

基于云计算的景区数据仓库应用研究

谢福伟,梁昌勇,马银超

(合肥工业大学 管理学院,安徽 合肥 230009;
过程优化与智能决策教育部重点实验室,安徽 合肥 230009)

摘要:云计算、物联网、大数据等新兴信息技术的发展与应用在提高景区信息化服务水平的同时,也对景区海量信息资源的有效利用提出了严峻挑战。面对超大规模、非结构化的海量数据,传统的基于关系型数据库的数据仓库已很难有效支持景区的数据存储与分析工作。基于此文中提出了一种基于云计算技术的景区数据仓库,通过采用 HDFS 对数据进行分布式存储管理,利用 MapReduce 设计海量数据的分析模式,使用 HiveQL 语言实现数据仓库与前端表现层的交互,能够有效解决景区海量数据的数据管理问题。以黄山风景区为实际背景的实验结果表明了该数据仓库的正确性和有效性。

关键词:云计算;数据仓库;MapReduce;ETL

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2014)09-0198-04

doi:10.3969/j.issn.1673-629X.2014.09.046

Research on Data Warehouse Application of Tourist Areas Data Based on Cloud Computing

XIE Fu-wei, LIANG Chang-yong, MA Yin-chao

(School of Management, Hefei University of Technology, Hefei 230009, China;

Key Laboratory of Process Optimization and Intelligent Decision-making of Ministry of
Education, Hefei 230009, China)

Abstract: The emergence of new information technologies, such as cloud computing, internet of things, big data, etc, greatly enhances the level of area of information technology services. However, how to effectively utilize the scenic area of information resources is a great challenge. Faced large scale and unstructured mass data, the data warehouse based on the traditional relational database has been difficult to effectively support the data storage and analysis in scenic area. Based on this, propose a scenic area data warehouse based on cloud computing technology, adopting HDFS for distributed storage of data, using MapReduce to design massive data analysis model, with HiveQL language to implement the interaction between data warehouse and front-end presentation layer, which can solve the data management problem of massive data in scenic area. Taking Huangshan as example, the experimental results indicate the data warehouse is correct and feasible.

Key words: cloud computing; data warehouse; MapReduce; ETL

0 引言

随着信息化技术在旅游景区管理中的广泛应用,越来越多的景区建立起了自己的各种信息系统,这些信息系统为景区提供了海量信息资源,大大提高了景区的服务质量和管理水平。同时,为了从这些信息资源中挖掘、获取有效信息,以为景区管理和服务工作提供决策支持,很多景区建立了对信息系统资源进行统一存储、管理的数据仓库。然而,随着云计算、物联网等新兴信息技术的飞速发展与应用,人类社会的数据

种类和规模呈指数级增长,大数据的到来势不可挡^[1],传统的基于关系型数据库的景区数据仓库体系已很难有效管理 PB 级别的大数据,而其较高的数据移动代价、不灵活的适应变化能力等弊端也很难适应当下景区的大数据管理需求^[2]。基于此有必要对大数据时代景区数据仓库体系的重构进行研究和实践。

Hadoop 是由 Apache 基金会开发的可运行在大量低端商用 PC 机集群的云计算技术,可以在有效解决数据仓库的源数据超大规模膨胀问题的同时,提供高

收稿日期:2013-11-15

修回日期:2014-02-17

网络出版时间:2014-07-17

基金项目:国家自然科学基金重点项目(71331002);智慧景区客流量预测系统项目(10120106011)

作者简介:谢福伟(1989-),男,河南信阳人,硕士研究生,研究方向为云计算;梁昌勇,教授,博士,博士生导师,研究方向为决策、云计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140717.1226.002.html>

可扩展性的计算能力和存储能力^[3-4]。基于 Hadoop 的分布式文件系统 (Hadoop Distributed File System, HDFS)^[5]以流式数据访问模式来存储超大规模的文件,它采用的 MapReduce^[6]分布式编程架构,能够对海量数据集进行高效并行分布式处理,与传统的处理模式相比简单易用,且具有自动进行负载均衡、高容错性等优点。而建立在 HDFS 基础上的 Hive^[7]数据仓库,具有很高的可扩展性,对事务性要求不高,在性能方面可面向千万行级以上海量数据的查询与分析^[8]。鉴于以上分析,基于 Hadoop 云计算技术的 HDFS 和 Hive 实现的数据仓库,不仅可以为景区海量数据提供全新的、高效的、可扩展的分布式数据存储中心,解决海量数据存储问题,而且利用云计算技术的动态可扩展特性,能够大大提高数据处理效率。

