

基于组通信技术的数据库复制技术研究与应用

王成良¹, 孙永红¹, 杨 斌¹, 王 珏²

(1. 73672 部队, 江苏 南京 210016; 2. 镇江船艇学院, 江苏 镇江 212000)

摘 要:针对分布式数据库系统中多节点数据复制存在的响应时间长、通信消耗大的问题,利用组通信技术中的消息定序特性,提出一个基于组通信技术的数据库复制协议。该协议将各节点的事务与组通信中的原子广播技术相结合,使节点上所有事务的执行达到可串行化,严格保证分布式数据库系统多节点数据的一致性。试验结果表明,相对于传统的两阶段锁 2PL 的同步复制策略,该同步模型在保证组成员节点的数据一致性前提下,减少了系统中数据库复制的通信量,避免了复制过程中的死锁,提高了同步复制的效率。

关键词:数据库复制;组通信;原子广播;一致性

中图分类号:TP311.133.1

文献标识码:A

文章编号:1673-629X(2013)05-0108-05

doi:10.3969/j.issn.1673-629X.2013.05.028

Research and Application of Database Replication Based on Group Communication Technology

WANG Cheng-liang¹, SUN Yong-hong¹, YANG Bin¹, WANG Jue²

(1. Unit 73672 of the PLA, Nanjing 210016, China;

2. The PLA Boats College, Zhenjiang 212000, China)

Abstract: In view of problems of long response time and large communication consumption existed in node data copy in distributed database system, using the news sequencing characteristics of group communication technology, put forward a database replication protocol based on group communication technology. This principle combines the transaction on each node with the technology of atomic broadcast, provides one-copy-serializability for execution of all affairs and guarantees the data consistency of the system strictly. Experimental results show that compared with traditional 2PL synchronization replication strategy, this model reduces the traffic of replication, avoids the deadlock in the process of reproduction, and improves the efficiency of eager replication.

Key words: database replication; group communication; atomic broadcast; consistency

0 引言

近年来,基于分布式数据库的应用随着计算机网络的飞速发展而日益增加,像银行管理系统、铁路售票系统等。这些应用中,通常在分布式数据库系统的设计和实现时引入冗余数据,多个数据副本在不同地点同时提供服务,减少通信代价的同时提高系统的性能和容错能力,从而提高数据的可用性和系统的可靠性^[1,2]。

如何保证这些数据副本的一致性分布式数据库系统中的一个核心问题。

1 数据库复制

1.1 基本概念

在分布式数据库系统中,数据库复制是一项解决数据副本一致性的核心技术。数据库复制是指在两个或多个不同的数据库节点之间进行数据交换,以使其中一个数据库发生的数据变化,在其他数据库节点中会相应地表现出来。数据库复制可以是单向的,也可以是双向的^[3]。单向复制又称主从复制,即将复制系统中的数据库划分为主数据库和从数据库。数据以主数据库为主,只有主数据库的变动会复制到从数据库中去。双向复制是指数据库系统中节点的关系看作互为主从,任一数据库发生变化,其他节点相应同步。

1.2 复制协议

数据复制协议是用于解决数据库系统中各节点的数据一致性问题的协议,目前的同步协议根据事务更新时间上的不同可分为同步复制和异步复制^[1]。

收稿日期:2012-08-11;修回日期:2012-11-25

基金项目:国家“973”重点基础研究发展计划项目(2012CB315901)

作者简介:王成良(1982-),男,山东烟台人,助理工程师,硕士,主要研究方向为数据库、信息安全;孙永红,硕士,高级工程师,主要研究方向为数据库技术、信息安全。

同步复制是一种实时存取和实时更新数据的同步分发技术,也称实时复制。同步复制是指在提交事务之前,更新操作必须在所有节点副本上完成,事务才会被真正提交。这种机制保证了任意时间各节点副本上数据的严格一致性,同步数据在任何时间、任何节点均保持一致。但是也带来了一系列的问题,如系统通信消耗的增加、死锁、事务响应时间延长等,同步复制一般采用2PL(两阶段提交锁)实现^[4]。

异步复制是指当事务提交之后,更新操作才被传播到其他节点副本,因此也称为存储-转发复制。节点副本之间的数据可以是暂时不同步的,但最终将保证所有节点副本上的数据保持一致。异步复制的优点是减少了系统通信消耗并且缩短了响应时间。但是事务的延迟提交导致系统节点副本数据的暂时不一致,有时会产生数据冲突。

