XML 相关标准综述

A Survey of Specifications Correlative with XML

杨建武　陈晓鸥

（北京大学计算机研究所　文字信息处理技术国家重点实验室　北京100871）

Abstract　With the rapid development of Internet, XML has been accepted widely and has become a pop topic of study and application. XML specification is sample, but its theory is profound, its system is complicated and lots of specifications are correlative with XML. An overview of XML's development and characteristics are given in the paper. In allusion to complexity of the system, authors summarize the XML system architecture, class the XML foundation specifications and discuss their relation. Some questions are list in the end of paper.

Keywords　XML, SGML, Structured, Markup language

1　引言

随着 Internet 的迅速发展，XML（eXtensible Markup Language，可扩展置标语言）以其自身的独特优点，经过短短几年的发展，已被广泛接受并成为研究与应用的热点。

XML 1.0[1]是国际组织 W3C 于1998年2月正式推出的。它是为适应 Web 的应用，对国际标准 ISO8879 SGML（标准通用置标语言）进行简化形成的通用置标语言。通用置标语言是指一套通用置标的表示方法[2,3]。所谓置标是指插入文件数据中的正文，用于表述关于该文件的信息。

通用置标是相对过程性置标而言的，它是以两个前提为基础的：a）置标描述的是文件的结构及其它属性；b）置标应该是严格的。通用置标是将文件归约成一种已知语法的正规表达式。这使得可以使用计算机语法学与编译程序设计中已建立起来的技术进行文档的处理。

2　通用置标语言的发展

2.1　标准通用置标语言——SGML

在计算机发展的早期，电子文档中包含"特殊编码"，以使文档按某种特定的方式格式化。60年代末期出现的"描述编码"则使用描述性的标签来描述文档的格式。

1969年，IBM 的技术人员 Charles Goldfarb 等发明了通用置标语言 GML（Generalized Markup Language）用来解决 IBM 公司内部大量文档和刊物的交换和存储。它在各文档之间共享一些相似的属性，允许文本编辑、格式化和信息检索等子系统共享文件。使用 GML 置标的同一源文件，通过不同的批处理程序格式化为不同格式的输出形式，从而避免了不同文档格式间的转换。

1986年，标准通用置标语言 SGML 成了国际标准 ISO8879：信息处理——文本和办公系统——标准通用置标语言（Information processing---Text and office systems --- Standard Generalized Markup Language (SGML)）。

SGML 是一个十分规范、结构化、可扩展的，足以用它建立大型的跨平台的信息资料库，但也正因为它的太完备、可定义性太强使得实现这样的系统很困难。SGML 最初的最大应用是用在美国国防部的 CALS 系统中，作为美国国防部内部技术资料的交换标准，目前 SGML 主要作为数字式图书馆、大型资料库的数据基础。

SGML 文件的组成主要有三部分，即 SGML 声明、DTD 和文件实例。SGML 声明是说明 DTD 和文件实例所使用的语法，其中包括文件和语法元素使用的字符集、定界符和命名规则、名字字符的使用、保留名的替换、可选的特征等，它体现的是"抽象语法"的思想，对于不同的系统环境、民族习惯及键盘的不同 SGML 的语法是可以改变的。DTD（Document Type Definition）称文档类型定义，与 XML 中的 DTD 一样，主要作用是定义一类文件的结构。文件实例是文件实际要表达的信息，是由文件数据内容和描述结构的置标组成。

2.2　超文本置标语言——HTML

超文本置标语言 HTML（HyperText Markup Language）是一种用于建立超文本/超媒体文档的置标语言，是 SGML 的一种应用。

80年代末，WWW 的发明人 Tim Berners-Lee 借用了 SGML 的一个例子 DTD 中置标的表示方法，发明了一套标记，并加入了超链接，形成 HTML。随着 Mosaic 的出现，HTML 在 Internet 上迅速蔓延开来，使 WWW 成为最受欢迎的 Internet 服务，至今 HTML 也经历了2.0、3.0、4.0，并向 XHtml 发展。

HTML 在发展中，结合了许多其他的 Web 技术，如：CGI、脚本语言（Script）、部件技术、Java 应用小件（Java applet），使 HTML 具有了交互性，发展出基于 Web 的应用，浏览器不再是单纯的信息接收者。HTML3.2加入了样式 CSS（Cascading Style Sheet），使 HTML 的版式越来越灵活，特别是 CSS2中有了对象的绝对定位、相对定位，可以解决以前许多对 HTML 版式控制的难题。W3C 的 DOM 标准是对动态 HTML 的规范，使得浏览器中的 HTML 文档可动态操纵。

