一种基于视觉特征的 Deep Web 信息抽取方法*

孙 璐 陈军华 廉德胜
(上海师范大学 上海 200234)

摘 要 随着 Web 数据库的不断增长,大量网络信息通过普通搜索引擎难以满足用户的需求,需要用户提交表单查询并从后台数据库中返回结果页面才能获取到想要的信息,称为 Deep Web。因此如何有效地抽取这些实体信息成为一个值得研究的问题。论文通过分析 Deep Web 结果页面的特点,结合人的视觉特征,提出了一种基于视觉特征的 Deep Web 信息抽取方法。该方法充分利用了人的视觉特征,在解析器将 Web 文档解析成语法树之前,将 Web 页面一些与主题无关的信息(例如导航栏、广告)等去除,并对优化后的 DOM 树利用 VIPS 算法对其进行语义分块,分块后根据位置特征首先寻找到基准视觉块,以该基准视觉块作为中心位置逆序和顺序遍历 DOM 树寻找所有相似的视觉块并对其进行抽取。从实验效果来看,该方法从提取信息速度和提取信息的准确率和完整率方面与传统方法相比都有一定的提高。

关键词 Deep Web; 视觉特征; DOM 树; 语义分块; 信息抽取
中图分类号 J653 **DOI**:10.3969/j.issn.1672-9722.2016.06.026

Deep Web Information Extraction Method Based on Visual Features

SUN Lu CHEN Junhua LIAN Desheng
(Shanghai Normal University, Shanghai 200234)

Abstract With the constantly development of Web database, a large number of information can not be got by ordinary search engine. The results which users want to get need them submit the form query so that the information can be got from the database behind called Deep Web. Thus how to effectively extract these information become a problem which worth of study. This paper propose an improved method by analyzing the characteristics of the results pages combining with human visual sense. This method makes full use of human visual characteristics, before the parser parsed the Web document into a syntax tree, and removed some information which has nothing to do with the theme such as navigation, advertising, etc. After that, division the DOM tree into semantic block using VIPS algorithm. Sw we can find the standard block according to the block's position, then put the standard block as center block which used to find all similar visual blocks by reversing and suquential traversal the DOM tree. These result blocks are the information blocks which we want to extraction. According to the experimental results, this method has some improvement from the aspects of accuracy rate and complete rate to some extent compared with traditional method.

Key Words Deep Web, visual characteristics, DOM tree, semantic block, information extraction
Class Number J653

1 引言

信息抽取可以理解为从一段待处理文本中抽取指定的一类信息,并将其以结构化的形式表示(如 XML 等)供用户查询和使用的过程。针对 Web 信息抽取工作目前国内外已展开了大量的研究,并且取得了一定的成果。其中按照抽取技术的不同可以分为基于自然语言处理方式的实体抽取;基于包装器归纳法的信息抽取;基于模板的信息抽取;基于视觉特征的信息抽取和基于 DOM 树的实体抽取技术。其中基于视觉特征的信息抽取和基于 DOM 树的信息抽取是目前应用比较广泛的方法。

* 收稿日期:2015 年 12 月 5 日,修回日期:2016 年 1 月 23 日
作者简介:孙璐,女,硕士研究生,研究方向:数据库。陈军华,男,硕士,副教授,研究方向:数据库。廉德胜,男,硕士研究生,研究方向:人工智能。