目前,实际应用中,商业上的数据库产品通常注重系统效率,因而多采用异步复制的数据库复制协议,本质上是以牺牲数据一致性为代价而达到的弱一致性^[5]。因此,需要一种有效率的同步复制技术来满足对数据一致性和实时性要求严格的分布式数据系统。文中结合同步复制和异步复制技术的特点,基于组通信技术中的原子广播机制(Atomic Broadcast),在减小系统通信量的同时,保证了系统各节点副本间数据的一致性。

2 组通信技术

2.1 组通信概念

组通信^[6~8](Group Communication)技术是指将计算机网络中存在的若干节点作为一组,当组内的某个节点向其他节点发送信息时,每个节点都可以通过相应的通信模块接收到该节点发送的信息。

数据库复制系统中需要在分布式条件下处理很多问题:故障检测、副本协调、信息的可靠有序传播等等。这些问题在组通信中也同样存在并有很多相关研究^[6],这些需求也可以视为组通信中的问题。一些学者^[7~9]基于这些问题的解决方案实现了组通信系统,并为分布式应用程序提供了相关功能接口,具有通信可靠性和数据有序性等功能,能够很好地与数据库复制架构相结合。

2.2 组通信特性

在组通信技术的所有特性中,文中重点关注群组管理、可靠交付及消息的定序问题^[8~11]。

(1)群组管理(Membership Manage)。

组通信系统的组成员管理可以动态维护组成员信息和当前活动的组成员视图。应用程序可以通过发送加入/退出(join/leave)请求来加入/退出该组。当组

中一个成员由于故障而出现错误时,该组成员的连接信息会动态地反应在当前活动的组成员视图中。

(2)可靠交付(Reliable Delivery)。

组通信系统中的可靠交付是通过可靠组播来实现的。可靠组播将一个信息通过组播的方式发送给当前可用的组成员时,具有可靠的特点。可靠交付的特点是指如果一个消息正确发送到一个活动的组成员上,那么其他组节点上也会收到该消息。可靠组播主要有以下特性:

①正确性:一个正确的进程(没有失败的进程)组播一条消息 m ,则它终将交付 m 。

②一致性:如果一个正确的进程交付了信息 m ,则在组中所有成员的正确进程最终都会交付 m 。

(3)消息定序(Message Ordering)。

组通信系统提供以下几种不同类型的转发次序机制来解决系统中关于信息传播的次序问题^[6~11]:

①基本服务(Basic Service)。当节点接收到一条信息时,信息立即被交付到本地。由于信息到达组中每个成员节点顺序可能不同,因此基本服务不能保证接收信息的次序。

②先来先服务(FIFO Service)。又称FIFO序,如果一个进程先组播消息 m ,后组播消息 m' ,那么每个传递 m' 的正确的进程将在 m' 前交付 m 。

③因果定序服务(Causal Order Service)。若信息 m 的因果性优先于 m' ,那么在组通信系统中的所有节点成员上, m 均先于 m' 被交付。

④全序服务(Total Order Service)。又称原子组播(Atomic Multicast),指对于组通信系统中的所有成员节点,信息按照相同的次序交付到节点上。即组通信系统中任意两个节点 N_i 和 N_j 接收到信息 m 和 m' ,不管系统是先交付 m 后交付 m' ,还是系统先交付 m' 后交付 m ,在节点 N_i 和 N_j 上信息的交付次序是完全一致的。

在分布式数据库复制系统中,采用组通信系统作为通信模型,通过其原子组播的消息定序特性,可以保证系统中的组成员节点按照相同的次序接收到信息。在这种条件下,所有节点上的信息顺序一致,多副本的问题即可转换为单副本上面的问题,即一个节点上的所有事务达到可串行化,则所有节点副本上的执行顺序和这个节点相同。与其它通信模式相比,分布式数据库复制系统利用组通信的原子广播机制进行消息通信可以带来的好处如下^[7]:

