

云计算环境中结构化文档形式化建模

熊金波*, 姚志强, 金彪

(福建师范大学 软件学院, 福州 350108)

(* 通信作者电子邮箱 jinbo810@163.com)

摘要:结构化文档是云计算环境中实现信息交互与传播的载体,针对已有研究工作缺乏能够描述这种载体的结构化文档模型的问题,在深入分析云计算环境中结构化文档特征的基础上,对文档元素、原子文档元素、连接、文档片段、复杂文档元素等进行形式化定义,建立能够满足这种特征的结构化文档形式化模型。最后,指出云计算环境中结构化文档进一步的研究方向。

关键词:云计算;结构化文档模型;形式化方法;文档元素

中图分类号: TP393 **文献标志码:** A

Formal modeling for structured document in cloud computing

XIONG Jinbo*, YAO Zhiqiang, JIN Biao

(Faculty of Software, Fujian Normal University, Fuzhou Fujian 350108, China)

Abstract: Structured document plays a vital role in information carrier for information interaction and propagation in cloud computing. However, the existing research works lack structured document model that can describe this information carrier. In order to tackle this issue, the authors formalized the definitions of document element, atomic document element, connection, document fragment, and complex document element based on the in-depth analysis of the special characteristics of structured document in cloud computing. Furthermore, the authors proposed structured document model (referred to as the SDoc) satisfying the special characteristics in cloud computing. Finally, several open issues were indicated for structured document.

Key words: cloud computing; structured document model; formalization method; document element

0 引言

计算技术的快速发展以及面向服务体系结构的广泛应用推动计算模式的不断演化,云计算^[1]应运而生,它有效地将大规模且分布式的计算、通信、存储、平台和基础设施等资源虚拟化^[2],以云服务的形式按需提供给广大用户,不断满足用户对资源的多方位需求。

云计算环境中,云服务的供给、交互以及各服务间的组合与协同等需要数据媒介来实现,而结构化文档^[3]正是云计算环境中服务提供、信息交互和传播的数据载体。建立合适的结构化文档模型对于云计算环境中信息交互、服务协同等具有非常重要的理论意义,对于促进云计算技术的快速发展具有关键推动作用。

国内外对文档已有较深入的研究,覆盖多个研究领域,如文档模型、文档数据处理、文档结构、文档分析、文档检索、文档存储和文档安全等。结构化文档由物理结构和逻辑结构描述^[3],前者反映文档的布局结构和展示效果,后者反映文档的内容组织,体现文档的组织结构,也是本文关注的重点。已有关于结构化文档模型的研究主要为相似性检测文档模型^[4-5]、保护机密文档的文档模型^[6]、业务文档模型^[7-8]和多媒体文档模型^[9-10]等。而以上文档具有内容相对固定且版本较单一等特征。与之不同的是,云计算环境中的结构化文档不再是静态内容和单一版本^[11],而是具有动态内容、多用

户参与、多媒体交互与多版本共存的“活”文档;且存在多文档之间、文档内部不同元素之间存在动态关联及组合,同一文档被不同安全域的多用户协作编辑,在线编辑与实时呈现,在线文档与物理文档的紧密耦合等新型复杂特征。

据大范围查阅国内外科技文献,尚无云计算环境中有关结构化文档模型的研究。因此,本文首先着重研究云计算环境中的结构化文档特征,在此基础上,建立云计算环境中结构化文档形式化模型(Structured Document formal model, SDoc),最后针对结构化文档模型,指出进一步的研究方向。

1 云计算环境中的“活”文档

云计算环境中,结构化文档不再是内容静态和单一的文档^[11],相反,从文档的创建、编辑、修改、传播、使用、存储到最终的销毁都具有内容动态、多用户参与、多媒体交互、多安全等级和多版本共存等“活”文档特征。

内容动态 用户可以产生各种类型的动态文档,多个文档之间、单文档元素之间或者单元素与多文档之间的关联、聚合、组合和重组等都能产生动态内容。

多用户参与 为了某一特定目的,一个文档在创建之后可能吸引多个用户或者潜在的参与者共同编辑、修改和交互信息。

多媒体交互 结构化文档可能体现出多媒体特征,不仅仅是文本、图片、图像、文件、声音、视频和图表等单一的多种

收稿日期:2012-11-08;修回日期:2012-12-13。

基金项目:福建省自然科学基金资助项目(2011J01339);福建省教育厅科技项目(JA12078, JB12022, JB11034)。

作者简介:熊金波(1981-),男,湖南益阳人,讲师,博士研究生,CCF会员,主要研究方向:文档工程及文档安全、访问控制;姚志强(1967-),男,福建莆田人,教授,CCF会员,主要研究方向:信息安全;金彪(1985-),男,安徽六安人,助理实验师,硕士,主要研究方向:信息安全。

