,有些数据可能是不可信的。因此需要对数据的来源进行仔细的甄别,否则通过分析这些数据得到的结果可能是不准确的甚至是错误的。

攻击者可能通过修改数据采集软件、篡改数据本身或 ID 克隆攻击等手段来刻意伪造数据。或者修改数据中的一些关键属性信息(如数据大小、创建时间等),使得分析者对这些数据分析后得出错误的结论,从而达到攻击者的目的。由于大数据的低信息密度的特性,从大量信息中鉴别出虚假信息往往非常困难。

(2) 数据存储安全。大数据是一种超大规模和高并发的非结构化数据,无法用传统的关系型数据库存储,因此往往被存储在非关系型的数据库中,如 Google 的 BigTable、Apache 的 HBase 等。然而相对于较成熟的关系型数据库,非关系型数据库的发展刚刚起步,其安全性还有待完善。一方面,验证和鉴权机制较为薄弱,使得数据库容易遭受暴力破解和来自内部的攻击,攻击者可能窃取或篡改数据,造成敏感数据被泄露。另一方面,非关系数据库也易受各类注入攻击,如 JSON 注入、REST 注入、schema注入等,攻击者可以利用这些注入手段向数据库中添加垃圾数据。

另外,大数据的存储是一种分布式的存储,其事务处理的一致性较弱。根据 CAP 理论,一个分布式系统无法同时满足一致性、可用性和分区容错性,而且一致性和可用性是一对矛盾,所以分布式存储可能无法在任何时刻都提供一致的数据查询结果。

(3) 数据传输安全。数据在传播过程中可能失真或被破坏[2]。原因之一是某些数据采集的过程需要人工干预,其中可能引入误差。原因之二是早期采集的数据由于现实情况发生了变化而已经变得过时。原因之三是攻击者可能通过执行中间人攻击 MITM(Man In The Middle)或者重放攻击等手段,在数据传输过程中破坏数据。

数据在传输过程中也可能被拦截和泄露[3]。客户与服务器之间的数据传输没有加解密处理,攻击者就可以在传输的过程中窃取数据。例如,配备 GPS 定位跟踪装置的移动电话可能泄露用户的位置信息。泄露的数据还往往会被多方利用。而用户无法知道自己的数据是在哪个环节被泄露,以及是谁泄露的,从而加大了用户的担忧。

1.2　分布式计算安全问题

大数据由于其数据量巨大,需要用分布式的方式来处理。比如 MapReduce[4] 就是业界常用的一个分布式计算框架,它能够处理大数据量问题,被应用在许多行业和科研领域中。但是在应用环境中,分布式计算并非是安全可靠的,实际中存在一些不安全因素。

分布式处理的函数可能被黑客修改或伪造,用于一些不可告人的目的。比如对云架构实施攻击、监听请求、篡改计算结果、发送虚假数据或改变工作流程,使得最终的数据分析结论不符合事实,或造成用户数据的泄漏。也可能集群中的一个工作节点发生某种故障而导致错误的计算结果。而在大量的工作节点中很难找出有问题的节点,从而对安全隐患的探测造成更大的困难。

分布式处理的工作集群缺乏完善的安全认证机制和访问控制机制,使得黑客可以冒充他人,并非法访问集群,恶意提交作业,或者随意地篡改数据节点上的数据,甚至可以任意修改或杀掉任何其他用户的作业,造成安全隐患。

1.3　数据挖掘安全

大数据的核心是数据挖掘技术,从数据中挖掘出信息,为企业所用,是大数据价值的体现。然而使用数据挖掘技术,为企业创造价值的同时,随之产生的就是隐私泄露的问题。

数据挖掘技术使得人们能够从大量数据中抽取有用的知识和规则。然而,这些知识和规则中可能包含一些敏感的隐私信息,数据分析人员往往可以利用数据挖掘算法,找出非隐私信息和隐私信息之间的关联。从个人的非隐私信息推理出他的隐私信息,从而造成用户隐私信息的泄露。一个典型的例子是某零售商通过分析销售记录,推断出一名年轻女子已经怀孕,并向其推送相关广告信息,而这名女子的家长甚至还不知道这一事实[5]。虽然可以采用数据加密、数据匿名等方法在数据挖掘时保护隐私信息,但是一方面分析、处理大规模的加密数据变得困难,影响了数据挖掘的性能;另一方面,仅通过匿名技术并不能很好达到隐私保护目标。例如,AOL 公司曾将部分搜索历史中的个人相关信息匿名化,并将之公布供研究人员分析。即使如此,还是有分析人员通过数据挖掘技术识别出其中一位用户的详细信息[6]。这位用户是一位 62 岁妇女,编号为 4417749,家里养了三条狗,患有某种疾病等等。

