DOI: 10.3969/j.issn.1673-4785.201112018
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120608.1805.001.html

物联网与数据挖掘云服务

何清
(中国科学院计算技术研究所 智能信息处理重点实验室,北京 100190)

摘 要:物联网与云计算是目前信息技术的研究热点,探讨数据挖掘在其中扮演的角色,以及与这2项技术相结合的方式.分析了数据挖掘在物联网中的地位和作用,指出了云计算是物联网的基石,剖析了分布式数据挖掘与并行数据挖掘的异同,说明了物联网中数据挖掘服务的提供方式.

关键词:物联网;云计算;数据挖掘;云服务
中图分类号:TP391 **文献标志码**:A **文章编号**:1673-4785(2012)03-0189-06

The Internet of things and the data mining cloud service

HE Qing
(The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The Internet of things and cloud computing are a hot topic in the field of information technology. Their roles in the application of data mining, their combination, and technologies of data mining were discussed in this paper. Moreover, the status and role of data mining in the Internet of Things were analyzed; it was pointed out that cloud computing is the cornerstone of the Internet of Things. Additionally, the similarities and differences of distributed data mining and parallel data mining were analyzed. Lastly, a way to offer the data mining service in the Internet of Things was described.

Keywords: Internet of Things; cloud computing; data mining; cloud service

所谓物联网就是物物相连的互联网,也有人把它缩写成CPS(cyber physics system),"The Internet of things"是对其直观的解释.物联网实际上通过射频识别(RFID)装置、红外感应器、全球定位系统、激光扫描器等信息传感设备,按约定的协议,把任何物品与互联网相连接,进行信息交换和通信,目标是实现智能化识别、定位、跟踪、监控和管理的一种网络,人们称之为物联网[1].

1 物联网的现状

目前物联网的现状包括以下几方面.

1)国内比国外热——行业需求旺盛.我国人口众多,每一个行业所涉及的人员也比较多,因此行业需求比较旺盛.

2)政府比市场热——跨部门、跨行业.这是由我国的国情所决定的,因为物联网涉及到跨部门、跨行业问题,面对这样的难题,只有政府才有协调的能力.因此政府看到了物联网技术发展的趋势,并且正在主导物联网的发展.

3)教育比科研热——渴求技术和就业压力.这种情况一个具体的表现就是有的高校已经开始试图设立物联网专业.物联网专业所涉及的上下游技术比较多,所以可以说是一个相当综合的专业.从追求科学技术方面和就业压力方面看,对于物联网方面的知识是渴望掌握的.

4)应用比基础热——技术集成创新.因为物联网应该是以技术集成创新为基础的,所以说目前的研究更多的是如何有效地集成多种技术和进行技术

收稿日期:2011-12-28. 网络出版日期:2012-06-08.
基金项目:国家自然科学基金资助项目(60933004,60975039,61175052,61035003,61072085);国家"863"计划资助项目(2012AA011003).
通信作者:何清. E-mail: heq@ics.ict.ac.cn.

集成创新,因此应用比基础更热门.

5)硬件比软件热——可见、可检.目前可以看到传感器及传感器网络是非常热门的研究方向,并且这些相关的产品都是可见的,它的效果也是可检的,因此目前大家看到的研究状况是硬件比软件热.

6)采集比处理热——存储在先挖掘在后.物联网的技术处理必须以信息、数据为基础,首先必须先采集信息,之后才会进行处理,存储在先,挖掘在后,因此大家能看到的是采集比处理热.

2 物联网面临的挑战

物联网目前正面临着以下一系列的挑战[1].

1)物联网的商业模式有待清晰:因为物联网涉及到上下游的很多行业,在这种情况下采用什么样的商业模式,各行业如何去切分这块蛋糕,是有待解决的问题.

2)物联网的安全性、可靠性、可管理性有待加强.信息共享与保护隐私的矛盾未得到解决,这个问题在云计算方面已经得到了很大的重视.我国在云安全方面也已经投入了很多的资金和力量来解决这个问题.

3)物联网行业性太强,其公众性和公用性不足,目前的物联网还没有强大到让公众能够用起来.

4)物联网的产业链长但分散,每一环节的规模效益不够.