1 基于云计算的景区数据仓库体系结构

基于云计算的景区数据仓库体系结构主要由四层组成:分析层、数据仓库层、协调层、原始数据层,如图 1 所示。

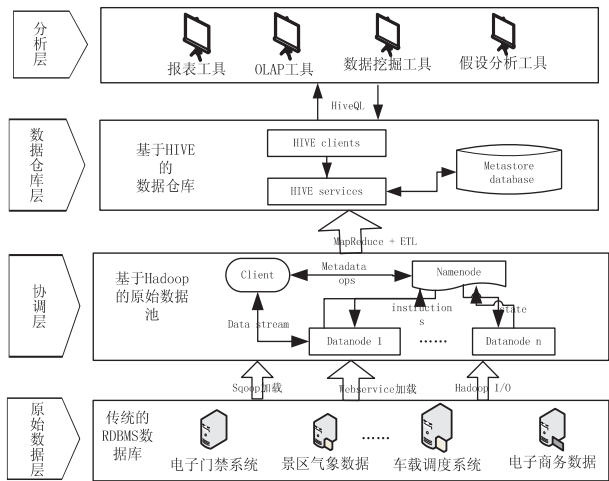


图1 景区客流预测数据仓库体系结构图

具体功能描述如下:

(1)原始数据层。云计算技术近年虽然得到了长足的发展,但目前旅游景区的信息系统基本上依旧采用的是传统的信息系统架构。基于云的景区数据仓库在很长一段时间的原始数据信息依旧需要从各个关系型数据库中获取。

(2)协调层。在基于 RDBMS 的数据源层和基于 Hive 的数据仓库层之间加入原始数据池,通过 Sqoop、Web Service 等技术从各源数据库中提取数据,并对其进行实体化集成和数据源清理,使得数据类型和格式统一,数据增量更新步调一致,然后在此基础上再进行数据的提取、转换、加载到数据仓库层。

(3)数据仓库层。在构建基于 Hive 的数据仓库过程中,建立数据文件和 Hive 表结构的关联关系,从而

既去掉了大量数据传至 Hive 的处理过程,也得以实现分布式存储数据,更好地支持业务数据的分析过程。

(4)分析层。根据各级决策者的需求,从数据仓库中提取相关的数据,然后确定数据分析的方法,并把分析结果通过前端展示工具提供给决策者。其中的功能包括提供报表、即席查询、OLAP 多维分析、数据挖掘、查询统计等。

2 基于云技术的 ETL 研究

ETL 是专门解决数据仓库同质化和数据清洗、加载问题的一类工具^[9],它由从数据源抽取数据、按需求转换数据、加载数据到数据仓库三个阶段组成。整个 ETL 过程设计作为构建数据仓库最复杂的工作之一,其工作量占整个数据仓库建设工程的 40% 到 60%^[10]。而在整个 ETL 过程中数据清洗的主要工作是对数据奇异值进行处理,比如空值处理;数据转换阶段主要解决将原始数据转换成数据仓库基石的事实表过程;而数据加载更多的是通过对数据事实表的聚类、聚集等操作生产所需的数据报表。文中研究基于 Hive 和 MapReduce 的 ETL 方法来处理数据,并主要分析了奇异值处理、事实表生成和数据聚集三个 ETL 关键问题。

(1)奇异值处理。

在数据预处理中,原始数据记录中一些属性值为空值等奇异值可能会影响后续数据挖掘算法的识别和处理,必须对属性值为空的进行填写。空值处理的原始数据记录中的空值将由指定的特定值或者有相关算法填补的值进行替换。通过用多个计算节点同时对这些块中空缺值进行填充,实现并行高效处理,加快填充速度。通过使用 Hive 中的条件函数 if (value = null, newValue, value) 填充空值,类似地也可以对数值进行最大最小值限定和数据分频操作,由于这些操作未对原始数据进行聚集等操作故无需 Reduce 过程。