2　大数据安全防范的关键技术

针对大数据所面临的数据安全、分布式计算安全、数据挖掘安全问题,国内外学者开展了许多关键技术研究。这些安全技术从不同方面解决大数据的安全问题。在数据自身的安全防范技术中,主要有数据溯源和数据扰乱技术来保证数据的可信性和隐私性。安全计算框架的开发是基于计算框架的安全防范技术。隐私保护的数据挖掘技术保证了数据挖掘时不泄露隐私。本节分别选取其中的一些主要技术予以介绍。

2.1　数据自身的安全防范技术

(1) 数据溯源。面对大数据应用中数据被篡改的危险,可引入数据溯源技术保证数据的可信性。数据溯源是一种记录从原始数据到目标数据演变过程的技术,用于评估数据来源的可信性,或在灾难发生后对数据进行恢复。在大数据前期处理过程中,如果将数据溯源技术用于大数据处理中,则能为后期的数据处理提供验证和清理的支持。数据溯源的主要方法是标记法[10-12],即对数据进行标注,记录原始数据的出处、演算过程等。此方法又可细分为 why、where、who 等类别,分别记录数据的演算过程、出处、相关使用者等。除此之外,数据溯源技术还可用于流数据与不确定数据[13]。

Muniswamy-Reddy 等人在数据溯源技术的基础上,提出了一种在统一环境下追踪数据起源的感知起源存储系统 PASS(Provenance Aware Storage System)[14],它能自动收集、存储、管

理并查询文件的起源信息。PASS 利用修改过的 Linux 内核,在操作系统层对起源信息进行收集,并对读写操作记录详细的信息流和工作流描述。

(2) 数据扰乱。为了降低数据泄露隐私风险,一种较常用的方法是对原始数据进行一定的处理,隐去其中的敏感数据。数据扰乱技术是对数据本身进行一些修改,以删除或弱化其中隐私敏感的部分。数据扰乱有多种方式,比如数据乱序、数据交换[16]、数据扭曲[17-22]、数据清洗[23,24]、数据匿名[25,26]、数据屏蔽[27,28]、数据泛化[29,30]等。即将原始数据重新排列、对多条记录的某些属性值进行交换、在原始数据上叠加一个噪声、删除或修改某些记录、对某些记录的关键属性作删除或泛化、将某些属性值用概率分析法修正、将属性值替换为一个更抽象的值(比如"北京人"、"南京人"替换成"中国人")。

数据扰乱技术虽然能够一定程度保护隐私,但同时由于数据本身被修改,会对数据挖掘结果造成影响,因此使用数据扰乱技术需要在隐私保护程度和数据挖掘精度上作一个权衡。

2.2 基于计算框架的安全防范技术

如 1.2 节所述,分布式计算框架的安全隐患主要在于不可信的计算节点及认证授权机制。因此解决计算框架安全问题的主要途径是建立安全的认证授权机制和减少不可信计算节点的影响。