由于构成网页的 HTML 语言在很大程度上是用来显示数据而不是展示其内容结构的,所以从用户的视觉角度对 Web 页面进行分析有其一定的合理性。文献[10]提出了基于视觉特征的 VIPS 算法,该算法充分利用了 Web 页面的视觉特征,例如颜色、字体大小、图片等,把 Web 页面划分为许多视觉块,根据视觉块之间的相似度重构页面的内容结构,从而对信息进行抽取。但是该方法基于许多启发式的规则,有时会受人的视觉误导,把页面一些无用的信息当作视觉块处理,例如广告信息等。文献[6~8]提出了基于 DOM 树的实体抽取技术。在该方法中,首先利用解析器将 Web 文档解析成语法树,然后深度遍历整棵 DOM 树,利用 DOM 树节点之间的相似度确定正文区域,从而对文本信息进行有效抽取。但是该方法是把文本节点和标签节点放在一起对整个 DOM 文档进行遍历分析,加大了遍历 DOM 树的时间复杂度。本文通过观察大量 Deep Web 结果页面,首先运用启发式规则对原始页面进行去噪处理,使得去噪后解析 DOM 树的节点数大大减少,然后在 DOM 树结构基础之上,运用文献[10]提到的 VIPS 算法,把 Web 页面分割成许多大小不等的视觉语义块,利用页面中心位置的坐标确定出基准视觉块,然后根据 Web 页面正文信息的位置分布特征和正文视觉块之间的视觉相似性,以该视觉块作为中心位置,顺序和逆序递归遍历整棵 DOM 树,寻找出所有相似视觉块,即要提取的正文信息。实验表明该方法与传统的方法相比有一定的优势。

2 基于基准视觉块的信息抽取算法

2.1 Web 页面去噪

一般的网页可以分为导航型网页和内容型网页两种,由于本文主要是针对特定领域的关键词搜索结果研究,所以不对导航型网页做研究。对于一个已抽取到的 Deep Web 结果页面,需要提取的数据区域往往集中于页面的某个区域,称之为正文区域。而普通的 Deep Web 页面往往包含标题、广告栏、导航链接等许多噪声信息,一些针对特定领域的数据查询(如图书查询),因为它们有规律地分布在页面的特定部分,使得这些无用的噪声信息占了整个页面的一定比重,这样不利于页面的信息抽取,所以对初始页面做去噪处理是非常有必要的。本文通过观察大量的网页后台 HTML 代码并结合文献[9]提到的网页信息去噪技术,得出如下一些启发式规则:

规则一:如果一个节点周围含有大量的链接节点,如〈link〉等,即链接节点数超过了该区域总数的一定比例,在这里取 95%,那么倾向于把这片信息块看作噪声信息,反之则为正文信息;

规则二:如果一个节点的 position 属性为 fixed,并且该节点下还包括 img、object 或 iframe 节点,那么把该节点作为噪声节点;

规则三:如果一个文本节点的文本字数低于版权信息节点所含文本的字数(这里把版权信息的字数作为一个阈值)那么倾向于把它看作噪声节点或无用节点。基于以上一些规则,可以初步对原始 Web 页面做一些优化处理。

本文采用 HTMLPaster 的词法分析器对页面的 HTML 代码进行分析,通过提交关键字查询获取 Deep Web 页面作为实验数据的来源。解析到原始页面的 HTML 代码后,利用上一节提到的启发式规则对页面的噪声进行过滤处理。可以看出,经过处理后 DOM 树的节点数大大减少了。

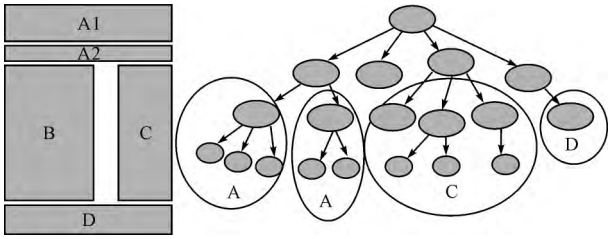

图 1 页面结构和去噪节点对比图

2.2 VIPS 算法

VIPS 算法主要是利用 Web 页面的视觉特征如背景颜色,字体的颜色和大小等把页面分成许多合适的视觉块,根据视觉块之间的逻辑间距重构语义 DOM 树,进而对页面信息抽取的过程。下面对该算法做一个简单介绍。在 VIPS 算法中,一个 Web 页面由 Ω 表示,$\Omega=(O,\Phi,\delta)$。其中 $O=\{\Omega^1,\Omega^2,\cdots,\Omega^N\}$ 是一系列有限的页面块的集合,$\Phi=\{\Phi^1,\Phi^2,\cdots,\Phi^T\}$ 是一系列有限的分隔符的集合,$\delta=O\times O\rightarrow\Phi\cup\{NULL\}$,它表示 O 中每两个块之间的关系[10]。例如,假设 Ω_i 和 Ω_j 是 O 中的两个对象,$\delta(\Omega_i,\Omega_j)\neq NULL$ 表明 Ω_i 和 Ω_j 之间是有联系的,即它们有可能是 *DOM* 树中的两个相邻的节点。另外,在 Ω 中,每一个页面块都可以看作一个子页面,所以可以递归地对它作同样的处理,直到当前页面块不能再分割为止。