(2)事实表生成。

在构建数据仓库的过程中,需要建立事实表和维度表,而事实表和维度表通常是通过主外键关联。因此,根据原本的关联模式,在建立一个数据仓库时需要根据维表对原表中与之关联的值用维表 id 替换,生成事实表,以方便从不同的维度和粒度进行数据处理和分析。并行事实表生成的基本思路:两表原始数据记录分块,使两个表中可数据连接的记录以被分配到相同的计算节点中,多个计算节点同时在两个表的记录连接操作,可以在使用 Hive 中的 left join 语句后用维表 id 替换源表中相关的属性,最后将替换 id 后的数据加载到新表中,以此实现事实表的生成。主要 MapReduce 设计如图 2 所示。

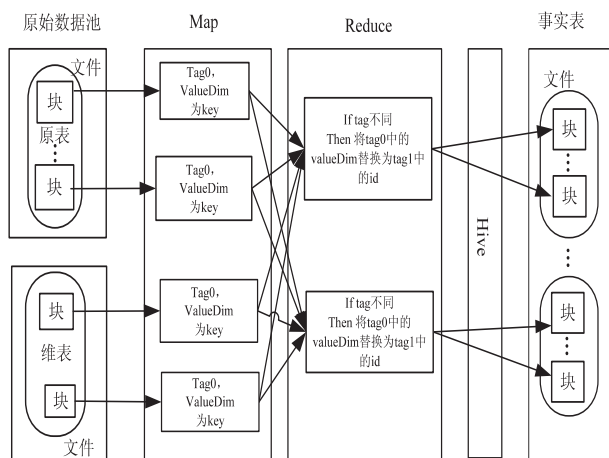


图 2 事实表生成

①Map 根据表的元数据和 NameNode 中的元数据,将源表和维表中对应数据块中的信息一一获取,生成 Map 类型能够识别的<key,value>键值对,并标识源表和维表的数据行号;

②根据表的元数据,解析 value 值,取出各列的属性值,将源表中与维表关联的属性值作为 Map 的 key 值,同时,也将维表中该属性作为 Map 的 key 值;

③将处理结果形成 Map 的输出<key,value>,包括标签 tag,其中 key 包含表的关联属性值,value 为包括其余属性值的对象,tag 标明了数据行来源;

④Reduce 收集具有相同 key 值的数据,将 tag 标签不同的数据进行连接,用维表中的 id 替换源表中的相关属性值,形成 Reduce 的输出<key,value>,其中 key 为空文本,value 为处理过后的每行文本,包含表的所有列值;

⑤将 Reduce 处理后的结果输出到 HDFS 中新表所在目录,即为事实表。

(3) 数据聚集。

通过对沉淀了大量历史数据的原始数据库的数据提取和分析关联操作后生成了事实表,但事实表是按数据的最新粒度来组织的数据。在实际的数据挖掘应用中需要不同的粒度和维度的数据,因此在构建数据仓库时,聚集生成可用报表是一个必不可少的环节。聚集其实质就是用 group by 操作对选取的特定维度用所需的度量方式对表中的数据进行整合的操作。并行数据聚集的基本思路是:将数据按所选的维度分组,不同组的数据分配至不同计算节点,各节点因数据互不相关,故可以同时使用 Hive 的 group by 语句和聚集函数进行数据聚集。

3 实验研究

黄山风景区是世界自然文化遗产、国家首批 5A 级风景区,景区信息化建设是全国旅游风景区信息化

建设示范单位,已建成了 27 个应用系统,正在开展基于云计算技术的“黄山云”规划和建设,而笔者所在实验室在前期合作项目“智慧黄山景区客流预测系统”的基础上,已参与“黄山云”的建设工作。文中利用原客流量预测项目的数据仓库为依托,建立基于云的景区数据仓库并在实际环境下运行,最终验证了该架构数据仓库的切实有效性。

3.1 数据仓库设计

在收集景区可提供的各信息系统数据的基础上,通过对景区客流量预测的影响因素分析,最终选择了电子门禁系统、电子商务系统、景区气象中心三大景区系统的数据。首先从数据源中分析并选择了检票和订票两个事实,依据数据源所含信息建立每个事实的属性树,其中属性树的每个节点对应着一个数据源模式属性。但是建好的事实的属性树所包含的属性不一定是客流预测这一目标所需要的,需结合实际需求对属性树进行修剪和移植,最终得出符合实际需求的属性树。为了在决策过程中显著地聚合相关事实,对属性树中有价值的属性进行离散化处理作为维度,对于可以标识事实的根子节点属性作为度量。在本数据仓库设计中选择检票数作为检票事实的度量,选择订单人数作为订单事实的度量,选择检票口、日期天气等作为维度,最后将属性树转换成如图 3 所示的包含以上维度和度量的事实模式。