德克萨斯大学的 Indrajit Roy 等人基于流行的 MapReduce 框架,开发了一套分布式计算系统 Airavat[8],主要为了解决 MapReduce 的安全问题。Airavat 在 SELinux 中运行,并利用了 SELinux 的安全特性,防止系统资源泄露。在认证授权机制方面,开发人员采用了 Kerberos 认证。Kerberos 协议是一种计算机网络授权协议[7],为网络通信提供基于可信第三方服务的面向开放系统的认证机制,是一种应用对称密钥体制进行密钥管理的系统。同时 Airavat 整合了强制访问控制 MAC(Mandatory Access Control)和差分隐私技术。其中,MAC 是由系统强制确定访问主体能否访问相应资源的一种访问控制机制,可以提供细粒度的访问控制。差分隐私技术是由 Dwork 等人在 2006 年提出[9],解决了传统的匿名保护方法易受背景知识攻击的缺点,它通过在分析结果中加入噪声的手段使得攻击者无法分析出原始数据中的隐私信息。Airavat 系统结构如图 2 所示,它包括三个角色:计算提供者、数据提供者和 Airavat 计算框架。其中计算提供者使用 Airavat 编程模型编写 MapReduce 代码,数据提供者指定隐私策略的参数。

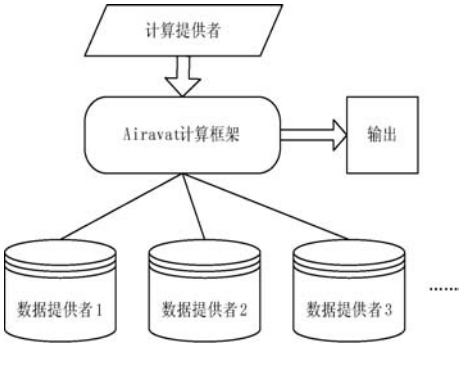

图 2 Airavat 系统结构

2.3 数据挖掘中的隐私保护的技术

由于数据挖掘可能泄露用户的隐私,因此促使学者开始研究数据挖掘中的隐私保护方法,即在控制数据隐私泄露的情况下进行数据挖掘,同时保证数据挖掘的精度不受很大影响。隐私保护的数据挖掘技术 PPDM(Privacy Preserving Data Mining)由 Agrawal 在 2000 年首次提出[15],经过十年的研究已经产生了大量的方法。PPDM 按照数据的隐藏技术分,可分为基于同态加密、基于不经意传输和基于安全多方计算的方法等。

数据加密技术是用某种算法对数据进行加密,攻击者如果强行破译密码需要很大的代价,从而保护数据的隐私安全。虽然在数据挖掘时对数据进行加密可以提高数据安全性,但由于需要处理海量的加密数据,计算代价提高,降低了数据挖掘的效率。由此产生了同态加密技术[31],它使得加密后的数据可以进行与原始数据一样的代数运算,运算的结果还是加密数据,并且该结果就是明文经过同样的运算再加密后的结果。这项技术可以用于加密数据的检索、比较等操作,无需对数据解密就能得出正确的结果。

不经意传输 OT(Oblivious Transfer)协议是一种可保护隐私的通信协议,它最早由 Rabin 提出[32]。它的思想是接收者以一定概率得到发送者发出的某些消息,从而可以在通信的过程中保护双方的隐私。OT 协议最初由 1 个消息的传输,发展到 2 选 1 消息的传输,随后扩展至 n 选 1 不经意传输[33,34],即发送者发送 n 个消息,接收者只能以一定概率收到其中的 1 个,而发送者不知道接收者收到哪一个消息。这一协议可以使用在 PPDM 中,比如 Yehuda Lindell 提出了一种基于不经意传输的隐私保护分类挖掘[35]。

安全多方计算 SMC(Secure Multi-Party Computation)最早由姚期智提出[36]。它是指多个参与方需要用各自的秘密数据进行一项协同计算,在保证每个参与方得到的计算结果正确性的同时,保护每个参与方的秘密数据不被泄露。安全多方计算被用于数据挖掘中,达到保护隐私的目的。比如,文献[37,38]分别提出了基于 SMC 的 K-means 聚类方法。文献[39]提出了一种隐私保护的水平分布数据的关联规则两方挖掘方法。文献[40]提出了一种高性能的安全多方计算的框架,用于数据挖掘应用。文献[41,42]提出了基于同态加密 SMC 协议的 ID3 和 C4.5 算法。

3 大数据安全性评估

大数据的安全技术是否有效,能否阻挡黑客的攻击,需要相应的评估手段来验证。如前所述,大数据安全的两个重要方面是数据的可信性和隐私保护。因此评估大数据的安全性也可从这两个方面入手,即数据的可信性和隐私保护程度。如图 3 所示,数据的可信性主要包括相关性、准确性、及时性、完整性、一致性、有效性等;数据的隐私保护程度主要可以从差异度、方差、信息熵、匿名化程度、数据泄露风险度等方面来计算。本节就对数据的可信性和隐私保护程度的相关评估指标进行论述。