下面以当当网为例具体阐述整个分割过程。在当当网首页输入“计算机”,点查询,可以得到如图 2 结果页面。

图 2　当当网页面

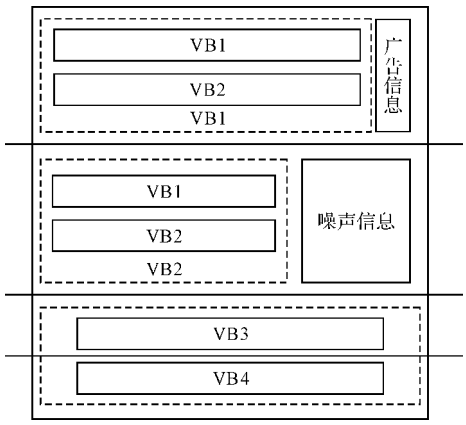

图 3　当当网视觉分块图

根据 VIPS 算法，把该页面分割成如图 3 所示的视觉块，其中 VB1 中主要是查询信息和导航信息，还夹杂了一些广告信息，VB3 和 VB4 是底下一些服务指南和版权信息，VB2 是想要提取的正文信息。可以看到要提取的信息，即 VB2 主要集中在页面的某一特定部位，以 VB2 为例简单说明 VIPS 分块过程。VB2 的 DOM 树结构如图 4 所示。

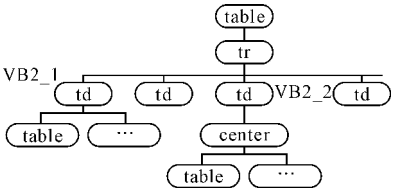

图 4　VB2 的 DOM 树结构

首先得到〈table〉标签，它有孩子节点〈tr〉，且孩子节点的背景颜色和它父亲节点的背景颜色不同，所以分隔这个节点，这样就得到两个节点块，然后分别对这两个节点块进一步分析。它有四个〈td〉节点，其中两个是无效节点，取出剩下的两个有效节点即 VB2_1 和 VB2_2 所在视觉块。分别对两个视觉块深度遍历，得到〈table〉子节点，它有可能是想要的文本信息，所以把它放到分块池中等待进一步被分析。等到所有的节点都被分析完放入池中后，再递归地对分块池中的节点块作同样的分析，直到得到合适的视觉信息块。至此，整个 DOM 树的分块过程完毕。

2.3　页面信息提取算法

VIPS 算法是对页面所有信息进行分块，而信息提取只需要提取与主题有关的正文信息，本文讨论的是针对特定领域的 Deep Web 结果页面信息，这些信息大都集中在 Web 页面的特定位置（一般在正中间）。并且这些信息块具有相似的层次结构，大小和颜色，所以可以根据页面视觉特征和 DOM 树的层次结构找出一个基准视觉块，并逆序和顺序遍历整棵 DOM 树，找出页面所有相似视觉块，若存在形似的视觉块，再递归地对相似视觉块做以上同样的操作。直到找到所有想要抽取的信息。抽取流程图如图 5 所示。

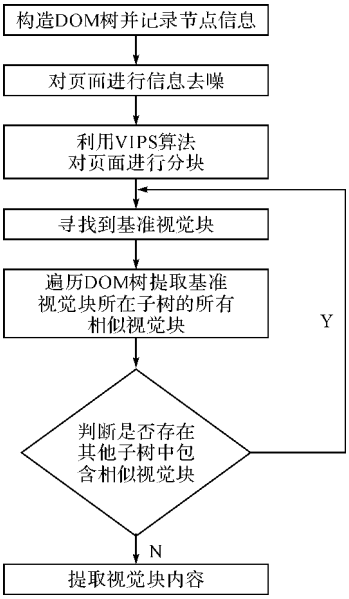

图 5　信息抽取流程图

下面具体说明正文信息提取算法。以图 6 为例定义网页左上角顶点为坐标原点，网页中心坐标为$(Center_X, Center_Y)$，定义每个视觉块的中心坐标为$(Block_Xi, Block_Yi)$，其中 $i=\{1,2,3,\cdots,n\}$，$n\in Z$。页面信息提取过程可以描述如下：