但 HTML 的最大缺点是，它只是 SGML 的一个应用，是由有限的标记组成的置标语言，无法描述应用特定的结构，使 Web 上交换的数据只能用来显示，不能进行更复杂些的处理；仅对于显示而言，HTML 也不能涵盖所有的描述，如数学公式和化学公式。

2.3　可扩展置标语言——XML

1996年，随着 Internet 的发展，HTML 的缺点逐渐暴露

杨建武　博士研究生，主要研究方向为 SGML/XML 与数据挖掘。陈晓鸥　副教授，主要研究方向为彩色图像处理、XML 数据交换与表现。

出来。人们开始致力于描述一个置标语言，它既具有 SGML 的强大功能，同时又具有 HTML 的简单性。1998年2月，W3C 正式推出了 XML 的1.0版本。

XML 是 SGML 的一个简化而严格的子集，是特别为 Web 应用而设计的。它去掉了 SGML 中很少使用而且处理起来很麻烦的特征，但继承了 SGML 具有的可扩展性、结构性及可校验性，这使 SGML 的优秀品质能方便而直接地被用在 Web 开发上。

XML 具有简单、开放、中立、可扩充、国际化、高效、易管理、无二义等特点。与 HTML 语言相比，区别主要在三方面：

可扩展性方面：HTML 不允许用户自行定义他们自己的标识或属性，而在 XML 中，用户能够根据需要，自行定义新的标识及属性名，以便更好地从语义上修饰数据。

结构性方面：HTML 不支持深层的结构描述，XML 的文件结构嵌套可以复杂到任意程度，能表示面向对象的等级层次。

可校验性方面：HTML 没有提供规范文件以支持应用软件对 HTML 文件进行结构校验；而 XML 文件可以包括一个语法描述，使应用程序可以对此文件进行结构校验。

下图为 SGML、HTML、XML 文件组成的比较[4]：

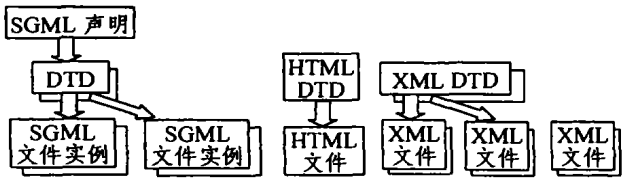

图1 SGML、HTML、XML 文件组成的比较

3 XML 相关标准分类

3.1 XML 标准体系框架

虽然 XML 标准本身简单，但与 XML 相关的标准却种类繁多，W3C 制定的相关标准就有二十多个[5,6]，采用 XML 制定的重要的电子商务标准就有十多个。这一方面说明 XML 确实是一种非常实用的结构化通用置标语言，并且已经得到广泛应用；另一方面，这又为学习了解这些标准带来一定的困难，除了标准种类繁多外，标准之间通常还互相引用，特别是应用标准，它们的制定不仅仅使用的是 XML 标准本身，还常常用到了其他很多标准。笔者通过分类总结，给出了 XML 标准的体系框架，并从体系角度讨论了各标准的功能及相互关系。

XML 标准的体系与 SGML 标准的体系非常相似，如图2所示。XML 相关标准也可分为元语言标准、基础标准、应用标准三个层次。

元语言标准(meta-Language)：描述的是用来描述标准的元语言。在 XML 标准体系中就是 XML 标准，是整个体系的核心，其他 XML 相关标准都是用它制定的或为其服务的。

基础标准(Foundation Standards)：这一层次的标准是为 XML 的进一步实用化制定的标准，规定了采用 XML 制定标准时的一些公用特征、方法或规则。如：XML Schema 描述了更加严格地定义 XML 文档的方法，以便可以更自动化处理 XML 文档；XML Namespace 用于保证 XML DTD 中名字的一致性，以便不同的 DTD 中的名字在需要时可以合并到一个文档中；XSL 是描述 XML 文档样式与转换的一种语言；XLink 用来描述 XML 文档中的超链接；XPointer 描述了定位到 XML 文档结构内部的方法；DOM 定义了与平台和语言无关的接口，以便程序和脚本动态访问和修改文档内容、结构及样式，等等。本文的下节内容将更为详细地讨论这些基础标准及其相互关系。