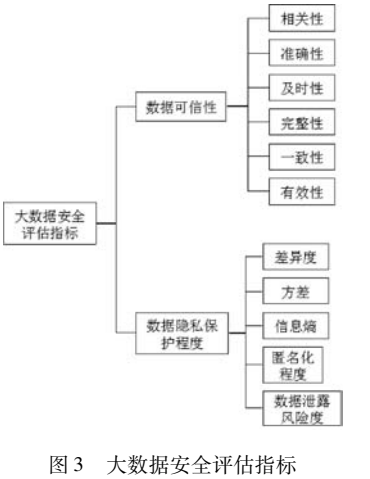

图 3 大数据安全评估指标

3.1 数据的可信性

数据可信性可以在许多方面进行定义,并与不断变化的用户需求有关。同一个数据的可信性可能被一个用户所接受而另

一个用户无法接受,在 2010 年可信的数据可能在 2013 年是不可信的。通常会参照高质量的数据特征来分析数据是否可信,一般通过表 1 中所述的几个方面评估数据可信性。

表 1　数据可信性指标

指标名称	说明
相关性	指数据是否满足其用户的需求
准确性	指数据能否反映基本现实,并且有足够的精确度
及时性	指数据是否是当前最新的数据
完整性	指数据是否包含了用户需要的所有信息
一致性	指数据之间能否很好地整合在一起,并保持与数据本身一致
有效性	指数据是真实的,并且可以满足相关方面的标准诸如准确性、及时性、完整性和安全性

数据可信性差的一个必然结果是,用这些数据得出结论并做出决策会产生风险。这些数据用于指定的用途时也可能会产生意想不到的后果,导致实际损失。

3.2　数据的隐私保护程度

前述的数据可信性的评估指标主要用于定性地评估数据来源是否可靠,其衡量标准可能会随着时间和需求而变化。本节所述的数据的隐私保护程度指标则是定量地评估处理后的数据的质量和隐私保护程度。2.1 节已经介绍了一些基于隐私保护的数据处理方法,不同的数据处理方法有不同的评估指标,这些指标分别从不同的角度来衡量隐私保护的效果。现举其中有代表性的几种评估指标加以论述,包括差异度、方差、信息熵、匿名化程度、数据泄露风险度,具体每项指标的计算方法和说明如表 2 所述。

表 2　数据隐私保护程度指标

指标名称	计算公式	公式说明	关注点
差异度	$\mathrm{Diss}(D,D')=\dfrac{\sum_{i=1}^{n}\lvert f_D(i)-f_{D'}(i)\rvert}{\sum_{i=1}^{n}f_D(i)}$	i 是原数据集 D 中的一个数据项,$f_D(i)$ 是数据项 i 在数据集 D 中出现的频率,$f_{D'}(i)$ 是 i 对应的处理后数据在处理后的数据集 D' 中出现的频率	评估数据信息损失程度,适用于评估数据扰乱技术的保护效果
方差	$\dfrac{Var(X-Y)}{Var(X)}=\dfrac{\frac{1}{N}[(X_i-Y_i)-(\bar{X}-\bar{Y})]^2}{\frac{1}{N}\sum_{i=1}^{N}[X_i-\bar{X}]^2}$	其中 X 表示数据的原始值,Y 表示扰乱后的值,N 表示数据数量,和表示 X 和 Y 的平均值	评估乘性噪声扰乱技术的保护效果
信息熵	$h(X)=-\sum p(x)log_2(p(x))$	X 是一个随机变量,它根据概率分布 $P(X)$ 在一个有限范围内取值	评估一个数据值的不可预测性,即预测经过隐私保护处理的数据的原值的难度
匿名化程度	$1/P(r(QI)\mid r'(QI))$	假设数据集合 D 被匿名化为数据集合 D':变量 r 是 D 中的任意一条数据记录,变量 r' 则是 D' 中 r 的匿名化形式,$r(QI)$ 代表数据记录 r 中的准标识数据,P 是可能从 $r(QI)$ 中推测出 $r'(QI)$ 的概率	评估从匿名化的数据中推测出原始数据的难易程度
数据泄露风险度	$HF=\dfrac{\#R_P(D')}{\#R_P(D)}$	表示从原始数据集 D 中发现的敏感数据,表示从处理后数据库 D' 中发现的敏感数据的数量	评估某条信息和一个特定的个人相关联的风险度

