维普资讯 http://www.cqvip.com

一种 XML 概念模型—XUML*)

刘洪星[1,2] 卢炎生[1] 陈 明[2]
(华中科技大学计算机科学与技术学院 武汉 430074)[1]
(武汉理工大学计算机科学与技术学院 武汉 430063)[2]

摘 要 由于 XML 已成为 Web 上表示结构化和半结构化数据的标准,设计 XML 模式的方法变得更加重要。为了设计或集成 XML 模式,常常需要基于合适的概念模型。本文分析了对 XML 概念模型的需求;提出了一种新的 XML 概念模型:XUML,并说明了 XUML 的主要特征和优点:能更明确地表示"包含"语义,支持"业务组件"概念,能在多级上下文中说明数据依赖,基于 UML2 标准;最后结合实例介绍了一种实现 XUML 的方法。

关键词 XML,概念模型,XML 模式,UML2

An XML Conceptual Model: XUML

LIU Hong-Xing[1,2] LU Yan-Sheng[1] CHEN Ming[2]
(School of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 430074)[1]
(School of Computer Science & Technology, Wuhan University of Technology, Wuhan 430063)[2]

Abstract As XML has become the standard for representing structured and semi-structured data on the Web, the methods for designing XML schemas is becoming more and more important. In order to design or integrate XML schemas, it is necessary to first design the conceptual structures with a proper conceptual model. The requirement to XML conceptual model is analyzed. Then a new XML conceptual model, XUML, has been introduced, which has following characteristics and advantages: expressing the containment semantics more explicitly, supporting the concept of Business Components, specifying the data dependencies in multiple contexts, and based on the UML2 standard. Lastly an approach to the implementation of XUML is explained by using an example.

Keywords XML,Conceptual model,XML schema,UML2

1 引言

XML 已成为 Internet 上各种应用系统和数据库之间数据表示和交换的标准。为了描述 XML 文档的语法和结构,可以采用 DTD 或 W3C XML Schema 语言(WXS)来定义文档的模式。然而,这样的模式表示的是文档的逻辑结构,而不是概念结构,故难以表示语义。为了设计 XML 文档或 XML 数据库,有必要采用合适的概念模型,来描述现实世界中的数据语义以及它们之间的复杂联系。目前已提出了一些 XML 概念模型,比如 Semantic Network[1], AOM[2], ORA-SS[3], X-Entity[4]和 C-XML[5],这些模型通过扩展语义网或 ER 模型来支持 XML。文[6]定义了一些专门的 Profile,通过扩展 UML1.x 来支持 WXS。相比逻辑级 XML 模式,这些概念模型能捕获更丰富的语义和约束;虽然一定程度地改进了 XML 模式的设计,但仍然不足以支持 XML 概念建模。

本文提出了一种新的 XML 概念模型——XUML,它继承了上述模型的优点,而在几个重要方面有了增强。XUML 主要用于设计 XML 文档或 XML 数据库,也可支持 XML 信息集成;这些在大型电子商务环境和内容管理系统中是非常重要的。

本文第 2 节分析了对 UML 概念模型的需求;第 3 节描述了 UXML 模型及其主要特点;第 4 节介绍了 XUML 的实现,并结合实例说明了其应用方法,最后是结束语。

2 XML 概念模型的需求分析

概念模型用于描述业务领域中的对象及对象之间的关系,它们独立于实现平台,便于交流。人们对结构化数据(如:关系数据)的概念建模已有了多年研究,形成了成熟的概念模型(如:ER 模型)和相应的概念建模方法。但 XML 是自描述、半结构化和可扩展的标记语言。作为一种标记语言,它将数据和对数据的描述(元数据)结合在一起,因而具有比关系模型更灵活的描述能力——不仅能表示结构化数据,还能表示半结构化数据。XML 不仅能作为信息的持久存储格式,更能作为信息流的格式,其应用领域十分广泛。这些都对 XML 模式的概念设计提出了新的要求。

为了适应 XML 自身的特点和其应用特点,一个好的 XML 概念模型应该具有如下性质:

1)支持"包含"语义,应将包含联系提升到一级建模原语的地位。XML 具有天然的层次结构,XML 表达对象之间联系的最主要形式就是元素之间的层次嵌套。这种嵌套自然地表达了包含(Containment)语义。然而,已有的 XML 概念模型并不能准确地表示这种包含语义。在图 1 所示的 UML 类图模型中,共有 3 个 2 元组合联系;虽然都关联到相同的"整体"(Order),但 3 个联系彼此之间却是独立的。第 3 节将进一步说明,普通的类图不能很好地表达 XML 中所需的"包含"语义。