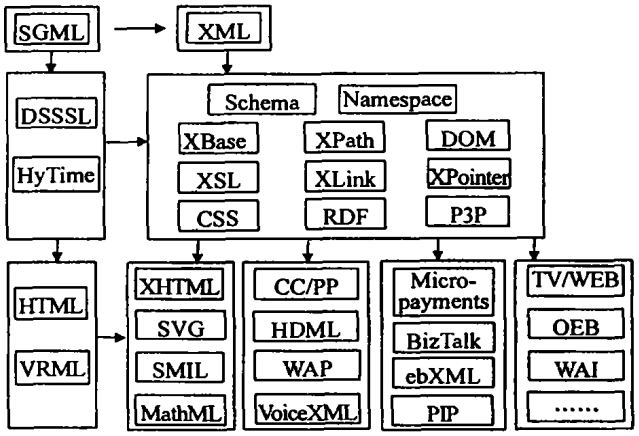

图2 XML 标准体系

应用标准(Application Standards)：XML 已开始被广泛接受，大量的应用标准，特别是针对 INTERNET 的应用标准，纷纷采用 XML 进行制定。有人甚至认为，XML 标准是 Internet 时代的 ASCII 标准。在这 Internet 时代，几乎所有的行业领域都与 Internet 有关。而它们一旦与 Internet 发生关系，都必然要有其行业标准。而这些标准往往采用 XML 来制定。当前较为重要的应用标准主要包括：用于 XML 显示的标准：XHTML(采用 XML 对 HTML 的重新定义)、SVG(有关矢量图形的)、SMIL(有关多媒体同步显示的)、MathML(有关数学公式符号的)；用于移动设备的标准：CC/PP(移动设备的内容协商与信息交换)、HDML(手持设备)、WAP(无线应用设备)、VoiceXML(通过语音进行 Web 访问)；用于电子商务领域的标准：Micropayments(W3C 制定的)、BizTalk(Microsoft 发起的电子商务的 schema 库)、ebXML(联合国 UN/CEFACT 小组和 OASIS 共同发起的)、PIP(由诸多 IT 业的巨子组成的一个标准化组织 RosettaNet 的应用网络标准)、cXML、xCBL、tpaML 等等；以及其他领域的，如：TV/Web(Web 电视)、OEB(电子图书)、WAI(方便残障人进行 Web 访问)；等等。

3.2 XML 基础标准及其相互关系

从 XML 标准体系中可以看到 XML 基础标准是相当多的，而且这些标准在标准体系中占有重要的地位，是 XML 应用标准的基础。它们是在 XML 标准的基础上，进一步对 XML 中一些公共的特性、方法、规则并保证它们之间的一致性作了更为详细明确的规定。应用标准通常都要使用到这些标准的内容或者遵照其中的约定。在使用 XML，阅读利用 XML 应用标准时也需要理解这些标准的内容。由于篇幅限制，本文不讨论这些标准的细节内容，而是从整个体系的角度，讨论各标准在体系中的地位作用及其相互关系，以便能从整体上对 XML 标准体系有个清晰的认识。

图3是 XML 基础标准的框架图。XML 基础标准根据其功能可分为五组：外围标准、核心标准、操作标准、样式与链接标准、内容描述标准。它们分别从不同的方面为 XML 的应用作了更为详细明确的规定。

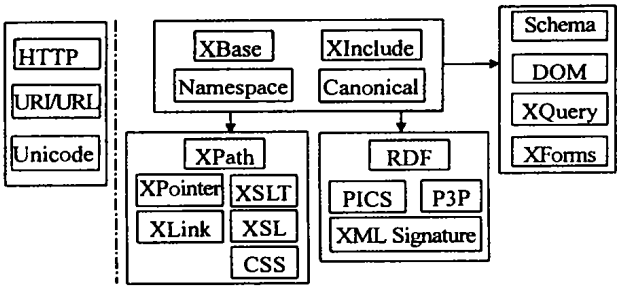

图3 XML 基础标准的框架图

1)外围标准 图3的虚线左侧的三个标准是 XML 相关标准的外围标准,这些标准并不是针对 XML 标准应用或采用 XML 标准制定的,但它们是 Web 应用的基础,几乎在 Web 应用的任何地方都会使用到它们。这些标准对 Web 应用具有确定体系框架意义。