步骤一：提取基准视觉块。计算所有提取视觉块的中心坐标与页面中心坐标差值的绝对值$|Block_Xi-Center_X|$，$|Block_Yi-Center_Y|$，其中 $i=\{1,2,3,\cdots\}$，$n\in Z$。并把这两个值求平均$\dfrac{|Block_Xi-Center_X|+|Block_Yi-Center_Y|}{2}$，$i=\{1,2,3,\cdots\}$，$n\in Z$，取出所有数值中最小的所在视觉块作为基准视觉块。并记录该视觉块的大小、

颜色和其所在 *DOM* 树的层次数。从图 6 可以看出,该页面基准视觉块为 VB2_2_3。

步骤二:提取相似视觉块。通过观察可以发现,处于正文位置的视觉块具有相似的视觉特征,并且它们在 DOM 树中有相似的树层次结构和相同的父节点信息,所以可以以该基准视觉块即 VB2_2_3 所在树层次作为中心位置,遍历该视觉块所在层次的所有的兄弟节点,得到 VB2_2_1 和 VB2_2_2 并把它们和 VB2_2_3 作比较,它们具有相似的视觉大小和颜色,并处在相同层次的结构树中,所以把这三个视觉块其作为要提取的正文信息存储在目标池中。

步骤三:提取其他可能视觉块。尽管在 Web 页面中 DOM 结构树为基本的对象提供了一种层次结构,但是 DOM 结构树主要是用来显示而不是组织内容的,所以具有相似语义的视觉块可能存在不同的 DOM 树中,因此需要对 DOM 树进行进一步遍历以便找到所有可能的视觉块。这里采用文献[11]提到的逆序遍历方法。首先逆序遍历 DOM 树节点,找出目标池中所有视觉块 VB2_2_1、VB2_2_2 和 VB2_2_3 对应 DOM 树层次结构所在节点的公共父节点,即 VB2_2,再逆序向上找出该公共父节点的根节点 VB2,对此节点进行顺序遍历,得到 VB2_1 和 VB2_3 两个子节点,它们为 VB2_2 所在 DOM 树结构的所有兄弟节点。如果还有相似的正文目标视觉块存在,那么他们应该存在于 VB2_1 和 VB2_3 的子节点中,否则,说明不存在其他 DOM 树中包含相似的目标视觉块。在这里遍历到 VB2_1_1、VB2_1_2、VB2_3_1 和 VB2_3_2 四个孩子节点,把他们分别和基准视觉块 VB2_2_3 作比较,从图 6 中可以发现,这四个节点的大小和 VB2_2_3 相差较大,所以舍弃这些节点。

步骤四:根据步骤三的结果,如果提取到相似的视觉块信息,那么以提取到的视觉块作为新的基准视觉块递归作同样的操作,直到找到所有可能的视觉块。至此,正文信息视觉块提取结束。