*)湖北省自然科学基金资助,项目编号:2004ABA040。刘洪星 博士生,副教授,主要研究方向:XML 与数据库设计,信息集成。卢炎生 教授,博士生导师,主要研究方向:新型数据库技术,软件测试。

维普资讯 http://www.cqvip.com

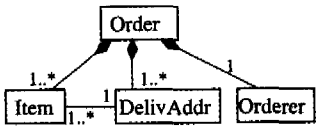

图 1　三个 2 元聚集联系

2)明确支持业务组件的概念。1 个 XML 文档可以记录 1 个简单的原子对象,但更一般的情况下,它表示的是 1 个由若干对象组成的复杂业务组件,所以 XML 概念模型应该支持业务组件的概念。

3)支持对"顺序"的建模。XML 文档中各元素之间的顺序常常需要明确指定。

4)能明确表达上下文中的数据依赖。文档中的元素可以组成多级的层次结构,而元素之间的一些约束条件可能只在局部范围内有效,这就需要概念模型能表示多级上下文中的数据依赖。

5)同时支持 XML 文档和 XML 数据库的建模。一个 XML 文档通常是信息自包含的,用来记录一份业务文档或一个业务对象。XML 数据库(或 XML 文档数据库)是 XML 文档的集合,它们需要持久保存,以便查询。设计 XML 数据库时,除了考虑同一文档内的联系,还要考虑跨文档的联系。

6)基于标准。只有基于标准,才可能有更广泛的应用。

3　XUML 特征

XUML 是一种 XML 概念模型,它能满足前面说明的主要需求。下面说明其最重要的几个特征。

3.1　"一般聚集"联系

UML1. x 和 2. 0 要求聚集联系必须是 2 元的[7]。但在 XML 中,一个联系常常有 2 个以上的参与者,比如,1 个元素包含多个不同类型的子元素。定义这种联系的不变式(Invariant)涉及到联系的所有参与者,而不仅仅是其中的 2 个,故几个 2 元联系并不等于一个非 2 元联系。图 1 中,共有 3 个 2 元组合联系,它们虽然具有相同的"整体"(Order),但每个联系却是相互独立的,即,模型并没有表示 3 个部件的实例同时关联同一个 Order 实例。为了表示"整体"与"部件的集合"相关,我们参考文[8],在 XUML 中定义了"一般聚集",并将其作为基本的联系类型。

定义 1　对于一个一般聚集(generic Aggregation),其中一个参与者("整体")的特性,部分地,由其它参与者("部分")的特性来确定。一个聚集(aggregate)类型对应一个或多个部分类型,而一个聚集实例对应每个部分类型的 0 个或多个实例。

一般聚集比 UML 中普通的 2 元聚集更一般化,所以称之为"一般的"。它是非对称的,分成了整体和部分二种成分,所以也不同于基本 UML 中的(对称的)n 元关联。图 2 中有一个"非 2 元的一般聚集"联系,它表示 1 个 Order 由至少 1 个 Item,1 个 DelivAddr 和 1 个 Orderer 所"共同组成";而 Item 和 DelivAddr 之间的关联是一个 Order"内部的"联系,称为内部关联(Inner association)。一般聚集的重要意义,在于它更准确地表示了 XML 中的"包含"语义,描述了 XML 中最重要的联系,即,元素和子元素之间的层次联系。

在 XUML 中,"2 元一般聚集"的图形表示还是采用传统 UML 中的菱形符,而"非 2 元一般聚集"则采用了新的图形表示,如图 2 所示。注意,图 2 中的树型表示不同于基本 UML 中的树型表示,后者只是多个 2 元聚集的一种简化表示形式。

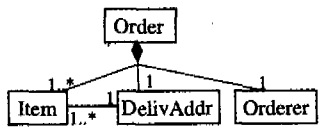

图 2　一个非 2 元一般聚集

一般聚集可分为共享的(非层次化的),或组合的(层次化的)。组合型一般聚集(图形表示为实心菱形)是缺省情形。图 2 中的一般聚集只有 2 层,但一般聚集可以嵌套并形成一个多级的聚集层次。一般聚集是传递的。