(1)减小系统通信响应时间。原子广播技术在应用层面上消除了明确的回复信息,因此可以减少执行事务产生的通信等待时间。

(2)保证数据一致性。原子广播的全有或全无

(all-or-nothing)的特性保证了组内节点副本的数据一致性。

(3)消除全局复制。由于所有的站点的进度均已经在本地处理好,因此处理死锁时不会发生全局性的复制。

3 基于组通信的数据库复制协议

文中基于组通信机制的原子广播技术,提出了一种基于组通信的数据库复制协议,将组通信机制与事务机制较好地结合,可以降低数据复制中的系统通信开销,提高数据复制系统的性能。

3.1 系统模型

在分布式环境下,各个节点数据库副本以对等模式进行数据交互,相互关联(存放同一数据)的数据库节点构成一个组,数据信息在组内通过组通信系统进行传播,并通过中间件复制系统保持组内各节点数据的一致性。系统的结构如图 1 所示。

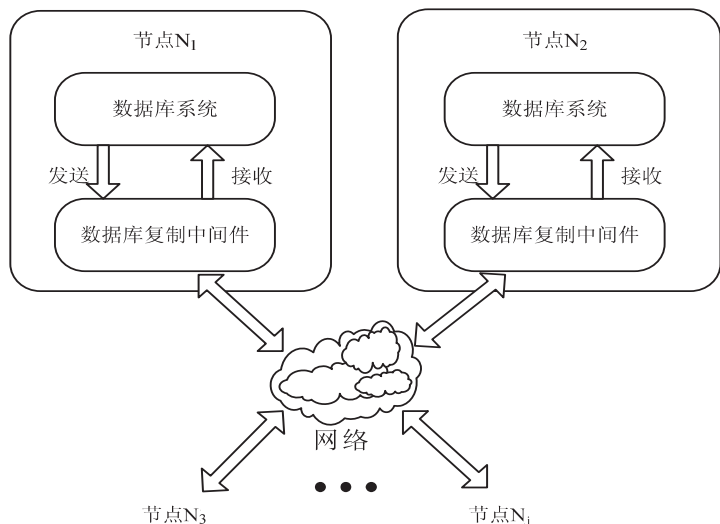


图 1 系统结构图

每个站点均由三个部分组成:

(1) 节点。

在面向分布式数据库的应用中,通过计算机网络连接的每个逻辑单位称为节点^[1]。节点用于保存、管理和使用数据资源,每个节点都有一个与之关联的身份标识和数据库。

(2) 数据库复制中间件。

分布式数据库系统是由多个节点组成的,数据库复制中间件进行事务监控,对本地数据库的活动进行实时监控,当有对数据库中的复制表发生更新操作的事务时,调用复制功能,对并发事务进行冲突检测,保证系统中各节点中的数据库的一致性。

(3) 数据库系统。

在每个节点中,用于存放节点业务的数据信息,对

于用户而言,称为本地副本(或本地数据库),其他节点中的数据库相对称为远程副本(远程数据库)。

3.2 协议描述

由系统模型可知,文中通过数据库复制中间件来实现基于组通信技术的数据库复制功能,并通过中间件完成数据库复制协议来达到对数据库和用户透明的目的。

3.2.1 协议执行流程

数据库复制中间件在用户发起事务的提交请求时获取到用户提交的事务,执行数据库复制协议,协议的执行流程主要分为三个阶段:本地截获阶段、冲突检测阶段、执行阶段。

(1) 本地截获阶段。

中间件截获节点 N_i 的用户提交的事务请求 T_i ,对 T_i 进行解析,提取出读操作集合 $\text{readSet}(T_i)$ 和写操作集合 $\text{writeSet}(T_i)$,并通过对请求类型的判断进行不同的操作。若只读事务,即 $\text{writeSet}(T_i)$ 为空,则可以直接在本地执行事务并将执行结果返回给用户;若事务请求 T_i 为更新事务,即写操作集合 $\text{writeSet}(T_i)$ 不为空,则进入冲突检测阶段。

(2) 冲突检测阶段。

令队列 $Q_i^w(i=1,2,\dots,n)$ 存放节点 N_i 上所有副本的已经通过冲突检测的写操作集合,由组通信的原子广播机制可以确保 $Q_i^w = Q_j^w(i,j=1,2,\dots,n,i \neq j)$ 。