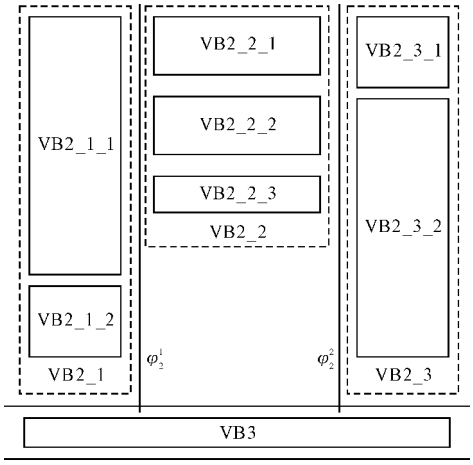

图 6 页面视觉分块图

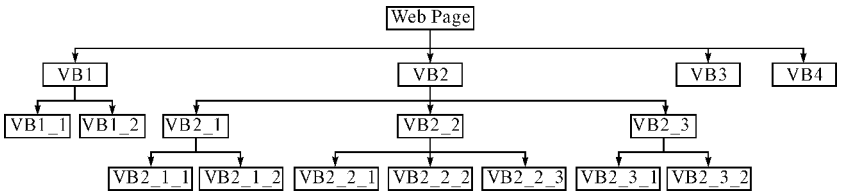

图 7 页面视觉块对应的 DOM 结构图

3 实验

本文实验分别实现传统的基于 DOM 树的网页信息抽取方法和本文提出的基于基准视觉块的逆序提取网页正文信息的抽取方法,并把这两种方法做比较,以体现本文提出的方法的优势。传统的基于 DOM 树的网页信息抽取方法主要是用一些开源工具如 NekoHTML、Jtidy 等把 Web 页面解析成一棵 DOM 树,然后深度遍历 DOM 树节点提取出页面正文信息。该方法实现简单,并具有一定代表性。本文通过对当当网、淘宝网等一些特定领域网站提交关键词查询获得大量的结果页面,把这些结果页面作为实验数据的来源。实验环境采用的是:主机 ASUS,处理器 Intel(R) Celeron(R) CPU 1.50GHz,内存 4GB,硬盘 250GB,操作系统为 Window 7。

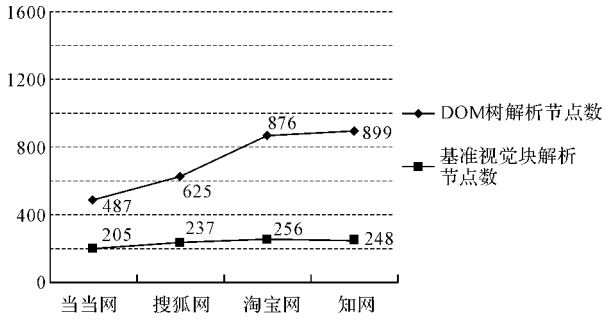

图 8 两种算法使用的节点数比较

图 8 显示的是使用两种不同的算法提取正文信息需要解析的 DOM 树节点数,从图中可以看出,不管是提取哪一类的网页,经过去噪处理的基于基准视觉块的逆序提取方法都只需解析几乎只包含正文信息的 DOM 节点。

另外，为了进一步验证该算法的性能和可行性，本文进行了信息抽取的准确率实验。分别对每类网站抽取 10 个页面，人工提取出关键正文信息，并估算总共包含的正文信息个数，与本文提出的抽取出正文信息块方法抽取出的正文信息以及正文信息数量进行人工比对，结果如表 1 所示。

表 1　Web 页面信息抽取性能

数据来源	Web 个数	总共包含正文信息个数	抽取到正文信息个数	准确抽取到正文信息个数	完整率	准确率
当当网	10	4800	4780	4780	0.99	1.00
搜狐网	10	1700	1300	700	0.76	0.53
淘宝网	10	7200	7180	7180	0.99	1.00
知网	10	3000	2890	2890	0.96	1.00

其中完整率＝(抽取到正文信息个数/总共包含的正文信息个数)×100%；准确率＝(准确抽取到正文信息个数/抽取到正文信息个数)×100%。从实验结果可以看出，本文提出的基于基准视觉块的 Web 页面抽取方法可以较准确并完整地抽取到所需要的正文信息，证明了该方法的可行性。由于一些网站，例如搜狐，并不属于纯粹的针对特定领域的网页，所以该类网页中包含的干扰视觉信息块较多，并且正文信息与基准信息不具有相似的视觉特征和位置特征，所以有部分正文信息块会被遗漏，导致抽取的完整率和准确率会有所下降。