HTTP(超文本传输协议)是在 Web 中应用最为广泛的一种应用层的协议,采用请求/应答方式,客户端发送请求信息到服务器端,这些信息包括:请求方式、URI 和协议版本以及客户端信息等。服务器端返回状态信息、实体信息以及可能有的实体内容。当前浏览器进行网站页面的浏览都是采用这一协议。URI/URL(统一资源标识符/统一资源定位器)是用来定位 Internet 上资源,以便在庞大的 Web 信息系统中能唯一地标识任何一个资源。这种标识是在 Web 上进行信息访问的前提和基础。Unicode 是在 Web 应用中广泛采用的一个字符编码标准,它将几乎世界上所有的文字都包括进去了。它的制定者 Unicode 策进会与相关国际组织密切合作,Unicode2.0版和 ISO10646-1使用完全相同的字库与编码。XML 标准要求 XML 分析器必须至少支持 UTF-8/16编码的 Unicode 字符。

2)核心标准 在 XML 基础标准中,仅次于 XML 标准本身,居于核心地位的是一组 XML 的核心标准,如图3中上部所示。它们是几乎被其它所有 XML 相关标准采用的一组标准。这组标准是由 XML 核心工作组(XML Core Working Group)制定的,为 XML 标准提供最为基础的支持。

核心标准主要包括:XML Base,用于定义 XML 文档的 URI 的基础部分标准,与 HTML BASE 相似。XML Inclusions (XInclude),用于规定文档中包含物的处理模型与语法规则,包括规定如何合并这些包含物的信息,如何使用类似 XML 的语法进行合并控制等。Namespaces in XML,提供了一种简单的方式,用来在 XML 文档中通过与由 URI 引用标识的命名空间相关联来限定元素和属性,提供了解决多 DTD 的 XML 文档中元素名、属性名冲突的基本方法。由于 XML 标准越来越丰富,命名空间变得越来越重要。Canonical XML 是一个还在制定中的标准,提供了一种判断文档等效的方法,它描述了一种对输入的 XML 文档生成范式的方法,这个范式不会因为文档采用的句法形式的改变而改变。对被一个应用改变的 XML 文档,如果它的范式没有改变的话,那么对多数应用来说,改变前后的两个文档是等效的。另外,XML 核心工作组还在制定 XML 语法、片断、信息集等核心标准。

3)操作标准 图3中的右侧四个标准为 XML 文档的处理提供有效的方法与规则。Schema 是对 DTD 的补充,提供了一种更为严格的描述 XML 文档的结构、属性、数据类型等的方法,以便可对 XML 文档进行更加严格的自动化处理。DOM 定义了一组与平台和语言无关的接口,以便程序和脚本能够动态访问和修改 XML 文档内容、结构及样式。XQuery 和 XForms 是两个正处于工作草案阶段制定中的标准。XQuery 的目的是为从 Web 中的实际的或虚拟的文档中提取数据,提供一种灵活的查询机制。为 XML 文档提供一种数据模型,基于数据模型的查询操作,在这些操作基础上的查询语言。它的需求文档已经发布,但在其下定义数据模型却是困难的工作,为此,虽然 W3C XML Query 工作组早已成立,但该标准还处于工作草案需求阶段。XForms 是从 HTML 的"表单"发展抽象而来的。其关键思想是将用户界面和表现与数据模型和逻辑分开,以便同一个表单可被广泛地应用于手持设备、桌面设备或基于语音的浏览器等各种情况。XForms 将 XML 的优点带入到 Web 表单中,采用 XForms 进行数据传输可以减少脚本语句,使得不必为实现表单的布局而将表单嵌入表格中等等。

4)样式与链接标准 XSL、XLink 是一组描述 XML 文档链接、显示、转换的标准。这组标准在 HTML 标准中已有其雏形:显示与链接,并且是 HTML 中最为重要与常用的内容。这组标准的内容充分继承了 SGML 标准中 DSSSL 与 HyTime 相关内容。

在这组标准中 XPath 描述如何识别、选择、匹配 XML 文件中的各个构成元件,包括元素、属性、文字内容等。该标准最初是从 XSL 标准中分离出来的,但由于其定义的是 XML 中一种常用的功能,为了 XML 标准本身的一致性,该标准不再仅仅为 XSL 标准服务,当需要进行 XML 文档内部元素定位时都采用该标准规定的方法与规则。XPointer 充分地利用了 XPath 的内容,并在它基础上进行了扩展。XPointer 和 XLink 标准,继承了 HyTime 标准中有关定位、链接方面的内容,链接采用单独的元素形式,并在标准中定义了"元元素",以便作为模板或父元素类型,链接可以有多种形式等。同时将 XPointer 与 XLink 分开制定为两个标准。