将事务 T_i 的写操作集合 $\text{writeSet}(T_i)$ 与 Q_i^w 进行冲突检测,如果在 Q_i^w 中发现冲突,则中止该事务并将检测结果返回给用户。如果不存在冲突,则将写操作集合 $\text{writeSet}(T_i)$ 通过组通信系统全序广播到

所有副本节点中。

(3) 执行阶段。

每个节点通过组通信系统接收事务 T_i 的写操作集合 $\text{writeSet}(T_i)$,并将其加入到本地执行队列中。由组通信的全序的特性可知,所有副本接收到的写操作集合顺序一致,所以所有副本节点中的冲突检测结果一致,即若不存在冲突则所有副本节点中的判定结果都为不冲突,若存在冲突则所有副本节点的判定都为冲突,同时执行不存在冲突的事务。

3.2.2 协议执行具体步骤

令 R_i^t 表示节点 N_i 上的数据副本, $R_i^t \in R, R = \{R_1, R_2, \dots, R_n\}$ 为组内所有节点副本的集合,协议的工作流程如图 2 所示,协议具体步骤如下:

Step1: 事务获取与记录。

(1) 获取用户提交的事务请求 T_i ,在系统中开始

处理事务;

(2) SQL 语句获取。若为数据库操作语句中的读写操作(select,update,insert,delete),则提交操作,并等待结果,若操作为数据库语句中的中止(abort)操作,则中止事务 T_i 并返回;

(3) SQL 语句提取。若为提交(commit)操作,则表示该事务已经达到用户提交结尾,提取语句中的写操作集合 T_i . $WS = writeSet(T_i)$;

(4) 事务类型判断。若 T_i 为只读事务(即 $writeSet(T_i) = \emptyset$),则将事务 T_i 在 R_i^L 上提交并结束,并返回结果,否则进入事务冲突检测阶段。

Step2:事务冲突检测。
在本地节点中检测是否存在冲突,若 $\exists T_j \in Q_i^w \wedge T_i, WS \cap T_j, WS \neq \emptyset$,则说明存在冲突,中止事务并返回,否则就进入事务组播阶段。

Step3:事务组播。
(1) 将 $writeSet(T_i)$ 通过组通信系统广播到其他组成员节点;

(2) 组成员节点接收 $writeSet(T_i)$ 。
Step4:事务执行。

(1) 远程副本节点 N_j 接收到事务 T_i 的更新操作,将其加入本地队列 Q_j^w ,在副本节点 N_j 上执行事务操作;

(2) 若本地副本接收到事务 T_i 的写操作集合,则将本地队列中对应的整个事务提取并执行,返回执行结果。

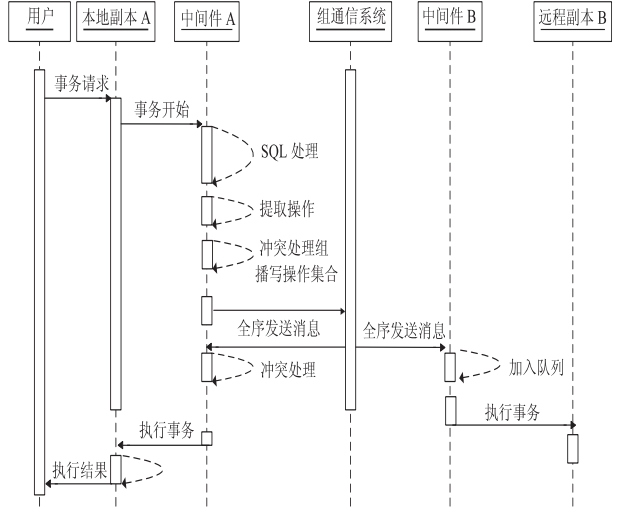


图2 协议流程图

通信支撑服务,将该支撑服务部署在每台组成员节点上。模拟环境由四台服务器作为组成员相互通信,网络结构如图3所示,服务器机配置相同:CPU采用Intel的Pentium4处理器,1G内存,2.93GHz主频,100G硬盘。数据库采用应用比较广泛的Oracle 9i,前端采用一台PC机模拟客户端向组成员的工作节点发送更新事务请求。