4　结语

本文在基于 DOM 树结构的 Web 页面基础之上，利用人的视觉特征，首先根据一些启发式规则对原始页面去噪优化，然后利用 VIPS 算法把 Web 页面分成不同的视觉块，根据正文信息在页面的分布特征提取出基准视觉块，并根据基准视觉块的视觉特征逆序和顺序遍历整棵 DOM 树，递归提取出所有相似的视觉块。从实验结果来看，本文提出的方法在提取速度方面与传统方法相比有了一定的提高，并且有较高的准确率和完整率。但是本文的方法还有许多有待改进的地方。比如该方法比较适用于主题单一的网站，即整个网页只含单个文本区域的网站。如果页面结构较复杂，文本块较多，那么使用该方法有可能丢失一些有用的信息。另外，基准视觉块的大小也是一个关键，视觉块过大或过小都会影响实验的准确性和提取效率，下一步将对这方面做进一步研究，以达到更好的抽取效果。

参 考 文 献

[1] 吴茜，刘嘉勇. 基于 VIPS 算法和模糊字典匹配的网页提取技术研究[J]. 技术研究，2014(10)：49-53.
WU Qian, LIU Jiayong. Web Page extraction technology research Based on VIPS algorithm and fuzzy dictionary matching[J]. Netifo Security Technology Research, 2014(1): 49-53.

[2] 安增文，徐杰锋. 基于视觉特征的网页正文提取方法研究[J]. 微型机与应用，2010(3)：38-41.
AN Zengwen, XU Jiefeng. Web Page text extraction technology research Based on Visual feature[J]. Micro Computer and Application, 2010(3): 38-41.

[3] 郭迎春，刘一伟，陈召旭. Deep Web 数据抽取的分析与研究[J]. 南开大学学报(自然科学版)，2012，45(3)：9-14.
GUO Yingchun, LIU Yiwei, CHEN Zhaoxu. Analysis and Research on Deep Web Data Extraction[J]. Journal of Nankai University(Natural Science Edition), 2012, 45(3): 9-14.

[4] Wachirawut Thamviset, Sartra Wongthanavasu. Information extraction for deep web using repetitive subject pattern, World Wide Web 2014 DOI 10.1007/s11280-013-0248-y.

[5] 顾韵华，高原，等. 基于模板和领域本体的 Deep Web 信息抽取研究[J]. 计算机工程与设计，2014，35(1)：327-332.
GU Yunhua, GAO Yuan, et al. Deep Web information extraction research Based on template and domain ontology[J]. Computer Engineering and Design, 2014, 35(1): 327-332.

[6] 田建伟，李石君. 基于层次树模型的 Deep Web 数据提取方法. 计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP，2011，48(1)：94-102.
TIAN Jianwei, LI Shijun. Deep Web data extraction method based on hierarchical tree model[J]. Computer Research and Development ISSN 1000-1239/CN 11-1177/TP, 2011, 48(1): 94-102.

[7] 李朝，彭宏，叶苏南，等. 基于 DOM 树的可适应性 Web 信息抽取[J]. 计算机科学，2009，36(7)：202-210.
LI Chao PENG Hong, YE Sunan, et al. Adaptive Web information extraction based on DOM Tree[J]. Computer Science, 2009, 36(7): 202-210.

(下转第 1126 页)

7 结语

近年来,数据存储泄密事件层出不穷,为解决数据存储安全问题,本文提出了可控网络系统中的存储安全控制模型。模型以可控网络为理论指导,可控网络系统为载体,保证了数据存储的安全可控性;引入了加密控制机制,用户通过选取加密算法对存储数据进行加密,以保证数据的保密性;引入了数据鉴别控制机制,通过不同的数据鉴别方法实现对数据的签名,保证了数据的完整性和不可否认性;引入了存储备份控制机制,通过形成副本进行备份,保证了数据的持续可用性。提出了实现该模型的物理结构,形成网络存储安全动态防护体系。可控网络中各个部分在安全控制中心的协调和控制下,形成安全控制回路,从而实现对可控网络系统中的存储安全控制。在后续的工作中,将对可控网络理论及存储安全控制技术进行更深一步研究,以指导对该模型做进一步改进,使其更加安全与高效。

参 考 文 献

[1] 吴倩,范莉. 一种主动安全存储控制技术的实现与应用[J]. 中央民族大学学报,2010,19(1):57-62.
WU Qian, FAN Li. An Active Access Control Technology and Its Applications[J]. Journal of MUC,2010,19(1):57-62.