3.2　业务组件

组件是系统中的一个模块化部件,它现在在 UML2 中是一个逻辑概念[7]。

定义 2　业务组件(Business Component, BC)代表一个能独立存在的单元,它由一个对象或一组相关的对象组成。有 2 种业务组件:业务对象(business object, BO)和业务文档(business document, BD)。BO 在业务过程中扮演一个角色;而 BD 是在业务过程中,不同 BO 之间交换的信息[2]。

传统的概念建模方法,例如 ER 方法,倾向于将现实世界抽象成一系列基本的、最小冗余的实体(规范化实体),以及这些实体之间的联系。这样建立的概念模式很容易转换成规范化的关系数据库模式。但是,在这样的概念模型中,关于 BC 的概念常常是模糊的。在 XML 中,一个实例文档通常记录并描述一个 BC,而对应的文档模式则描述了这种 BC 的结构。所以在 XML 应用中,BC 概念比在常规数据库中更为重要。XUML 引入了 BC 概念,并将其作为"一级"建模概念。

在 XUML 中,用 1 个一般聚集或 1 个多级的一般聚集层次,来表示 1 个 BC。一般聚集层次的根类,又称标识类(Identifying class),是对文档根元素的概念建模;一般聚集联系,是对元素间"包含"联系的概念建模;而聚集层次中的内部关联,则是对文档内 Key/Keyref 或 ID/IDRESF(S)的概念建模。

3.3　结构化类

为了更精确地对 BC 的内部结构进行建模,我们采用 UML2 新引入的结构化类(structured class)[7]。结构化类是具有内部结构的类元。它是一些部件(part)的集合,这些部件之间可通过连接器(connector)而连接起来。图 3 是一个结构化类的例子,它和图 2 表达了类似的结构语义。

部件在其容器内可以有类型和多重性。结构化对象是结构化类的实例,它"包含"了其中的所有对象,这些对象在同一个上下文中隐式相关。连接器是同一结构化类中 2 个部件之间的联系,即,它是一种在特定上下文中的联系。2 个部件之间的联系不同于 2 个普通类之间的联系,即使这 2 个类分别通过组合关联到同一个类;例如:图 3 中 Item 和 DelivAddr 之间实际上是一个连接器,而图 1 中 Item 和 DelivAddr 之间是一个普通的关联;理解这一区别十分重要。

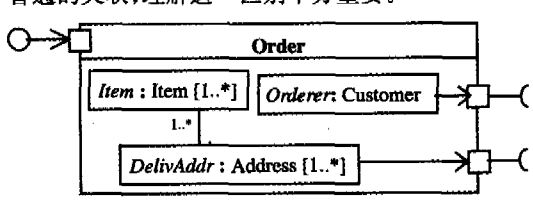

图 3　一个结构化类,其部件和端口

维普资讯 http://www.cqvip.com

结构化类可以被整体封装，使内部部件与外界的交互只能通过端口(port)来进行。端口是一个交互点，它具有一个定义好的接口(interface)。当在一种 BC 上定义查询或操纵运算时，或定义 BC 之间的参引(reference)联系时，端口是一个很好的机制。图 3 中有一个 provided 接口和 2 个 required 接口。

结构化类和一般聚集有如下区别：1)结构化类中的部件不是类，而一般聚集中的部件是类！2)1 个结构化类只能表示它的儿子部件之间的连接，不能在同一图中表示儿子和孙子部件之间的连接，但几个结构化类可以结合起来表示嵌套层次；而在 1 个一般聚集层次图中，可以同时表示跨层的部件类之间的内部关联。在 XUML 模型中，结构化类主要用来描述 1 个 BC 的实现，其接口可用来描述 BC 之间的联系。

在一个典型的 XML 数据库中，可能会有多种不同类型的文档。我们用一般聚集层次和结构化类来建模文档的内部结构；用接口和接口间的通信来模拟文档之间的 XLink 参引。

3.4 上下文中的数据依赖

数据依赖(特别是函数依赖)是最重要的语义。在传统的数据库设计中，数据依赖是在“扁平的”实体或类这样的上下文中进行分析。由于 XML 可能是多层的层次结构，因此一个 XML 业务组件中的数据依赖要更复杂，它们常常需要在多层上下文环境中说明。

结构化类是一个容器、一个命名空间，还可以是嵌套的，所以它也是一种说明数据依赖的、合适的上下文环境。

例 1 在图 1 所示的概念模式中，业务概念 Order 被分成了 4 个独立而又相互关联的概念，这 4 个概念处于同一层次。在这个例子中，业务规则“Order 中的一个 Item(货物)被发送到一个指定地址”，可表示成如下函数依赖：

Item (orderID1, itemID) → DelivAddr (orderID2, deliverAddr)，其中 orderID1=orderID2.