上述指标有些是基于传统的统计学方法,如基于差异度、方差和数据泄露风险度的指标;有些和特定的隐私保护技术相关,如匿名化程度;有些利用了信息论理论,如信息熵。因此这些指标适合用于不同的场合。

差异度反映了经过隐私保护处理后的数据集与原数据集的相似程度,由 Bertino 等人在文献[43]提出。它能够衡量数据信息损失程度,适用于评估数据扰乱技术的保护效果。差异度越小,信息损失越少,数据质量越好,但同时隐私保护程度越小。这是比较普适的指标,因为它的测量不需要对所分析的数据集作很多假设。

方差适用于评估乘性噪声扰乱技术的保护效果。方差越大,表示扰乱后的值与原数据差异越大,隐私保护程度也就越好,但相应的数据可用性就越低。

信息熵由 Bertino 等人提出[43],这个方法的基础是由香农定义的。信息熵用来度量数据取值的不确定程度,因此它可以用来评价一个数据值的不可预测性,即预测经过隐私保护处理的数据的原值的难度。因为熵表示数据的信息量,所以数据经过隐私保护处理之后的熵应该比之前的熵要高。信息熵是一种较通用的测量数据隐私级别的方法,它越大表明隐私保护程度越好。对于不同的隐私保护方法,需要根据不同方法的特性重新定义计算方法,这和不同隐私保护算法有关。在文献[43]中,信息熵被用来评价基于关联规则的隐私保护算法。

匿名化程度适用于评估匿名方法的保护效果。数据匿名方法主要针对数据的准标识属性(可唯一确定一条记录的一组属性)执行隐去或泛化的操作。匿名化程度用来度量从匿名化的数据中推测出原始数据的难易程度。一个好的匿名化方法应该使得用户难以从匿名化的数据中推测出原始的敏感关联。

数据泄露风险度适用于评估 PPDM 的隐私保护效果。有些 PPDM 算法允许使用者选择隐藏敏感信息的数量,因此数据泄露风险度可以通过计算隐藏失效参数来评估。它被 Oliveira 和 Zaiane 定义为在处理后的数据集中被发现的敏感信息的百分比[44]。数据泄露风险度表示某条信息和一个特定的个人相关联的风险度,所以数据泄露风险度越大,则隐私保护程度越低。

需要指出的是,在实际应用中往往仅凭单个指标难以全面

衡量大数据应用的安全性,因而需要用多种指标来综合评估。有些指标的计算方法也可能需要根据实际情况作出一些调整。

4 结 语

本文在梳理大数据安全研究现状的基础上,从数据和计算两个层面上分析了大数据所面临的安全问题,阐述了目前主要的解决大数据安全问题的研究成果。最后针对大数据安全性的评估提出了一些评估指标,并对这些指标的适用性作了比较。

参 考 文 献

[1] Temple K. What Happens in an Internet Minute? [EB/OL]. http://scoop. intel. com/what - happens - in - an - internet - minute/.

[2] Feng Dengguo, Zhang Min, Li Hao. Big Data Security and Privacy Protection[J]. Chinese Journal of Computers, 2014, 37(1): 246 - 258.

[3] Miller H E. Big-data in cloud computing: a taxonomy of risks[J]. Information Research, 2013, 18(1).

[4] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107 - 113.

[5] Duhigg C. How companies learn your secrets[EB/OL]. http://128. 59. 177. 251/twiki/pub/CompPrivConst/HowCompaniesLearnOurConsumingSecrets/How_Companies_Learn_Your_Secrets_ - _NYTimes. com. pdf.