5)技术上重视数据收集,而忽略数据挖掘与智能处理.其原因在于目前物联网发展过程的第一阶段就是要把它部署成一个物联网,通过这个建成的物联网把数据收集上来,之后才会进行数据挖掘和智能处理.但是就总体规划而言,必须首先认识到数据挖掘和智能处理是将来物联网智能水平的一个衡量标准.

应该说发展物联网的关键是看系统的智能体现在什么地方,只有突出智能服务的特征,才能建立起一个巨大的物联网产业.

3 物联网中的计算模式

物联网的计算模式分为云计算模式和物计算模式2种,只有这2种模式有机地结合起来才能实现物联网中所需的计算、控制和决策.

1)云计算模式.

云计算作为一种基于互联网、大众参与、提供服务方式的新型计算模式,其目的是实现资源分享与整合,其中计算资源是动态、可伸缩且被虚拟化的.大量复杂的计算任务,如服务计算、变粒度计算、软计算、不确定计算、人参与的计算乃至于物参与的计算,都是云计算所面临的任务[2].云计算模式就是通过分布式的架构采集物联网中的数据,然后采用上面的云计算模式集的方法进行数据和信息处理.此模式一般用于辅助决策的数据挖掘和信息处理过程,系统的智能主要体现在数据挖掘和处理上,需要较强的集中计算能力和高带宽,但终端设备比较简单[3].

2)物计算模式.

物计算模式更多的是基于嵌入式,强调实时控制,对终端设备的性能要求较高,系统的智能的外在表现主要在终端设备上;但这种智能是嵌入的,是智能信息处理结果的利用,不能建立在复杂的终端计算基础上,对集中处理能力和系统带宽要求比较低.

之所以在物联网中采用云计算模式,原因就在于云计算事实上具备了很好的特性,是并行计算、分布式计算和网格计算的发展.而物联网中就迫切需要这种分布式的并行,目前物联网采用的云计算模式正是这种分布式并行计算模式,其主要原因是:1)低成本的分布式并行计算环境;2)云计算模式开发方便,屏蔽掉了底层;3)数据处理的规模大幅度提高;3)物联网对计算能力的需求是有差异的,云计算的扩展性好,都能满足这种差异性所带来的不同需求;4)云计算模式的容错计算能力还是比较强的,健壮性也比较强,在物联网中,由于传感器在数据采集过程的物理分布比较广泛,这种容错计算是非常必要的.

4 数据挖掘是物联网中的重要环节

4.1 物联网架构

从物联网的架构来看,基本分为4层:感知层、传输层、信息处理层和决策控制层.

1)感知层:主要是通过传感器实现对物品的识别和信息数据的采集.

2)传输层:通过现有的2G、3G以及未来4G通信网络将信息进行可靠传输.

3)信息处理层:通过后台系统进行智能信息处理,其中一个重要方面就是数据管理.

4)决策控制层:根据数据挖掘结果和预案库来反馈控制和管理物联网,而数据挖掘是决策支持和过程控制的重要技术支撑手段.

4.2 数据挖掘在物联网中的作用

互联网将信息互联互通,物联网将现实世界的

物体通过传感器和互联网连接起来,并通过云存储、云计算实现云服务.物联网具有行业应用的特征,依赖云计算对采集到的各行各业、数据格式各不相同的海量数据进行整合、管理、存储,并在整个物联网中提供数据挖掘服务,实现预测、决策,进而反向控制这些传感网络,达到控制物联网中客观事物运动和发展进程的目的.

数据挖掘是决策支持和过程控制的重要技术制成手段,它是物联网中的重要一环[4].物联网中的数据挖掘已经从传统意义上的数据统计分析、潜在模式的发现与挖掘,转向物联网中不可缺少的工具和环节.

4.3　物联网中数据挖掘的新挑战

1)分布式并行整体数据挖掘.物联网的计算设备和数据在物理上是天然分布的,因此不得不采用分布式并行数据挖掘,需要云计算模式.

2)实时高效的局部数据处理.物联网任何一个控制端均需要对瞬息万变的环境实时分析并做出反应和处理,需要物计算模式和利用数据挖掘结果.

3)数据管理与质量控制.多源、多模态、多媒体、多格式数据的存储与管理是控制数据质量和获得真实结果的重要保证,需要基于云计算的存储.

4)决策和控制.挖掘出的模式、规则、特征指标用于预测、决策和控制.