CSS 开始制定时是用来进行 HTML 文档的显示,随着 XML 的出现与发展,它也被用来作为 XML 文档显示的样式标准,并将继续使用下去。但是,XML 出现的目的并不在于它的显示,而主要在于对数据内容本身的描述,其中数据转换是其重要内容之一。为此,很快出现了制定 XML 转换标准的需求,XSL 孕育而生。XSL 的制定借鉴了 DSSSL 和 CSS 的内容与经验,如 XSL 标准采用对 XML 文档形成树状结构,采用元素节点匹配的方式进行转换、采用格式化对象的方法进行显示格式的定义。

XSL 标准可分为两部分:显示部分与转换部分。由于 XML 显示的部分 FO 的复杂性与争议性,使得 XSL 仍未成为正式推荐标准。而用于 XML 转换的标准很快从 XSL 标准中分离出来,形成标准 XSLT,并率先成为正式推荐标准。在当前的浏览器中进行 XML 文档的显示,实际上采用的是其转换标准,将 XML 文档转换为 HTML 文档,进行显示处理。

5)内容描述标准 在图3中,还有一组标准是由 RDF、PICS、P3P、XML-Signature 组成。这组标准中除了 RDF 较常用之外,其他几个标准一般的 Internet 使用者很少直接使用这些标准。但它们是采用 XML 定义的,用来描述资源内容的元信息、安全信息、隐私信息等。

RDF (Resource Description Format)是采用 XML 语法格式处理元数据的应用,为描述图像、文档和它们之间的相互关系定义的一个简单数据模型。简单地说,RDF 是用来进行

资源描述，但并不是直接用来描述资源，而是定义了描述资源的规则。PICS (The Platform for Internet Content Selection) 提供了一种标注 Internet 内容特殊属性的方法。P3P (Platform for Privacy Preferences) 采用 XML 提供了一种进行隐私策略的描述格式，以便保护 Internet 使用者的个人隐私信息或其他保密信息，不会未经允许而被站点或他人获取。XML Signatures 提供的是一种描述对 XML 文档进行数字签名的方法。采用 XML 的语法描述数字签名的方法、计算和验证签名的处理方式，以便保证数据的完整性、可信任性和不可抵赖性。

4 问题与发展

SGML 由于具有坚实的理论基础和长时间发展积累的丰富的经验，已形成了较为完善的理论体系与标准体系。XML 以 SGML 为其基础，使得 XML 迅速形成其标准体系，并被广泛接受，其应用范围甚至超出 W3C 制定 XML 时的预期应用范围。在这些应用的刺激下，出现了许多新的问题亟待研究。

1) *数据模型* XML 在被广泛采用后，数据量急剧增大。如何对这些数据有效地存取、查询、挖掘，已成为 XML 被进一步使用的瓶颈问题。这其中的理论基础问题是 XML 的数据模型问题，以及在该模型上建立起来的存储、索引、查询机制。这方面的研究主要有 Pennsylvaniad 的 UnQL 计划、AT&T 的 StrucQL 计划与 XML-QL 计划、Stanford 在 Lore 计划等[7]。

2) *功能的完善与增加* 当前 XML 的许多标准都在不断的完善增强之中，如：正处于草案阶段的 DOM 3将比 DOM 2 具有更强的事件处理能力和文档操作能力；刚提出的标准草案 CSS 3比 CSS 2具有更丰富的图形处理、文字布局与媒介描述能力。还有许多新的功能需求的出现，使许多标准正在制定中，如：文档范式的生成与表达、文档的安全性保证等等。

3) *行业应用标准的制定* XML 最大特点之一在于它的可扩展性，各行业可根据自身的应用需求约定 DTD 或 Schema，制定行业标准。如电子商务标准。XML 要在各行业广泛应用，需要制定统一的行业标准。

4) *支持工具* 为了 XML 得到更高效的应用，需要相关的支持工具。特别是存储工具和编辑制作工具。当前在存储工具方面主要有两种模式：a) 以关系数据库为基础，增加转换接口。当前大型的数据库厂商主要采用这种方式支持 XML 的存储、管理，如 Oracle、Microsoft。b) 以新的数据模型为基础，进行直接的 XML 数据的存储、管理，如 Software AG 公司的 Tamino Server、eXcelon 公司的 eXcelon Sever。