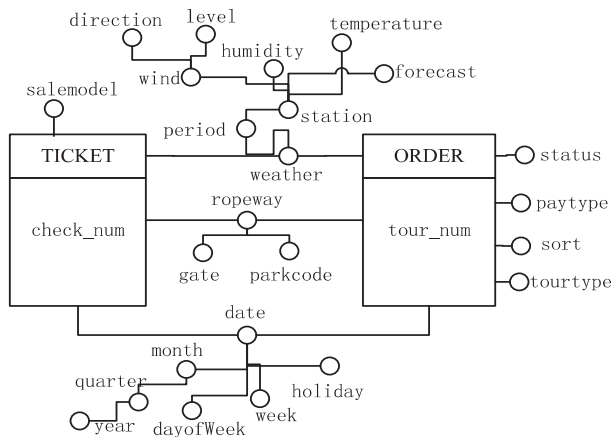


图 3 事实模式图

3.2 数据仓库实现

本实现平台是选取 6 台普通计算机主机利用 Hadoop 技术集群搭建而成的,其中性能较好的一台作为主机 Master 实现 NameNode 和 JobTracker 的功能,另外 5 台作为从机 Worker,实现相应的 DataNode 和 TaskTracker 的功能。整个搭建过程如下:

(1)节点设置。为了让集群网络中各计算节点可以互相通信传输数据等,需要修改配置各节点主机的/etc/hosts 文件添加集群中所有节点的 ip 和主机名。

(2)SSH(Secure Shell)设置。在 Hadoop 云计算平

台中,可以通过软件 SSH 实现各节点之间执行指令时无须输入登录密码。首先在 NameNode 节点上执行命令“`sudo apt-get install openssh-server`”来安装该软件,然后在/home 下创建 .ssh 目录,通过命令“`ssh-keygen - t rsa`”生成密钥,最后在 NameNode 主机上使用 scp 安全复制指令将 SSH 授权密钥 authorized_keys 复制到每台 DataNode 主机上。

(3)环境变量设置。因为 Hadoop 云计算基础架构是由 Java 语言编写而成的,所以必须要在 Hadoop 的 `hadoop-env.sh` 环境配置文档中配置 Java 环境变量,同时也需要配置自身环境变量。

(4)Hadoop 系统设置。Hadoop 系统设置主要包括 3 个部分,分别是 Hadoop 云计算系统设置文档、HDFS 设置文档和 MapReduce 程序设置文档^[11]。完成上述三个文档的设置之后,将 Master 中设置完成的 Hadoop 拷贝到其他 Slaves 主机上。

(5)Hadoop 启动。在启动一个新的 Hadoop 云计算系统之前,必须先对系统的 NameNode 进行格式化,用“`hadoop namenode -format`”命令让 Master 主机正式成为 NameNode 主机。最后用“`start-all.sh`”命令就可以正常启动 Hadoop 云计算系统。

(6)基于 Hive 的数据仓库创建。文中数据仓库的构建是基于 Hive 数据仓库工具,根据数据仓库架构逻辑模型的设计,创建相关的 Hive 表。Hive 表结构的定义与 SQL 类似,只要根据设计中的字段逐一定义表结构即可。

3.3 实验结果及分析

数据仓库作为商务智能的基石,其主要作用是为商务智能的数据挖掘和联机分析处理模块提供全面准确的数据支持。文中以“智慧黄山景区客流量预测系统”为案例从数据挖掘和联机分析处理两方面的应用情况来验证该数据仓库架构设计的有效性和可行性。