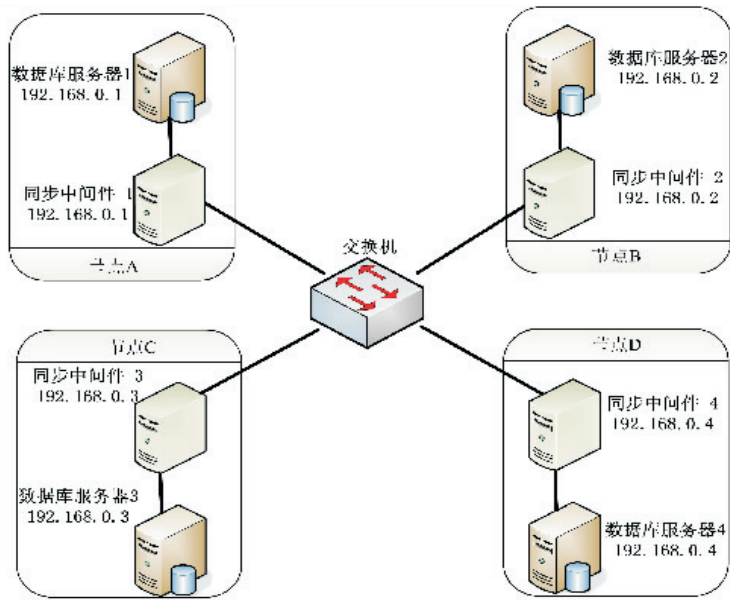


图3 系统验证环境网络结构图

在实验中,设计一张实验使用的复制表sync_test,在各个节点数据库中均存在此表,用于测试维护数据一致性的效果。表内包含主键共有5个数据字段,具体结构如表1所示。

表1 表sync_test结构

字段名	类型	大小	空值	约束条件
Student_id	number	64	N	主键
Student_name	varchar	4	N	
department	varchar	64	Y	
age	varchar	4	Y	
grade	number	4	Y	

4.2 性能测试

测试模拟了用户连续提交30、50、100、150、200、250、300个更新事务的请求的情况,观察了系统完成事务所需的时间和各节点副本达到数据一致性所需要的时间。

图4表明,由于采取了原子广播机制,事务中又除去了只读事务,传播的数据只包含了更新事务的写操作集合,减少了网络通信的消耗,所以随着事务数据量的增大,相对原同步复制2PL策略,系统通信量明显减少,完成事务所需的时间也明显减少。由此可见,在系统事务量较大的情况下,文中所使用策略的性能要明显好于原有同步复制策略。

4 实验测试

4.1 测试环境

基于组通信系统 Spread^[12]实现了一个简单的组

5 结束语

文中介绍了分布式数据库系统中数据复制方法的分类,针对目前的同步复制策略的响应时间过长,系统通信消耗大的缺点,基于组通信中的原子广播机制,提出了一种基于组通信的数据库复制协议,在保证组成员节点副本一致性的同时,减小了系统通信消耗,并在模拟测试中得到了很好的验证。下一步将进一步完善分布式数据库系统中事务并发执行的冲突解决机制,以期达到更好的数据库复制效果。

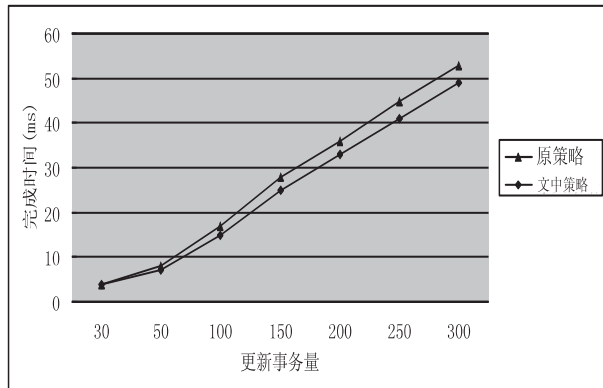


图 4 系统完成事务所需时间

参考文献:

- [1] 肖迎元. 分布式实时数据库技术[M]. 北京: 科学出版社, 2009.
- [2] Gray J, Helland P, O'Neil P, et al. The Dangers of Replication and a Solution[C]//Proceedings of SIGMOD 96. [s. l.]: [s. n.], 1996: 173-182.

n.], 1996: 173-182.