媒体组合。比如病人的电子医疗记录 (Electronic Health Record, EHR)^[12] 就是一个由门诊记录表、X 射线影像片、血液生化记录、CT 扫描影像片和 B 超视频记录等组成的多媒体文档。

多安全等级 结构化文档由组织成层次结构的元素组成,每个层次以及每层的元素都有相应的安全等级标记^[13]。因此,具有不同安全等级的元素可以组成多级安全文档。比如在 EHR 文档中,包含病人隐私信息,如姓名、年龄、单位、生日和传染性疾病或特殊病种等都应该分配高安全级别,其他一般病理特征的描述可以设定较低的安全等级。

多版本共存 结构化文档可能仅仅由于某元素或内容的细微改变或者布局的调整而产生多个版本,由于每个版本都非常重要不能覆盖,从而产生多版本共存^[14]。同一个文档对于不同用户、或者对同一个用户在不同时间和不同地点都应该呈现出不同的版式,从而亦可产生多版本。

综上所述,结构化文档在云计算环境中具有独特的“活”文档特征,已有的文档模型不能足够描述以上特征。因此,需要建立云计算环境中结构化文档新模型。

2 结构化文档的层次化组织和安全等级

为了更清晰地表达结构化文档元素之间的关系,本文将云计算环境中的结构化文档组织成层次化树型结构,如图 1 所示。文档元素和内容被组织成多个层次,如子文档、章、节、段、字段或区域等,每个层次包含多个元素,每个元素还可以包含多个子元素等。

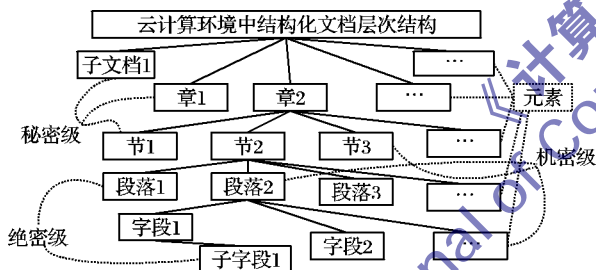


图 1 云计算环境中结构化文档树型结构

基于图 1 所示的树型结构,结构化文档的拥有者能够根据文档中各敏感元素的安全需求,灵活地分配安全等级。依据文献^[13]将安全等级分为绝密级、机密级、秘密级和无密级 4 类。图 1 中,第 2 章中第 2 节的段落 1 和段落 2 中字段 1 的子字段 1 被分配绝密级,第 2 章的第 3 节和第 2 节的段落 2 被分配机密级,子文档 1、第 1 章和第 2 章的第 1 节被分配秘密级,该文档中其余元素均为无密级的。从而,拥有绝密级权限的用户可以访问段落 1 和子字段 1 中的元素内容,以及安全等级低于绝密级的其他文档内容;拥有机密级权限的用户可以访问第 3 节和除子字段 1 之外的段落 2 的内容,以及安全等级低于机密级的其他文档内容。

3 云计算环境中结构化文档模型 SDoc

形式化方法^[15]能够对信息系统和文件系统进行精确的描述,因此,在给出 SDoc 基本概念的基础上,建立了 SDoc 的形式化框架。

3.1 SDoc 基本概念

由图 1 可知,云计算环境中 SDoc 以树型结构描述结构化文档^[9],树的节点表示文档元素 (document element, de),树

的边将文档元素层次化绑定 (bind) 在一起。每个文档元素都有 1 个能够向上连接到其他文档元素的绑定点 (bind point, bp), 也有 1 个或多个能够向下连接到其他文档元素绑定点的连接点 (connect point, cp), 此外, 每个文档元素还有 1 个或多个布局点 (layout point, lp), 连接布局信息, 用于指明文档元素的布局。图 2 是从图 1 抽象出来的云计算环境中 SDoc 中各元素的图解表示。