数据挖掘应用,利用数据仓库中的客流信息进行客流量影响因素相关性分析,选出八点前客流量、历史同期客流量、星期、节假日、天气等对本日客流量预测影响权重大的数据组成数据集,采用 SVR(支持向量回归)^[12]和 BP(神经网络)^[13]构建预测模型,并用 PSO(粒子群算法)对 SVR 模型的自由参数(c, ϵ, δ)进行优化以提高预测精度^[14]。该项目通过 SVR 的软件包 lib_svm.jar 将 SVR 预测模型嵌入系统,通过 Matlab 的 Matlab Builder for Java 工具和 MCRInstaller 编译器实现将 BP 预测模型集成至系统中,从而使得基于该数据仓库的客流预测模型能够在景区客流预测系统中得以实际应用;另外预测模型从 2012 年 4 月使用以来的月均误差率如表 1 所示,其月均误差率均低于 15%(黄山景区对本日预测模型精度的验收指标为月

均误差低于 15%),满足景区对预测精度的实际要求。可见本数据仓库在实际的数据挖掘应用中是有效可行的。

表 1 本日预测月均误差率

月份	SVR 模型/%	BP 模型/%
4	11.67	10.91
5	8.55	12.6
6	9.68	14.67
7	7.12	8.38
8	12.88	13.86
9	11.91	12.89
10	10.29	11.27
11	12.65	13.62
12	14.61	14.63

联机分析处理,项目通过采用开源的 JPivot+Mondrian^[15]架构实现对本数据仓库的多维分析处理。整个 OLAP^[16]系统由存储层、集合层、维度层和表现层组成,存储层负责创建集合的单元数据;集合层起到缓存的作用;维度层负责解析、验证和执行 MDX(多维表达式)来生成多维报表所需数据的形式;表现层使用 XML/XSLT 渲染 OLAP 报表。实验基于文中提出的数据仓库为基石,采用 JPivot+Mondrian 架构实现的日客流信息多维表展示图,该多维报表通过基于云的数据仓库实现了对电子门禁系统、电子商务系统等系统中相关客流量影响因素的集成展出,可以更直观地对其进行分析处理展示,为景区的服务工作提供支持。

4 结束语

文中针对传统的基于关系型数据库的景区数据仓库在大数据时代存在的海量数据存储和非结构化数据分析能力不足的问题,设计并实现了基于云计算技术的景区数据仓库系统结构。通过复用智慧黄山景区客流量预测系统中涉及到的黄山景区各信息系统的基础数据,验证了文中提出的数据仓库体系结构的可行性和有效性。

参考文献:

[1] 孟小峰,慈 祥.大数据管理:概念,技术与挑战[J].计算机研究与发展,2013,50(1):146-169.
[2] 王 珊,王会举,覃雄派,等.架构大数据:挑战,现状与展望[J].计算机学报,2011,34(10):1741-1752.
[3] White T.Hadoop:the definitive guide[M].[s.l.]:O'Reilly,2012.
[4] Iosup A,Ostermann S,Yigitbasi M N,et al.Performance analysis of cloud computing services for many-tasks scientific computing[J].IEEE Transactions on Parallel and Distributed

服务程序进行修改,就可实现整个系统的同步升级,使整个系统各模块之间的工作与更新保持独立、稳定。

4 结束语

随着物联网与移动互联网技术的发展,制造业企业也将会开启制造车间信息化的大潮,物联网技术与移动互联设备的互相融合将会是未来的主流。Android 操作系统给出了很好的提示,在移动设备上开发相应的智能程序来代替工厂传统的车间管理与数据采集模式^[13],可以简化传统车间信息交互的流程,达到生产信息上传下达的实时性要求;简易的组网、成熟工业产品的使用、相互独立的服务程序降低了系统中各元素、各服务模块之间的耦合程度,对于降低系统的维护成本,使系统的升级与更新变得更加简便快捷。

在快速变革的制造业信息化进程当中,制造业系统对新业务的实时更新能力是衡量系统的不可缺少的一个重要指标。如何解耦,如何提高系统的稳定性将会是该领域其中的一个重要课题。

参考文献:

[1] Wang M L,Dai Q Y,Zhong R Y,et al. RFID-enabled real-time mechanical workshop training center [J]. International Journal of Engineering Education,2012,28(5):1199-1212.

[2] 王美林,张湘伟,戴青云. 产学研相结合提高工程训练人才培养水平[J]. 广东工业大学学报(社会科学版),2010,10(S):7-9.

.....

(上接第 201 页)

Systems,2011,22(6):931-945.