4.4　物联网中数据挖掘算法的选择

物联网特有的分布式特征,决定了物联网中的数据挖掘具有以下特征.

1)高效的数据挖掘算法:算法复杂度低、并行化程度高.

2)分布式数据挖掘算法:适合数据垂直划分的算法、重视数据挖掘多任务调度算法.

3)并行数据挖掘算法:适合数据水平划分、基于任务内并行的挖掘算法.

4)保护隐私的数据挖掘算法:数据挖掘在物联网中一定要注意保护隐私.

5　分布式与并行数据挖掘的比较

云计算相关技术的飞速发展和高速宽带网络的广泛使用,使得实际应用中分布式数据挖掘的需求不断增长.分布式数据挖掘是数据挖掘技术与分布式计算技术的有机结合,主要用于分布式环境下的数据模式发现,它是物联网中要求的数据挖掘,是在网络中挖掘出来的.通过与云计算技术相结合,可能会产生更多、更好、更新的数据挖掘方法和技术手段.

5.1　分布式数据挖掘

1)分布式数据挖掘的优点.

考虑到商业竞争和法律约束等多方面的因素,在许多情况下,为了保证数据挖掘的安全性和容错性,需要保护数据隐私,将所有数据集中在一起进行分析往往是不可行的[5].分布式数据挖掘系统能将数据合理地划分为若干个小模块,并由数据挖掘系统并行处理,最后再将各个局部的处理结果合成最终的输出模式,这样做可以充分利用分布式计算的能力和并行计算的效率,对相关的数据进行分析与综合,从而节省大量的时间和空间开销.

2)分布式数据挖掘面临的问题.

a)算法方面.实现数据预处理中各种数据挖掘算法,以及多数据挖掘任务的调度算法.

b)系统方面.能在对称多处理机(symmetrical multi-processing, SMP)、大规模并行处理机(massively parallel processor, MPP)等具体的分布式平台上实现,考虑结点间负载平衡、减少同步与通讯开销、异构数据集成等问题[5].

3)分布式数据挖掘的系统分类.

分布式数据挖掘系统,按照不同的角度可以划分为以下几类[5].

a)根据结点间数据分布情况是否同构分为同构和异构2类,同构的分布式数据挖掘系统的结点间数据的属性空间相同,异构的分布式数据挖掘系统的结点间数据具有不同的属性空间.

b)按照数据模式的生成方式,分布式数据挖掘系统分为集中式、局部式和重分布式3类.①在集中式分布式数据挖掘系统中,先把数据集中于中心点,再生成全局数据模式,该系统适合模型精度较高、但数据量较小的情况;②在局部式分布式数据挖掘系统中,先在各结点处生成局部数据模式,然后再将局部数据模式集中到中心结点生成全局数据模式,该系统适合模型精度较低,但效率较高的情形;③在重分布式数据挖掘系统中,首先将所有数据在各个结点间重新分布,然后再按照与局部式系统相同的方法生成数据模式.

5.2　并行数据挖掘与分布式数据挖掘的比较

并行数据挖掘系统与分布式数据挖掘系统都用网络连接各个数据处理结点,网络中的所有结点构成一个逻辑上的统一整体,用户可以对各个结点上的数据进行透明存取.

并行挖掘与分布式挖掘的不同点主要有如下.

1)应用目标不同.并行数据挖掘中各个处理机

结点并行完成数据挖掘任务,以提高数据挖掘系统的整体性能;分布式数据挖掘实现场地自治和数据的全局透明共享,而不要求利用网络中的所有结点来提高系统的处理性能.

2)实现方式不同.并行数据挖掘中各结点间可以采用高速网络连接,结点间的数据传输代价相对较低;分布式数据挖掘的各结点间一般采用局域网或广域网相连,网络带宽较低,点到点的通信开销较大.

3)各结点的地位不同.并行数据挖掘的各结点是非独立的,在数据处理中只能发挥协同作用,而不能有局部应用,适合于算法内并行;分布式数据挖掘系统的各结点除了能通过网络协同完成全局事务外,每个结点可以独立运行自己的数据挖掘任务,执行局部应用,具有高度的自治性,适合不同算法之间的并行.