[6] Barbaro M, Zeller T, Hansell S. A face is exposed for AOL searcher no. 4417749[EB/OL]. http://w2. eff. org/Privacy/AOL/exhibit_d. pdf.

[7] Bhat S, Damle S, Chaudhari P, et al. KERBEROS: An Authentication Protocol[J]. International Journal, 2014, 2(2): 200 - 204.

[8] Roy I, Setty S T V, Kilzer A, et al. Airavat: Security and Privacy for MapReduce[C]//USENIX Conference on Networked Systems Design and Implementation, 2010, 10: 297 - 312.

[9] Dwork C, Roth A. The algorithmic foundations of differential privacy [J]. Theoretical Computer Science, 2013, 9(3 - 4): 211 - 407.

[10] Xu G, Wang Z, Yang L, et al. Research of Data Provenance Semantic Annotation for Dependency Analysis[C]//Advanced Cloud and Big Data, 2013 International Conference on. IEEE, 2013: 197 - 204.

[11] Bonatti P A, Hogan A, Polleres A, et al. Robust and scalable linked data reasoning incorporating provenance and trust annotations[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2011, 9(2): 165 - 201.

[12] Groth P. Provenance and Annotation of Data and Processes[C]//4th International Provenance and Annotation Workshop, Santa Barbara, CA, USA, June 19 - 21, 2012, Revised Selected Papers. Springer, 2012.

[13] Ming G, Cheqing J, Xiaoling W, et al. A survey on management of data provenance[J]. Chinese Journal of Computers, 2010, 33(3): 373 - 389.

[14] Muniswamy-Reddy K K, Holland D A, Braun U, et al. Provenance-Aware Storage Systems[C]//USENIX Annual Technical Conference, General Track, 2006: 43 - 56.

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

[16] Kantarcioglu M, Vaidya J, Clifton C. Privacy preserving naive bayes classifier for horizontally partitioned data[C]//IEEE ICDM Workshop on Privacy Preserving Data Mining, 2003: 3 - 9.

[17] Chen K, Liu L. Geometric data perturbation for privacy preserving outsourced data mining[J]. Knowledge and Information Systems, 2011, 29(3): 657 - 695.

[18] Islam M Z, Brankovic L. Privacy preserving data mining: A noise addition framework using a novel clustering technique[J]. Knowledge-Based Systems, 2011, 24(8): 1214 - 1223.

[19] Chhinkaniwala H, Garg S. Tuple Value Based Multiplicative Data Perturbation Approach To Preserve Privacy In Data Stream Mining[J]. International Journal of Data Mining & Knowledge Management Process, 2013, 3(3): 53 - 61.

[20] Patel A, Dodiya K, Pate S. A Survey On Geometric Data Perturbation In Multiplicative Data Perturbation[J]. International Journal of Research in Advent Technology, 2013, 1(5): 603 - 607.

[21] Oganian A. Multiplicative noise protocols[C]//Privacy in Statistical Databases. Springer Berlin Heidelberg, 2011: 107 - 117.

[22] Keyur D, Shruti Y. Classification Techniques For Geometric Data Perturbation in Multiplicative Data Perturbation[J]. International Journal of Engineering Development and Research, 2014, 2(2): 2380 - 2383.

[23] Rajalaxmi R R, Natarajan A M. A Novel Sanitization Approach for Privacy Preserving Utility Itemset Mining[J]. Computer and Information Science, 2008, 1(3): 77.

[24] Lee J, Ko H J, Lee E, et al. A Data Sanitization Method for Privacy Preserving Data Re-publication[C]//Networked Computing and Advanced Information Management, NCM'08. Fourth International Conference on. IEEE, 2008, 2: 28 - 31.

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

[26] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571 - 588.

[27] Ajayi O O, Adebiyi T O. Application of Data Masking in Achieving Information Privacy[J]. Innovative Systems Design and Engineering, 2014, 5(1): 27 - 35.

[28] Patel B R, Maheta J B. Survey on Privacy Preservation Technique: Data Masking[C]. International Journal of Engineering Research and Technology. ESRSA Publications, 2014, 3.