[2] 卢昱. 网络控制论概论[M]. 第一版. 北京:国防工业出版社,2005:15-17.
LU Yu. Cybernetics Introduction[M]. First Edition. Beijing: The National Defense Industry Press,2005:15-17.

[3] 游科友,谢立华. 网络控制系统的最新研究综述[J]. 自动化学报,2013,39(2):101-118.
YOU Keyou, XIE Lihua. The Latest Research Overview of Network Control System[J]. Acta Automatica Sinica,2013,39(2):101-118.

[4] 王双,卢昱,陈立云,等. 信息网络安全控制系统的研究与实现[J]. 军械工程学院学报,2015,27(1):35-38.
WANG Shuang, LU Yu, CHEN Liyun, et al. Research and Realization of Info-net Security Controlling System[J]. Journal of Ordnance Engineering College,2015,27(1):35-38.

[5] 饶兴. 基于 SSL 协议的安全代理的设计[D]. 武汉:武汉理工大学,2011.
RAO Xing. Design of Security Agent Based on SSL Protocol[D]. Wuhan: Wuhan University of Technology,2011.

[6] 卢昱. 信息网络安全控制[M]. 第一版. 北京:国防工业出版社,2011:26-27.
LU Yu. Info-Net Security Control[M]. First Edition. Beijing: The National Defense Industry Press,2011:26-27.

[7] 任恒. 基于密钥池的无线传感器网络密钥管理[D]. 长沙:湖南大学,2012.
REN Heng. Key Management Schemes for Wireless Sensor Networks Based on Key Pool[D]. Changsha: Hunan University,2012.

[8] 陈兴凯,卢昱,刘云龙,等. 信息网络安全可控技术研究[C]//2014 IEEE 工业应用前沿技术研究专题研讨会. 渥太华,加拿大,2014:919-922.
CHEN Xingkai, LU Yu, LIU Yunlong, et al. Research on info-net Security Controllable Technology[C]//2014 IEEE Workshop on Advanced Research and Technology Industry Applications . Ottawa, Canada, 2014:919-922.

[9] 牛德华,马建峰,马卓,等. 基于属性的安全增强云存储访问控制方案[J]. 通信学报,2013,34(Z1):276-284.
NIU Dehua, MA Jianfeng, MA Zhuo, et al. Enhanced Cloud Storage Access Control Scheme Based on Attribute[J]. Journal on Communications,2013,34(Z1):276-284.

[10] 陆丹. 基于 P2P 云存储备份系统设计及日志恢复实现[D]. 吉林:吉林大学,2012.
LU Dan. Design of Cloud Storage Backup System Based on P2P and Implementation of Log Recovery Module[D]. Jilin: Jilin University,2012.

(上接第 1111 页)

[8] 寇月,李冬. D-EEM:一种基于 DOM 树的 Deep Web 实体抽取机制[J]. 计算机研究与发展,2010,47(5):858-865.
KOU Yue, LI Dong. A Deep Web entity extraction mechanism based on DOM Tree[J]. Computer Research and Development,2010,47(5):858-865.

[9] 付涛. 基于 DOM 和显示属性的网页信息除噪技术研究[J]. 商丘师范学院学报,2010,26(9):90-93.
FU Tao. Web Information noise cancellation technology research Based on DOM and Display attributes[J]. Journal of Shangqiu Normal College,2010,26(9):90-93.

[10] Deng Cai, Shipeng Yu. Extracting Content_Structure for Web Pages based on Visual Representation Microsoft Research Asia.

[11] 张瑞雪,宋明秋. 逆序解析 DOM 树及网页正文信息提取[J]. 计算机科学,2011,38(4):213-215.
ZHANG Ruixue, SONG Mingqiu. Reverse parsing the DOM tree and informaiton extraction on the web page[J]. Computer Science,2011,38(4):213-215.