但这两种模式都有其缺陷，第一种模式由于是基于转换的间接方式，其功能与效率都受到极大的限制。而第二种模式由于其数据模型的理论基础与技术经验的限制，还很不成熟，但它是 XML 存储的发展方向。

当前 XML 的编辑制作工具已经较多了，如浏览工具：Microsoft IE、Mozilla、Amaya；分析工具：IBM XML4J、MSXML；编辑工具：XMLWrite、XML Spy 等。但这些工具在功能、实用性、易用性方面都有待改进。

除了以上这些问题外，语义 Web (Semantic Web) 概念的提出，将在理论体系、标准制定以及应用模式等方面都将为 XML 带来新的研究课题。

结束语 本文首先概述了 XML 的发展与特点。针对 XML 标准的复杂体系，笔者通过总结，给出了 XML 相关标准的体系框架，对基础标准进行了分类，并讨论了它们之间的相互关系。文章最后讨论了 XML 相关的研究热点与发展方向。

参 考 文 献

1 W3C. "Extensible Markup Language (XML) 1.0" W3C Recommendation. February 10, 1998. http://www.w3c.org/TR/1998/REC-xml-19980210.html
2 Goldfarb C F. The SGML Handbook. Clarendon Press. Oxford 1990
3 Goldfarb C F, Prescod P. The XML Handbook. Prentice Hall PTR, 1998
4 Wang Yong-Qun, Chang Ming. Structured Markup Language and Its Application in Database System. Computer Applications (in Chinese), 1998(10):13～16
5 W3C. http://www.w3c.org/
6 XML. 中国论坛. http://www.xml.net.cn
7 McHuge J. Data Management and Query Processing For Semistructured Data:[Dissertation]. Stanford University, 2000.3

(上接第39页)

结束语 本文运用自适应控制论建立 QoS 自适应控制模型，提出形式化评价体系。这是跨领域的结合，为研究 QoS 自适应机制提供一个崭新的思路。今后的工作包括：进一步用现有 QoS 自适应机制来验证 QoS 自适应控制和评价模型的有效性；进一步研究各性能参数之间的相互关系；应用提出的模型和控制理论来指导实际的策略和算法设计。

参 考 文 献

1 Yeadon N J. QoS Filtering for Multimedia Communications. Lancaster:[PhD thesis]. 1996.5
2 McCanne S R. Scable Compression and Transmission of Internet Multicast Video. Berkeley:[PhD thesis]. 1996.10
3 Semret N, et al. Market Pricing of Differentiated Internet Services. In: Proc. of the 7th Intl. Workshop on Quality of Service (IEEE/IFIP IWQoS'99), London: UK, June 1999
4 Semret N. Market Mechanisms for Network Resource Sharing: [PhD thesis]. Columbia Univarsity, 1999. Available at: http://comet.columbia.edu/~nemo/work.html
5 Courcoubetis C, Siris V A. Managing and pricing service level agreements for differentiated services. In: Proc. of 6th IEEE/IFIP Intl. Conf. of Quality of Service (IWQoS'99), London: UK, May-June 1999
6 Kelly F P, Maulloo A, Tan D. Rate control for communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research Society, 1998, 49(3):237～252
7 Amir E. An Agent-based Approach to Real-time Multimedia Transmission over Heterogeneous Environments. Berkeley University, 1998
8 Li Baochun, Nahrstdt K. Configurable Adaptors for Multimedia Delivery an End System Middleware Solution. University of Illinois at Urbana-Champaign, 1997
9 Li Baochun. Adaptive Behavior of Quality of Service in Distributed Multimedia Systems. Urbaba, Illinois, 1997
10 Li Baochun, Nahrstedt K. A Control Theoretical Model for Quality of Service Adaptations. Urbaba, Illinois, 1997
11 卢志恒. 控制论引论. 北京师范大学出版社
12 Chatterjee S. ERDoS QoS Architecture:[Technical Report]. May 1998
13 Huard J-F, Lazar A A. On End-to-End QoS Mapping. Available at: http://comet.ctr.columbia.edu/
14 Kelly F P. Charging and rate control for elastic traffic. European Transactions on Telecommunications, 1997, 8(1):33～37
15 Roberts J W. Bandwidth sharing and admission control for elastic traffic