[29] Komishani E G, Abadi M. A generalization-based approach for personalized privacy preservation in trajectory data publishing[C]//Telecommunications (IST), 2012 Sixth International Symposium on. IEEE, 2012: 1129 - 1135.

[30] Hajian S, Domingo-Ferrer J, Farras O. Generalization-based privacy preservation and discrimination prevention in data publishing and mining[J]. Data Mining and Knowledge Discovery, 2014, 28(5): 1158 - 1188.

[31] Paillier P. Public-key cryptosystems based on composite degree residuosity classes[C]//Advances in Cryptology—EUROCRYPT'99. Springer Berlin Heidelberg, 1999: 223 - 238.

[32] Rabin M O. How To Exchange Secrets with Oblivious Transfer[EB/OL]. IACR Cryptology ePrint Archive, 2005. http://eprint. iacr. org/2005/187. pdf.

[33] Vasant S, Venkatesan S, Rangan C P. A code-based 1-out-of-n oblivious transfer based on mceliece assumptions[M]. Information Security Practice and Experience. Springer Berlin Heidelberg, 2012: 144 - 157.

[34] Corniaux C L F, Ghodosi H. A Verifiable 1-out-of-n Distributed Oblivious Transfer Protocol[J/OL]. IACR Cryptology ePrint Archive, 2013, https://eprint. iacr. org/2013/063. pdf.

[35] Lindell Y, Pinkas B. Privacy preserving data mining[C]//Advances in Cryptology—CRYPTO 2000. Springer Berlin Heidelberg, 2000: 36 - 54.

(下转第 71 页)

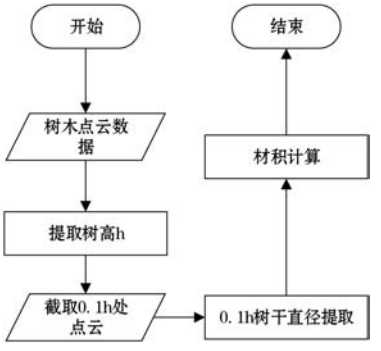

图 4　材积提取流程图

3　功能测试

树木点云数据处理软件是根据树木点云数据的空间结构等信息利用计算机可视化技术,以及数据处理算法实现树木结构的空间展示以及树木结构参数的求解。软件系统采用 Qt 作为图形界面开发工具并采用标准 C ++ 语言进行开发,所以该系统具有跨平台的特性,可以在多种操作系统上编译使用。

本文利用地面激光雷达扫描的块林木区域的树木点云数据对系统进行验证。该点云数据文件大约 1.3 GB。每个点包含了位置信息,颜色信息,扫描角度,反射强度等 11 个数据项。本系统在主频为 3.2 GHz 的英特尔酷睿双核,4 GB 内存的 PC 上可以流畅查看点云数据,并可以成功提取树木的胸径、树高、材积这三种树木结构参数。其运行效果如图 5 所示。

图 5　系统三维显示与参数提取结果显示图

4　结　语

本文设计的三维树木点云处理软件系统结合计算机可视化技术与雷达数据处理技术,可以快速有效地提取树木结构参数。可以快速准确地进行林业清查满足林业信息化的要求,同时可以对人力难以到达的林木区域进行林木测量,节省了人力物力。后续将继续研究树木定位,树木三维模型重建等工作,以使得本系统更加完善。