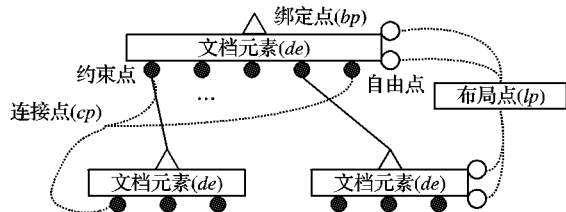


图 2 SDoc 的图解表示

文档元素 *de* 是 SDoc 中的通用元素,它们可以表示文本、图片、视频、动画和音频等文档数据,还可以表示时间、空间、布局和文档元素之间的关联等语义关系。考虑 1 个简单的结构化文档树,称为 SDoc 片段,如图 3 所示。图 3 中,1 份 EHR 文档,文档元素 *ehr*, 绑定门诊记录 (text) 和 X 射线片 (image) 到它的连接点 *cp*₁ 和 *cp*₃, 同时有个平行元素 *cp*₆ 和 *cp*₇ 连接到 *cp*₂, 分别同步 1 段 B 超视频 (video) 和 B 超诊断意见 (text) 数据。每个 SDoc 片段对元素的语义描述都是从根元素开始,依次呈现它的连接点 *cp*₁ 至 *cp*₅ 的数据。图 3 的 SDoc 片段表达了文档元素及语义的相互关系,其中 *ehr* 元素的绑定点还可以连接到其他文档元素中以组成更大更复杂的结构化文档树,而 *ehr* 元素的连接点 *cp*₄ 和 *cp*₅ 仍然为自由点,可以由文档创建者 (医生) 插入诊断结论和对患者的饮食及保健建议等信息。

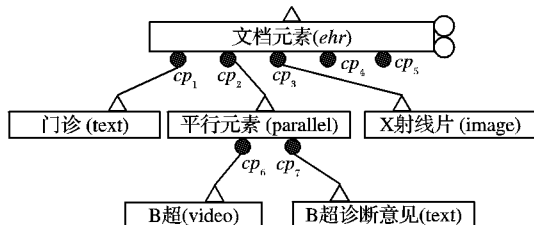


图 3 SDoc 片段示例

3.2 SDoc 形式化框架

SDoc 中用到的符号及描述如表 1 所示。

表 1 SDoc 中的符号及其描述

符号	描述
<i>Name</i>	所有文档名称的集合
<i>Type</i>	所有元素类型的集合
<i>DT</i>	所有文档类型的集合
<i>Attr</i>	所有文档属性的集合
<i>Sec</i>	所有安全等级的集合
<i>U_{ID}</i>	所有用户 ID 的集合
<i>Ver</i>	所有版本的集合
<i>BP</i>	所有绑定点的集合
<i>CP</i>	所有连接点的集合
<i>LP</i>	所有布局点的集合
<i>ODD</i>	所有原始文档数据的集合
<i>Con</i>	所有连接的集合
<i>DE</i>	所有文档元素的集合
<i>SDOC</i>	所有 SDoc 文档的集合

SDoc 中文档元素的形式化定义如下:

定义1 文档元素。文档元素 (de) 是1个元组 $de: \langle n_{de}, t_{de}, a_{de}, s_{de}, bp_{de}, CP_{de}, LP_{de} \rangle$, 其中 $n_{de} \in Name$ 表示 de 的名称, $t_{de} \in Type$ 表示 de 的元素类型, $a_{de} \in Attr$ 表示 de 的文档属性, $s_{de} \in Sec$ 表示 de 的安全等级, $bp_{de} \in BP$ 表示 de 的绑定, $CP_{de} \in CP$ 表示 de 的连接点集合, $LP_{de} \in LP$ 表示 de 的布局点集合。依据 t_{de} 类型的变化, 元组 de 还可以通过增加元组元素的形式进一步扩展^[9]。

SDoc 的基本单元是原子文档元素, 它不再向下连接其他的文档元素。原子文档元素是从原始文档数据中抽象出来用以表达文档元素和文档数据的特殊特征。原子文档元素的形式化定义由定义2给出。

定义2 原子文档元素。原子文档元素 (ad) 是1个元组 $ad: \langle n_{ad}, t_{ad}, a_{ad}, s_{ad}, bp_{ad}, CP_{ad}, LP_{ad}, d \rangle$, 其中 $t_{ad} \in DT = \{Text, Image, Video, Audio, rar, zip\} \subset Type$ 指明 ad 的类型; $CP_{ad} = \emptyset$ 表明 ad 没有连接点, 不再向下连接其他 de ; $d \in ODD$ 指明 ad 表达的文档数据。