而在图 3 中，由于 Order 已被表示成一个结构化类，它已建立了一个特定的上下文，因此上述函数依赖可以简洁地表示为：

Order: (Item → DelivAddr)

4 XUML 的实现与应用

UML 是一种可扩展的建模语言。UML 元模型是通过 OMG 的元对象设施(MOF)定义的，所以可以通过 MOF 来扩展它。使用 MOF 的全部能力的扩展被称为重型扩展(heavyweight extension)。UML 内置的扩展机制是 profiling，即通过在 profile 中定义 stereotype 和 tagged value，这种扩展被称为轻型扩展。完整的 XUML 实现需要采用重型扩展。目前，我们先进行轻型扩展。Sybase PowerDesigner v12(简称 PD)是一种功能很强的 UML 建模工具，它有良好的扩展机制，并对 UML2 有初步的支持。

我们在 PD 基础上初步实现了 XUML，实现的要点是：

1)针对 WXS 语言，对 PD 提供的 profile 进行适当的调整和扩充，以适应 XUML 的特点。

2)在不和原有定义相冲突的前提下，对 PD 中的内嵌类(inner class)和包(package)进行了一种应用方式上的“具体化(reify)”调整，要点是：a)将类与其内嵌类之间的联系(称为 inner link)看成是“包含”联系；即用内嵌类来模拟部分类，并将 1 个类和其所有内嵌类之间的联系看成是 XUML 中的 1 个“一般聚集”联系。b)关于 1 个 BC 和其部件类的全部定义都放置在 1 个包中，将 BC 中的标识类放在包的顶层；类的限定名(qualifiedname)，是根据该类在一般聚集层次中的地位来确定。

下面通过一个例子，来进一步说明 XUML 的一些实现特征。该例子也部分展示了基于 XUML 的 XML 概念建模方法。

例 2 设计一个学院文档模式。要求文档记录一个学院内的所有项目和教师(这是重点！)，并记录教师的论文、研究生，以及上述对象之间(在学院范围内)的联系。假设已有学校的企业级概念模式，用普通的 UML 类图表示，图 4 是该模式的部分。根据设计需求，我们对图 4 进行概念级的层次化分析和重构，得到了学院的 XUML 图(图 5)，这就是学院文档的概念模式(结构)。

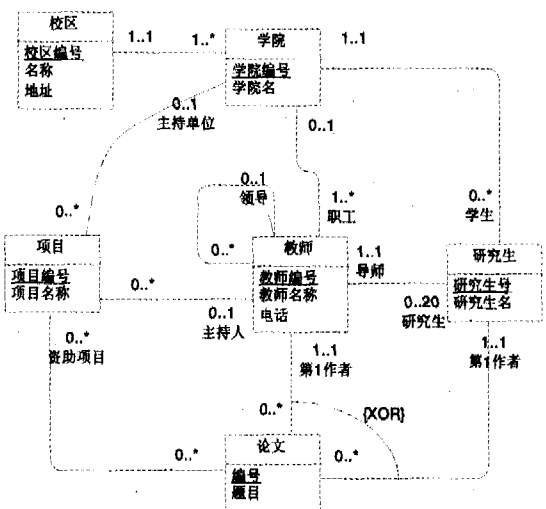

图 4 学校的企业级概念模式(普通的 UML 类图)

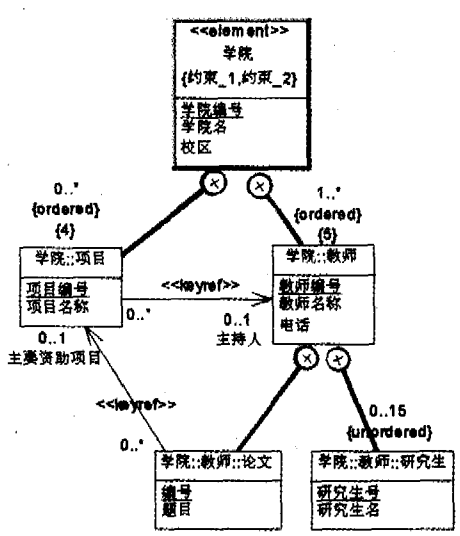

图 5 学院文档的概念模式(XUML 类图)