云计算通过廉价的PC服务器,可以管理大数据量与大集群,其关键技术在于能够对云内的基础设施进行动态按需分配与管理.云计算的任务可以被分割成多个进程在多台服务器上并行计算,然后得到最终结果,其优点是对大数据量的操作性能非常好.从用户角度来看,并行计算是由单个用户完成的,分布式计算是由多个用户合作完成的,云计算是在可以没有用户参与指定计算结点的情况下,交给网络另一端的云计算平台的服务器结点自主完成计算,这样云计算就同时具备了并行与分布式的特征.

6 数据挖掘云服务方式

数据挖掘在物联网中采取了云服务的方式来提供数据挖掘的结果用于决策与控制.云计算模式是物联网的基石,能够保证分布式并行数据挖掘,实现高效、实时挖掘.云服务模式是数据挖掘的普适模式,能够保证挖掘技术的共享,降低数据挖掘应用的门槛,满足海量挖掘的需求.国内中国科学院计算技术研究所于2008年底开发完成了基于Hadoop的并行分布式数据挖掘系统PDMiner.中国移动进一步建设了256台服务器、1 000个CPU、256TB存储组成的"大云"试验平台,并在与中国科学院计算技术研究所合作开发的并行数据挖掘系统基础上,结合数据挖掘、用户行为分析等需求,在上海、江苏等地进行了应用试点,在提高效率、降低成本、节能减排等方面取得了极为显著的效果[6].在此基础上中国科学院计算技术研究所2009年开发完成了面向云计算的数据挖掘服务平台COMS,现已用于国家电网与国家信息安全领域.数据挖掘云服务平台COMS作为无锡"感知环境,智慧环保"环境监控物联网应用示范工程重要的一环,2010年7月2日通过了环保部组织的专家论证,现正在落实中.

在国际上,CHU等采用Map-Reduce并行编程模式实现了机器学习算法[7],这是在多核环境下并行算法的实现.另外,在多节点的云计算平台上的开源项目Apache Mahout 0.5于2011年5月27日发布[8].

6.1 数据挖掘云服务平台要求

数据挖掘云服务平台包括以下几个方面的要求[9].

1)基础建设:专业人士成为服务的提供者,大众和各种组织成为服务的受益方,按领域、行业进行构建.

2)虚拟化:计算资源自主分配和调度.

3)需求:大众参与应对个性化和多样化的需求.

4)可信:算法通用、可查、可调和可视.

5)安全:隐私数据由客户自己在平台终端完成加密保护.

6.2 数据挖掘云服务平台结构

数据挖掘云服务平台的结构如图1所示.可以看出,1)硬件资源管理子系统和后台并行挖掘子系统紧密结合;2)平台对用户透明,资源抽象成提供数据挖掘服务的"云";3)用户通过前台的Web交互界面定制数据挖掘任务.

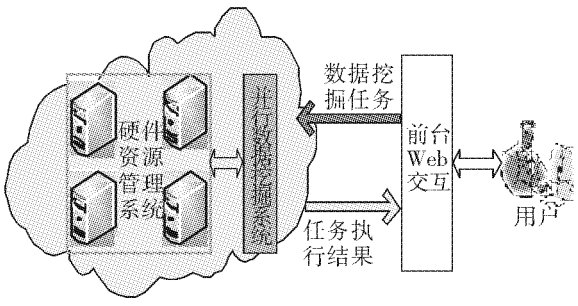

图1 数据挖掘云服务平台

Fig.1 Data mining cloud service platform

图2是数据挖掘云服务系统架构,既包括了数据挖掘预处理云服务[10],也包括了数据挖掘算法云服务,如关联规则云服务[11]、分类云服务[6,12-13]、聚类云服务[14]和异常发现云服务[15],总体上还有工作流子系统,对数据挖掘的任务进行多任务的组合,以达到数据挖掘的目标.