文档元素之间通过各自的 bp 和 cp 相互连接 (c), 组成复杂文档元素。连接 (c) 的形式化定义由定义3给出。

定义3 连接。连接是1个元组 $c = \langle bp_{de}, p' \rangle$, 表示连接 de 的 bp 与其他 de 的 cp 或者 lp 。其中 bp_{de} 为 de 的绑定, $p' \in CP_{de'} \cup LP_{de'}$ 表示 de' 的连接点或布局点, 且 $de \neq de'$ 。

文档元素连接的结果构成1个子树, 该子树表示1个可以进行文档重组或重用的单元, 称为SDoc片段, 如图3所示实例。1个有效的SDoc片段可以由定义4形式化描述。

定义4 SDoc片段。片段 $f = \langle DE_f, Con_f \rangle$ 是1个子树, 是非循环无向图, 描述的是1个完整或局部结构化文档。其中 DE_f 表示子树中所有 de 的集合, $Con_f \subseteq \{ \langle bp_{de}, p' \rangle \mid de, de' \in DE_f, de \neq de', p' \in CP_{de'} \cup LP_{de'} \}$ 表示子树中所有连接的集合。

对于1个有效的SDoc片段 $f = \langle DE_f, Con_f \rangle$, 下列条件必须满足:

1) 如果2个连接 $c_1, c_2 \in Con_f$, $de_1, de_2 \in DE_f$, $c_1 = \langle bp_{de_1}, p' \rangle$, $c_2 = \langle bp_{de_2}, p' \rangle$, 则满足 $de_1 = de_2$, 即每个连接点或布局点只能连接到唯一的绑定。

2) 如果2个连接 $c_1, c_2 \in Con_f$, $c_1 = \langle bp_{de}, p_1 \rangle$, $c_2 = \langle bp_{de}, p_2 \rangle$, $de \in DE_f$, 则满足 $p_1 = p_2$, 即每个绑定只能连接到唯一的连接点或布局点。

3) 不存在连接序列 $c_1, c_2, \dots, c_n \in Con_f$, 满足 $c_i = \langle bp_{de_i}, p_i \rangle$, $i = n-1, p_{i+1} \in CP_{de_i}, p_1 \in CP_{de_n}$ 表明每个SDoc片段都是非循环的。

4) $Uncon_f = \{ de \in DE \mid \neg \exists p \in \bigcup_{de' \in DE} (CP_{de'} \cup LP_{de'}) : \langle bp_{de}, p \rangle \in Con_f \}$, $|Uncon_f| = 1$, $root_f \in Uncon_f \wedge t_{root_f} \notin LT$ 。表明每个SDoc片段只能有唯一的文档元素 $de \in DE$, 其绑定 bp_{de} 没有和其他文档元素 de' 建立连接关系, 该文档元素称为SDoc片段的根元素, 表示为 $root_f$ 。根元素的绑定 bp_{root_f} 是自由点, 为该SDoc片段的入口。 $root_f$ 不能为布局元素。

5) $\forall lp \in \bigcup_{de \in DE} LP_{de} : \exists \langle bp_{de'}, lp \rangle \in Con_f \Rightarrow t_{de'} \in LT$, 表明文档元素的布局点只能连接布局元素。

6) $\forall cp \in \bigcup_{de \in DE} CP_{de} : \exists \langle bp_{de'}, cp \rangle \in Con_f \Rightarrow t_{de'} \notin LT$, 表明文档元素的连接点不能连接布局元素。

SDoc片段即为结构化文档的1个可重用的文档块, 可以实现不同结构化文档之间的关联和重组, 以满足云计算环境中复杂的结构化文档使用需求。为了降低SDoc片段重用的复杂度, 本文引入复杂文档元素 (cd) 的定义, 将SDoc片段 $f = \langle DE_f, Con_f \rangle$ 封装在1个单一的文档元素内而构成1个复杂文档元素。复杂文档元素的形式化描述由定义5给出。

定义5 复杂文档元素。复杂文档元素 (cd) 是1个元组 $cd: \langle n_{cd}, t_{cd}, a_{cd}, s_{cd}, bp_{cd}, CP_{cd}, LP_{cd}, f \rangle$, 表示将SDoc片段 $f = \langle DE_f, Con_f \rangle$ 封装到1个 cd 中, 其中 $t_{cd} = Complex \in Type$, $bp_{cd} = bp_{root_f}$, $s_{cd} = \min \{ s_{de_i} \mid de_i \in DE_f \}$, $CP_{cd} = \{ cp \in \bigcup_{de \in DE_f} CP_{de} \mid \forall de' \in DE_f : \langle bp_{de'}, cp \rangle \notin Con_f \}$, 且 $LP_{cd} = \{ lp \in \bigcup_{de \in DE_f} LP_{de} \mid \forall de' \in DE_f : \langle lp_{de'}, lp \rangle \notin Con_f \}$ 。