图 5 中，学院下的 2 个 inner link 被认为是 1 个整体，它们一起表示了 1 个“一般聚集”。图中有 2 个一般聚集，一起构成了 1 个“一般聚集层次”。该一般聚集层次刻画了学院内部的主要概念结构。注意研究生类的限定名：学院::教师::研究生，这也表达了研究生在该概念结构中所处的层次。XUML 中的一般聚集联系是传递的，这表示研究生和其导师属同一学院——该约束在图 4 中不能明确表示。图 5 中有 2

维普资讯 http://www.cqvip.com

个"内部关联":项目—〉教师,论文—〉项目;这分别是对文档内同层元素之间,和不同层元素之间参引联系的概念建模。进一步地,为了更准确地从 XUML 生成 XML schemas,可以在建立的 XUML 模型中指定各种 stereotypes(如《element》,《keyref》),以及 tagged values(如{ordered},{约束_1})。项目类上方的{4},指示在生成学院元素时,将项目元素放在 sequence 组的第 4 位。

从这个例子可以看出,所谓 XML 概念模式,实际上是企业概念模式的一种层次化视图,即从特定角度,对企业信息结构的一种看法。XML 概念模式通常是企业概念模式的一个局部视图。

当设计了 XUML 概念模型之后,逻辑设计就相对比较简单了。我们设计了一个 XUML2XSD 算法,该算法根据一组映射规则,自动将 XUML 模型转换成 XML 模式说明(XSD)。有了 XUML 图和 XUML2XSD 算法,用户可以不必精通 WXS 语言,就能设计高质量的 XML 模式。

结束语 本文提出了一种基于 UML2 的 XML 概念模型——XUML。通过隐藏关于实现的细节,强调语义相关的概念,XUML 能更清晰地描述 XML 的概念模式。它最重要的特征是能支持"一般的,非对称的聚集联系",这一特征对 XML 概念建模十分重要;就我们目前所知,其它 XML 概念模型还没有说明这样的特征。XUML 能明确地表示包含语义和业务组件概念,能在多级上下文中说明数据依赖。XUML 的元模型是对 UML2 元模型的扩展,所以能充分利用 UML2 的强大功能,例如:结构化类,部件,连接器和端口等。我们的前期研究表明,XUML 是一种对 XML 进行概念建模的合适模型。我们将在 2 个方面进一步扩展该研究。第一,在 Eclipse 环境中实现和集成一个支持 XUML 的 CASE 工具;第二,进一步开发并评估基于 XUML 的 XML 设计方法学。

参 考 文 献

1 Feng Ling, Chang E, Dillon T. A Semantic Network-Based Design Methodology for XML Documents. ACM Transactions on Information Systems, October 2002, 20(4): 390～421
2 Daum B. Asset Oriented Modeling (AOM). http://www.aomodeling.org. 2005
3 Dobbie G, Wu X, Ling T W, Lee M L. ORA-SS: An Object-Relationship-Attribute Model for Semistructured Data: [Technical Report TR21/00]. National University of Singapore, 2000
4 Lóscio B F, Salgado A C, Galvão L R. Conceptual Modeling of XML Schemas [C]. Proceedings of WIDM03, ACM Press, 2003. 102～105
5 Embley D W, Liddle S W, Reema Al-Kamha. Enterprise Modeling with Conceptual XML. In: Proceedings of ER 2004, LNCS 3288, Springer-Verlag, 2004. 150～165
6 Carlson D. Modeling XML Applications with UML. Addison-Wesley Professional, 2001
7 Rumbaugh J, Jacobson I, Booch G. The Unified Modeling Language Reference Manual (2nd Edition). Addison-Wesley, Reading, MA, 2004
8 OMG. UML Profile for Relationships Specification. http://www.omg.org/cgi-bin/doc? formal/2004-02-07

(上接第 73 页)

鉴于此,本文在对 Ad Hoc 网络的诸多动态特性分析的基础上,通过对传统网络测量技术和 NT 技术的比较,提出了基于 NT 技术的 Ad Hoc 网络性能测量架构,并指出其中所要解决的几个关键问题。由于 NT 技术和 Ad Hoc 网络本身正处于日益发展之中,因而基于 NT 技术的 Ad Hoc 网络测量需要进一步应用到具体网络环境中进行研究。