参　考　文　献

[1] 黄明,王晏民,付昕乐,等. 地面激光扫描数据处理系统的设计与实现[J]. 测绘通报,2014(8):55 - 58.
[2] 王炎松,谢飞. 古建保护对于三维激光扫描点云数据处理软件系统的用户需求——以古建测绘中的数据处理为例[J]. 华中建筑,2008,26 (4):130 - 132.
[3] 王晏民,王国利. 地面激光雷达用于大型钢结构建筑施工监测与质量检测[J]. 测绘通报,2013(7):39 - 42.
[4] 张腾波,罗德安,黄鹤,等. 基于地面激光雷达的土遗址保护研究[J]. 新探索,2013(4):67 - 72.
[5] 黄洪宇,陈崇成,邹杰,等. 基于地面激光雷达点云数据的单木三维建模综述[J]. 林业科学,2013,49 (4):123 - 130.
[6] 李丹,庞勇,岳彩荣,等. 基于 TLS 数据的单木胸径和树高提取研究[J]. 北京林业大学学报,2012,34(4):79 - 86.
[7] 李丹,庞勇,岳彩荣. 地基激光雷达在森林参数反演中的应用[J]. 世界林业研究,2012,25(6):34 - 39.
[8] Pueschel P,Newnham G,Rock G,et al. The influence of scan mode and circle fitting on tree stem detection,stem diameter and volume extraction from terrestrial laser scans[J]. ISPRS Journal of Photogrammetry and Remote Sensing,2013,77:44 - 56.
[9] 晏海平,吴禄慎,陈华伟. 基于 VC 和 OpenGL 的三维点云处理软件系统设计[J]. 计算机应用与软件,2014,31(6):177 - 180.
[10] 莫建文,邹路路,首照宇,等. 跟踪雷达三维场景显示系统的设计与实现[J]. 计算机应用与软件,2014,31 (5):141 - 144.
[11] 李中志,汪学刚. 基于 COM 技术的雷达数据处理软件系统设计[J]. 计算机应用与软件,2010,27(5):27 - 29.
[12] 李嘉,胡怀中,胡军,等. 可视化三维图形库 Visualization ToolKit3.2 的原理及应用[J]. 计算机应用与软件,2004,21 (2):5 - 6.
[13] 杨钦,徐永安,翟红英. 计算机图形学[M]. 清华大学出版社,2005.
[14] 王丽辉,袁保宗. 鲁棒的模糊 C 均值和点云双边滤波去噪[J]. 北京交通大学学报:自然科学版,2008,32(2):18 - 21.
[15] 姚宏忠,何军,刘祎. 一种基于 kd 树的实时大规模地形可视化算法[J]. 科学技术与工程,2012 12(2):338 - 341.
[16] 孟宪宇,测树学[M]. 北京:中国林业出版社,2006.
[17] 杨华,孟宪宇,程俊,等. 利用正形数估测立木材积方法的研究[J]. 林业资源管理,2005(1):39 - 41.

(上接第 38 页)

[36] Yao A C C. How to generate and exchange secrets[C]//Foundations of Computer Science,1986,27th Annual Symposium on. IEEE,1986:162 - 167.
[37] Beye M,Erkin Z,Lagendijk R L. Efficient privacy preserving k-means clustering in a three-party setting[C]//Information Forensics and Security,2011 IEEE International Workshop on. IEEE,2011:1 - 6.
[38] Zhukov V G,Vashkevich A V. Privacy-preserving Protocol over Vertically Partitioned Data in Multiparty K-means Clustering[J]. Middle-East Journal of Scientific Research,2013,17(7):992 - 997.
[39] Zhang F,Rong C,Zhao G,et al. Privacy-Preserving Two-Party Distributed Association Rules Mining on Horizontally Partitioned Data[C]// Cloud Computing and Big Data (CloudCom-Asia),2013 International Conference on. IEEE,2013:633 - 640.
[40] Bogdanov D,Niitsoo M,Toft T,et al. High-performance secure multiparty computation for data mining applications[J]. International Journal of Information Security,2012,11(6):403 - 418.
[41] Xiao M J,Huang L S,Luo Y L,et al. Privacy preserving id3 algorithm over horizontally partitioned data[C]//Parallel and Distributed Computing,Applications and Technologies,2005. PDCAT 2005. Sixth International Conference on. IEEE,2005:239 - 243.
[42] Xiao M J,Han K,Huang L S,et al. Privacy preserving C4.5 algorithm over horizontally partitioned data[C]//Grid and Cooperative Computing,2006. Fifth International Conference. IEEE,2006:78 - 85.
[43] Bertino E,Fovino I N,Provenza L P. A framework for evaluating privacy preserving data mining algorithms[J]. Data Mining and Knowledge Discovery,2005,11(2):121 - 154.
[44] Oliveira S R M,Zaiane O R. Privacy preserving frequent itemset mining[C]//Proceedings of the IEEE International Conference on Privacy,Security and Data Mining-Volume 14. Australian Computer Society,Inc.,2002:43 - 54.