定义5的语义含义为: cd 的类型为复杂类型 $Complex$, SDoc片段 f 中根元素 $root_f$ 的绑定 bp_{root_f} 即为 cd 的绑定 bp_{cd} , cd 的安全等级为 f 中所有文档元素安全等级中的最低值, f 中所有的自由点和没有连接的布局点组成 cd 的连接点和布局点。

将图3所示的SDoc片段封装成 cd 之后如图4所示, cd 犹如单一文档元素, 其连接点 cp_4 和 cp_5 可以连接其他文档元素或外部文档元素, 以方便组建大型结构化文档, 以及灵活实现结构化文档之间的关联和重组等云计算环境下复杂的结构化文档操作。



图4 SDoc片段封装成复杂文档元素

下面给出结构化文档SDoc的形式化定义。

定义6 结构化文档。结构化文档是1个元组 $SDoc_i: \langle DE_i, Con_i, UID_i, Ver_i, S_{Doc_i} \rangle$, 其中 $\langle DE_i, Con_i \rangle$ 为满足某一特定需求的所有文档元素、原子文档元素和复杂文档元素, UID_i 为用户的ID信息, 每次 de 或者 c 的改变都会在 UID_i 中增加1条记录, 相应地在 Ver_i 中增加1条记录, 体现出结构化文档的多用户和多版本特征。 $S_{Doc_i} = \min \{ s_{de_k} \mid de_k \in DE_i \}$, 表明SDoc的安全等级为SDoc内所有文档元素安全等级中的最低级, 拥有该安全等级的用户可以访问此SDoc文档, 遇到文档中安全等级高的文档元素时, 则该用户不能访问, 这需要设计多级安全的访问控制策略实现。

综上所述, SDoc形式化模型满足云计算环境中结构化文档的动态性、多用户、多媒体、多版本和多安全等级的“活”文档特征。XML (eXtensible Markup Language) 以其开放性、可扩展性、灵活性、结构化、可验证性和计算机系统与编程平台独立性等诸多优秀特征已经成为信息系统中数据交换与互操作的事实标准^[16], 尤其在面向服务的应用和Web服务中。因此, 下一步将结合XML和结构化文档形式化模型建立基于XML的结构化文档模型描述方法, 使该模型更好地表达云计算环境中结构化文档的特征, 抽象结构化文档的内容层次结构, 保护结构化文档安全, 以利于云计算环境中信息的表达与交互, 促进云计算技术的快速发展。对图4的复杂文档元素

进行 XML 描述如下所示:

```
<SDoc document >
  <title> Complex document element </title>
  <name> ehr </name>
  <type> complex document </type>
  <attribute> electronic health document </attribute>
  <security> secret </security>
  <data> ... </data>
  <bindpoint> bp </bindpoint>
  <data> ... </data>
  <connectpoint> cp4 </connectpoint>
  <operator> ... </operator>
  <layoutpoint> lp </layoutpoint>
  <operator> ... </operator>
</SDoc document >
```

4 开放问题及下一步工作重点

1) SDoc 中各种操作的定义、文档的存储与检索 SDoc 的操作可以分为以下几类:时间操作、空间操作^[9]、交互操作、重用操作、自适应操作等。需要对以上操作进行形式化定义。此外,还需要基于 XML 实现以上各类操作,并研究基于 XML 的结构化文档存储方法,结合云计算数据中心及 Oracle 等大型数据库研究结构化文档的存储实现。最后,研究基于 XML 的结构化文档快速检索和加密文档的检索算法,以提高云计算环境中对结构化文档的检索能力。

2) 云计算环境中结构化文档模型的多级安全需求。云计算环境中复杂的网络环境、结构化文档的特殊性以及用户对结构化文档访问方式和访问类型的多样性等对结构化文档的多级安全带来新的问题和安全需求^[17]。需要从用户安全等级、网络环境安全等级、用户访问行为安全等级和结构化文档安全等级等方面考虑云计算环境中结构化文档模型的多级安全需求。下一步的研究工作需要抽象提取多级安全需求的相关属性和核心因素,并对其进行形式化描述,从而为建立结构化文档的多级安全模型提供基础。