参 考 文 献

1 Karapantelakis T, Iacovidis G. Experimenting with Real Time Applications in an IEEE 802.11b Ad Hoc Network. In: The IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05), 2005, 1.554～559
2 Habib A, Fahmy S, Bhargava B. Monitoring and controlling QoS network domains. International Journal of Network Management, 2005, 15: 11～29
3 Awerbuch B, Holmer D, Rubens H. The Pulse Protocol: Mobile Ad hoc Network Performance Evaluation. In: Second Annual Conference on Wireless On-demand Network Systems and Services (WONS'05), 2005. 206～215
4 赵金晶,朱培栋. Ad Hoc 网络移动模型及其应用. 计算机工程与科学,2005,27(5):15～17
5 Nilsson A. Performance Analysis of Traffic Load and Node Density in Ad hoc Networks. In: Proc. European Wireless, 2004
6 Li N, Hou J C. Topology control in heterogeneous wireless networks: problems and solutions. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, 2004, 1(1): 233～243
7 Li N, et al. Design and analysis of a MST-based distributed topology control algorithm for wireless ad-hoc networks. IEEE Trans. on Wireless Communications, 2005, 4(1): 1195～1206
8 Rodoplu V, Meng T H. Growth of Ad Hoc Networks. Global Telecommunications Conference, 2003, 5 (1): 2819～2823
9 Agarwal A, Kumar P R. Capacity Bounds for Ad Hoc and Hybrid Wireless Networks. ACM SIGCOMM Computer Communication Review, 2004, 34(3): 71～81
10 D S. J. De Couto Daniel Aguayo Benjamin A. Chambers Robert Morris. Effects of Loss Rate on Ad Hoc Wireless Routing: [Technical Reports]. MIT-LCS-TR-836. MIT Laboratory for Computer Science, March 2002
11 吕淑琴,黄浩. 基于 SNMP 协议的网络性能分析. 电脑与信息技术,2005,1(1):60～62
12 Castro R, Coates M, Liang G, et al. Network Tomography: Recent Developments. Statistical Science, 2004, 19(3): 499～517
13 Duffield N G, Presti F L. Network Tomography From Measured End-to-End Delay Covariance. IEEE/ACM Transactions on Networking, 2004, 12(6): 978～992
14 Samar P, Wicker S B. On the Behavior of Communication Links of a Node in a Multi-Hop Mobile Environment. In: Proc. MobiHoc, 2004
15 Kang Seong-ryong, Liu Xiliang, Dai Min, et al. Packet-Pair Bandwidth Estimation: Stochastic Analysis of a Single Congested Node. In: 12th IEEE International Conference on Network Protocols (ICNP'04), 2004. 316～325
16 Lawrence E, Michailidis G, Nair V N. Maximum Likelihood Estimation of Internal Network Link Delay Distributions Using Multicast Measurements. In: Proceedings of the 37th Conference on Information Sciences and Systems (refereed). 2003
17 Tsang Y, Coates M, Nowak R. Passive Unicast Network Tomography using EM Algorithms. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, Utah, 2001, 3: 1469～1472
18 Liang G, Yu B. Maximum Pseudo Likelihood Estimation in Network Tomography. IEEE Trans Signal Processing, 2003, 51(8): 2043～2053
19 Padmanabhan V N, Qiu Lili, Wang H J. Passive network tomography using Bayesian inference. In: Proc. ACM SIGCOMM Workshop on Internet Measurement, New York: ACM Press, 2002. 93～94
20 Bettstetter C, Hartmann C. Connectivity of Wireless Multihop Networks in a Shadow Fading Environment. ACM/Springer Wireless Networks, 2005, 11(5): 571～579
21 Simon G, Stéger J, Hága P, et al. Measuring the Dynamical State of the Internet: Large Scale Network Tomography via the ETOMIC Infrastructure. The European Conference on Complex Systems / ECCS, 2005
22 Kwak B-J, et al. On the Scalability of Ad Hoc Networks: a traffic analysis at the center of a network. IEEE Communications Letters, 2004, 8: 503～505
23 Aguayo D, Bicket J, Biswas S, et al. Link-level Measurements from an 802.11b Mesh Network. ACM SIGCOMM Computer Communication Review, 2004, 34(4): 121～132
24 Zhao J, Govindan R, Estrin D. Sensor Network Tomography: monitoring wireless sensor networks. ACM SIGCOMM Computer Communication Review, 2002, 32(1): 64～64