[5] Baliga J,Ayre R W A,Hinton K,et al. Green cloud computing:balancing energy in processing, storage, and transport [J]. Proceedings of the IEEE,2011,99(1):149-167.

[6] Lee K H,Lee Y J,Choi H,et al. Parallel data processing with MapReduce:a survey [J]. ACM SIGMOD Record,2012,40(4):11-20.

[7] 李伟卫,李梅,张阳,等. 基于分布式数据仓库的分类分析研究[J]. 计算机应用研究,2013,30(10):2936-2939.

[8] 王德文,肖凯,肖磊. 基于 Hive 的电力设备状态信息数据仓库[J]. 电力系统保护与控制,2013,41(9):125-130.

[9] Vassiliadis P,Simitsis A,Skiadopoulos S. Conceptual modeling for ETL processes[C]//Proceedings of the 5th ACM international workshop on data warehousing and OLAP. [s. l.]: ACM,2002:14-21.

[10] Wang T,Hu J,Zhou H. Design and implementation of an ETL approach in business intelligence project[M]//Practical ap-

[3] 戴青云,钟润阳,张本军,等. 装备制造业 MES 系统的设计与实现[C]//2008 全国制造业信息化标准论坛论文集. 北京:制造业自动化,2008:78-81.

[4] 刘泽禧,戴青云,周科,等. 基于 MES 的嵌入式智能数据交互终端的设计[J]. 仪器仪表用户,2008,15(2):64-65.

[5] 王平,史少峰,卢超. 基于物联网技术的离散制造企业质量信息采集系统设计[J]. 中国制造业信息化:学术版,2012,41(23):5-8

[6] Komatineni S,MacLean D. Pro Android 4[M]. New York:Apress,2012.

[7] 钟润阳,戴青云,周科,等. 实时综合实验教学管理系统数据处理关键技术[J]. 现代制造工程,2008(12):122-126.

[8] 王朝华,陈德艳,黄国宏,等. 基于 Android 的智能家居系统的研究与实现[J]. 计算机技术与发展,2012,22(6):225-228.

[9] 钟润阳,戴青云,周科,等. 基于 RFID 的 Web 实时系统构建与实现[J]. 现代计算机:上下旬,2008(9):7-10.

[10] 胡星波,晏渭川. 基于 Android 的 NFC 实现与应用[J]. 电视技术,2011,35(21):84-88.

[11] 李佳,付强,丁宁. C#开发技术大全[M]. 北京:清华大学出版社,2009.

[12] 单东林,张晓菲,魏然. 锋利的 jQuery[M]. 北京:人民邮电出版社,2009.

[13] Bhattacharya S,Panbu M B. Design and development of mobile campus, an Android based mobile application for university campus tour guide [J]. International Journal of Innovative Technology and Exploring Engineering,2013,2(3):25-29.

.....

plications of intelligent systems. Berlin:Springer,2012.

[11] 刘鹏. 云计算[M]. 第2版. 北京:电子工业出版社,2011.

[12] Cheng Lu,Jiang Changsheng,Pu Ming. Online-SVR-compensated nonlinear generalized predictive control for hypersonic vehicles [J]. Science China Information Sciences,2011,54(3):551-562.

[13] Lee T L. Back-propagation neural network for the prediction of the short-term storm surge in Taichung harbor,Taiwan[J]. Engineering Applications of Artificial Intelligence,2008,21(1):63-72.

[14] Xiong Weili,Xu Baoguo. Study on optimization of SVR parameters selection based on PSO [J]. Journal of System Simulation,2006,18(9):2442-2445.

[15] Thomsen C,Pedersen T B. A survey of open source tools for business intelligence[M]//Data warehousing and knowledge discovery. Berlin:Springer,2005.

[16] 寿志勤,刘波. 数据仓库和 OLAP 技术在政府网站评估中的应用[J]. 计算机技术与发展,2011,21(10):133-136.

基于云计算的景区数据仓库应用研究

作者:

谢福伟, 梁昌勇, 马银超, XIE Fu-wei, LIANG Chang-yong, MA Yin-chao

作者单位:

合肥工业大学 管理学院, 安徽 合肥230009; 过程优化与智能决策教育部重点实验室, 安徽 合肥230009

刊名:

计算机技术与发展 

英文刊名:

Computer Technology and Development

年, 卷(期):

2014 (9)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjfz201409046.aspx