- [3] 姚路, 杨海涛, 王正华, 等. 基于 SyncML 协议的数据同步能力适应处理[J]. 计算机工程, 2009, 35(5): 68-72.
- [4] 刘腾. MySQL 复制技术的研究与改进[D]. 杭州: 浙江大学, 2011.
- [5] 王珏, 李立新, 张绍月, 等. 基于快照隔离的分布式数据库同步协议研究与实现[J]. 计算机应用研究, 2012, 29(8): 3012-3017.
- [6] Kemme B, Jiménez-Peris R, Patiño-Martínez M. Database Replication[M]. [s. l.]: Morgan and Claypool, 2010.
- [7] Amir Y, Tutu C. From total order to database replication[C]//Proceedings of International Conference on Distributed Computing Systems (ICDCS). [s. l.]: [s. n.], 2002: 494-503.
- [8] 舒后, 段成华. 高效同步复制模型的研究[J]. 计算机工程与应用, 2003, 39(8): 194-197.
- [9] Júnior A T C. Practical Database Replication[D]. Minho: Minho University, 2010.
- [10] Kemme B, Pedone F, Alonso G, et al. Using optimistic atomic broadcast in transaction processing systems[J]. IEEE Trans on Knowledge Data Engineering, 2003, 15(4): 1018-1032.
- [11] Kemme B, Alonso G. Database replication: a tale of research across communities[J]. Proc of VLDB, 2010, 3(1-2): 5-12.
- [12] Amir Y, Nita-Rotaru C, Stan-Ton J, et al. Secure spread: an integrated architecture for secure group communication[J]. IEEE Trans on Dependable and Secure Computing, 2005, 2(3): 248-261.

(上接第 107 页)

制. RBAC 是一种广泛使用的访问控制模型,但在有些环境中很难应用^[14]。目前基于模糊的 RBAC 受到关注,通过模糊关系,使授权的相关信息是模糊的,扩大了适用环境,是未来的发展方向。

参考文献:

- [1] 张敏,徐震,冯登国. 数据库安全[M]. 北京: 科学出版社, 2005.
- [2] 陈红梅,葛德江. SQL Server 中基于角色的访问控制应用[J]. 电脑知识与技术, 2008, 3(25): 1375-1377.
- [3] 王晓超,赵卫东,左青香. 基于元数据和角色控制的用户权限管理[J]. 计算机技术与发展, 2012, 22(3): 233-236.
- [4] 王海亮,林立新,焦大光,等. Oracle10g 快速入门[M]. 北京: 中国水利水电出版社, 2007.
- [5] 李岚. 基于角色的数据库安全访问控制的应用[J]. 通信技术, 2008, 41(10): 70-72.
- [6] Kim S, Kim Dae-Kyoo, Kim S. A feature-based approach for modeling role-based access control systems[J]. Journal of Systems and Software, 2011, 84(12): 2035-2052.
- [7] Richard D, Edward J, Timothy R. Adding Attributes to Role-

based Access Control[J]. IEEE Computer, 2010, 43(6): 79-81.

- [8] ScienceDirect. Practical Oracle Security[EB/OL]. 2012-07-11. <http://www.sciencedirect.com/science/>.
- [9] 冯凤娟. Oracle 数据库体系结构和管理[M]. 北京: 清华大学出版社, 2003.
- [10] Dewson R. SQL Server 2005 基础教程[M]. 北京: 人民邮电出版社, 2006.
- [11] England K, Powell G. Microsoft SQL Server 2005 Performance Optimization and Tuning Handbook[M]. [s. l.]: Digital Press, 2007.
- [12] 源码天空. SQL 用户自定义角色的创建[EB/OL]. 2012-07-11. <http://www.codesky.net/article/201010/144597.html>.
- [13] 田学志,邵保华. 浅析 Oracle 中的角色与权限[J]. 黑龙江生态工程职业学院学报, 2008, 21(4): 74-75.
- [14] Martínez-García C, Navarro-Arribas G, Borrell J. Fuzzy Role-based Access Control[J]. Information Processing Letters, 2011, 111(10): 483-487.

基于组通信技术的数据库复制技术研究与应用

作者: [王成良](#), [孙永红](#), [杨斌](#), [王珏](#)
作者单位: [王成良, 孙永红, 杨斌\(173672部队, 江苏 南京210016\)](#), [王珏\(镇江船艇学院, 江苏 镇江 212000\)](#)
刊名: [计算机技术与发展](#)
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2013(5)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201305030.aspx