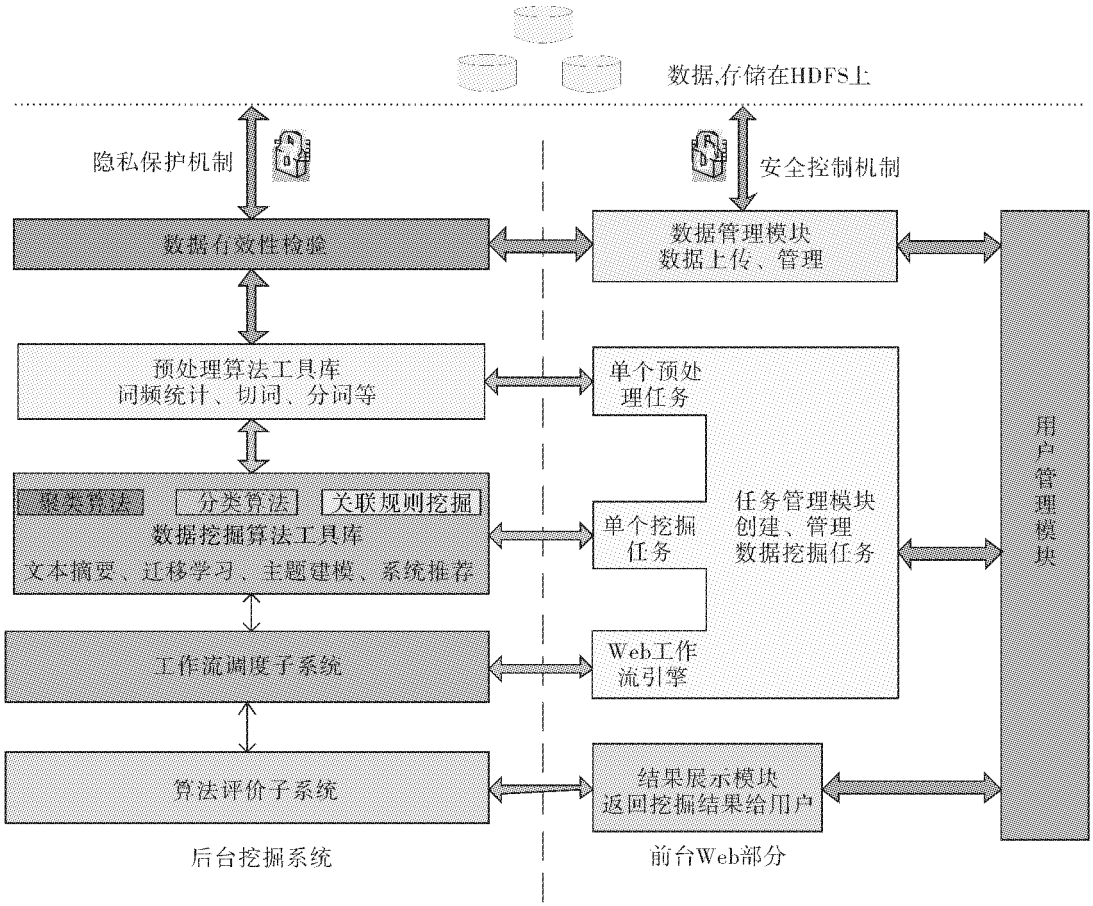

图 2　数据挖掘云服务系统架构

Fig. 2　System architecture of data mining cloud services

7　结束语

云计算是物联网的基石,数据挖掘是物联网不可缺少的重要一环. 物联网如果不加入智能信息处理和数据挖掘就不能体现智能,就只是传感器网. 而数据挖掘云服务是物联网中先进、实用、可持续、可推广的数据挖掘方式.

参考文献:

[1]邬贺铨. 中国物联网应用应该考虑中国国情[EB/OL]. (2010-06-29)[2011-11-25]. http://cloud. csdn. net/a/20100629/267886. html.

[2]李德毅. 2012 云计算技术发展报告[M]. 北京:科学出版社, 2012.

[3]马文方. 泛在计算:少谈些概念 多做些实事[N]. 中国计算机报, 2010-05-10(38).

[4]张诚,郭毅. 数据挖掘与云计算——专访中国科学院计算机研究所何清博士[J]. 数字通信, 2011, 38(3): 5-7.

[5]王媛媛. 基于概念格模型的关联规则挖掘算法研究及实现[D]. 合肥:合肥工业大学, 2005: 55-56.
WANG Yuanyuan. Research and implementation of algorithms of mining association rules based on concept lattice [D]. Hefei: Hefei University of Technology, 2005: 55-56.

[6]HE Qing, DU Changying, WANG Qun, et al. A parallel incremental extreme SVM classifier[J]. Neurocomputing, 2011, 74(16): 2532-2540.