3) 结构化文档多级安全模型的可能实现。多级安全模型根据安全标记实现主体对客体的访问控制^[13]。因此,需要在结构化文档多级安全模型中保存用户、用户访问行为^[18]以及文档元素和属性的安全标记。在结构化文档模型中,使用 Sec 来标记结构化文档元素和属性的安全级别,连同用户和用户访问行为的安全标记一起建立安全 XML 文档,以便扩展已有的 XML 文档系统,实现对结构化文档的多级安全管理。在此基础上,下一步需要研究云计算环境中结构化文档多级安全模型的可能实现方法。

5 结语

在云计算环境中,建立合适的结构化文档模型对于信息交互与传播、服务组合与协同等具有非常重要的理论和实际意义,对于促进云计算技术的快速发展具有关键推动作用。本文探索云计算环境中的结构化文档形式化建模问题,为结构化文档的进一步研究抛砖引玉。本文的主要工作如下:首先,深入分析云计算环境中结构化文档的特征,并构建结构化文档的层次结构;然后,形式化定义结构化文档 SDoc 的文档元素、原子文档元素、连接、SDoc 片段、复杂文档元素,构建 SDoc;最后,指明了云计算环境中结构化文档的若干开放课题

与进一步的研究方向。

参考文献:

- [1] ARMBRUST M, FOX A, GRIFFITH R, *et al.* A view of cloud computing[J]. Communications of the ACM, 2010, 53(4): 50 - 58.
- [2] 陈全,邓倩妮. 云计算及其关键技术[J]. 计算机应用, 2009, 29(9): 2562 - 2567.
- [3] 汤帆,仇睿恒,王毅. CEBX: 新一代结构化版式文档技术[J]. 北京信息科技大学学报, 2010, 25(Z2): 11 - 15.
- [4] WEI X, CROFT W B. LDA-based document models for Ad-Hoc retrieval[C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006: 178 - 185.
- [5] DENG H B, KING I, LYU M R. Enhanced models for expertise retrieval using community-aware strategies[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42(1): 93 - 106.
- [6] ZHENG S Y, LIU J. A secure confidential document model and its application[C]// Proceedings of the 2010 International Conference on Multimedia Information Networking and Security. Washington, DC: IEEE Computer Society, 2010: 516 - 519.
- [7] PICHLER C, HUEMER C, WIMMER M. Business document interoperability as a service[C]// Proceedings of the 6th International Symposium on Service Oriented System Engineering. Washington, DC: IEEE Computer Society, 2011: 318 - 326.
- [8] LIEGL P, HUEMER C, PICHLER C. Registry support for core component-based business document models[J]. Service Oriented Computing and Applications, 2011, 5(3): 183 - 202.
- [9] BOLL S, KLAS W. ZYX-A multimedia document model for reuse and adaptation of multimedia content[J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(3): 361 - 382.
- [10] TRAN N A T, DANG T K. An extended video database model for supporting finer-grained multi-policy and multi-level access controls[J]. Polibits, 2008, 38: 49 - 62.
- [11] BALINSKY H, SIMSKE S J. Secure document engineering[C]// Proceedings of the 11th ACM Symposium on Document Engineering. New York: ACM Press, 2011: 269 - 272.
- [12] ALHAQBANI B, FIDGE C. Access control requirements for processing electronic health records[C]// Proceedings of the 2007 International Workshops on Business Process Management, LNCS 4928. Berlin: Springer, 2008: 371 - 382.
- [13] 熊金波,姚志强,马建峰,等. 视频数据库多级访问控制[J]. 通信学报, 2012, 33(8): 147 - 154.
- [14] CHANG C C, CHENG M S, HSU P Y. Effective storage structure for multi-version XML documents[C]// Proceedings of the Second International Conference on Future Generation Information Technology. Berlin: Springer, LNCS 6485, 2010: 249 - 266.
- [15] BERSTEL J, BOASSON L. Formal properties of XML grammars and languages[J]. ACTA Informatica. 2002, 38(9): 649 - 671.
- [16] 吴洁明,范国梅. 基于 XML 的标准符合性测试方案[J]. 计算机应用, 2012, 32(2): 551 - 553.
- [17] 李风华,苏锐,史国振,等. 访问控制模型研究进展及发展趋势[J]. 电子学报, 2012, 40(4): 805 - 813.
- [18] 李风华,王巍,马建峰,等. 基于行为的访问控制模型及其行为管理[J]. 电子学报, 2008, 36(10): 1881 - 1890.