[7]CHU C T, KIM S K, LIN Y A, et al. Map-reduce for machine learning on multicore[C]//Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2006: 281-288.

[8]The Apache Software Foundation. 27 May 2011—Apache Mahout 0. 5 released[EB/OL]. [2011-12-25]. http://mahout. apache. org/.

[9]何清. 基于云计算的海量数据挖掘[EB/OL]. (2010-05-25)[2011-11-25]. http://cloud. csdn. net/a/20100525/267105. html.

[10]HE Qing, TAN Qing, MA Xudong, et al. The high-activity parallel implementation of data preprocessing based on MapReduce[C]//The Fifth International Conference on Rough Set and Knowledge Technology (RSKT). Beijing, China, 2010: 646-654.

[11]LI Ning, ZENG Li, HE Qing, et al. Parallel implementation of apriori algorithm based on MapReduce[C]//Proceedings of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and

Parallel/Distributed Computing. Kyoto, Japan, 2012(accepted).

[12]HE Qing, ZHUANG Fuzhen, LI Jincheng, et al. Parallel implementation of classification algorithms based on MapReduce[C]//The Fifth International Conference on Rough Set and Knowledge Technology (RSKT). Beijing, China, 2010: 655-662.

[13]HE Qing, WANG Qun, DU Changying, et al. A parallel hyper-surface classifier for high dimensional data[C]// Proceedings of the 3rd International Symposium on Knowledge Acquisition and Modeling. Wuhan, China, 2010: 338-343.

[14]ZHAO Weizhong, MA Huifang, HE Qing. Parallel K-means clustering based on mapreduce[C]//The 1st International Conference on Cloud Computing. Beijing, China, 2009: 674-679.

[15]HE Qing, MA Yunlong, WANG Qun, et al. Parallel outlier detection using KD-tree based on mapreduce[C]//Proceedings of the 2011 IEEE Third International Conference on Cloud Computing Technology and Science. Athen, Greece, 2011: 75-80.

作者简介:

何清,男,1965年生,研究员,博士生导师,中国计算机学会高级会员、人工智能与模式识别专业委员会委员,中国人工智能学会副秘书长、常务理事、知识工程与分布智能专业委员会秘书长、机器学习专业委员会常务委员,中国电子学会云计算专家委员会委员.主要研究方向为机器学习、数据挖掘、文本挖掘、基于云计算的分布式并行数据挖掘等.主持和参与国家"863"和"973"计划、国家自然科学基金等科研项目多项,发表学术论文近百篇.2008年底,何清研究员带领他的中科院计算所数据挖掘团队,受中国移动研究院委托,合作开发完成了基于云计算的并行数据挖掘平台,用于TB级实际数据的挖掘,实现了高性能、低成本的数据挖掘.

2012 2nd IEEE International Conference on Cloud Computing and Intelligence Systems

2012年第2届IEEE云计算与智能系统国际会议

2012年第2届IEEE云计算与智能系统国际会议(2nd IEEE CCIS2012)将于2012年10月30日—11月1日在杭州召开.会议由IEEE北京分会和中国人工智能学会主办,大会主席由中国人工智能学会理事长李德毅院士(中国工程院院士),美国明尼苏达大学David H. C. Du教授(IEEE Fellow),CISCO Fred Baker博士(CISCO Fellow)和北京邮电大学杨放春教授(IET Fellow)共同担任.

本次会议主要围绕云计算、云存储、云安全、网络与云计算基础设施、物联网以及人工智能等方面进行广泛的研讨.会议将聚集国内外相关领域著名学者共同探讨云计算和智能系统的发展方向,并邀请世界知名学术带头人或企业领袖做特邀报告,为本领域专家、学者以及在校学生提供一个学术交流的平台,促进相关技术和产业的进一步发展.本次会议论文将被EI和ISTP全检,部分优秀文章可推荐发表在SCI检索的国际著名杂志,特此邀请您在会议网站在线投稿.

会议网址及联系方式:

Website:http://conference.bupt.edu.cn/ccis2012/

E-mail: ccis@bupt.edu.cn

Tel: 86-10-62281360

Fax: 86-10-62282983

重要日期:

提交截止日期:2012年7月15日

论文接受通知:2012年8月15日

定稿上传:2012年8月30日

会议时间:2012年10月30日—11月1日