一种自适应网页结构化信息提取方法

淮晓永，韩晓东，高若辰，高焕新
（华北计算机系统工程研究所，北京 100083）

摘　要： 面向互联网信息采集挖掘应用，针对传统的网站信息整页采集方式存在采集信息混杂、无法直接使用，而人工结构化采集方式成本高、工作效率低的问题，研究提出了一种自适应网页结构化信息提取方法，实现了网页分类算法、基于子树的标题项、内容项的结构化信息提取算法。基于典型网站网页分类标注数据集进行分类模型的学习建模，可以自适应不同网站的差异，对网页进行分类，按照网页分类分别提取出网页中的列表项结构化信息、内容项结构化信息。该技术对提高网站信息结构化采集处理的自动化水平及处理效率具有重要作用。

关键词： 信息提取；结构化信息；分类模型；自适应
中图分类号： TN919.5；TP391.1　　**文献标识码：** A　　DOI：10.16157/j.issn.0258-7998.200160

中文引用格式： 淮晓永，韩晓东，高若辰，等. 一种自适应网页结构化信息提取方法[J].电子技术应用，2020，46(12)：97-102.
英文引用格式： Huai Xiaoyong，Han Xiaodong，Gao Ruochen，et al. An adaptive method for extracting structured information from web pages[J]. Application of Electronic Technique，2020，46(12)：97-102.

An adaptive method for extracting structured information from web pages

Huai Xiaoyong，Han Xiaodong，Gao Ruochen，Gao Huanxin
（National Computer System Engineering Research Institute of China，Beijing 100083，China）

Abstract： In order to meet the needs of Internet information collection and mining，aiming at the problems of traditional web site information collection methods，such as mixed collection information，unable to be used directly，and the high cost and low efficiency of manual structured collection method，this paper proposes an adaptive method for extracting structured information from web pages. We implement web page classification algorithm，subtree based title item and content item structured information extraction algorithm. Based on the classification annotated dataset of typical website pages，the classification model can adapt to the differences of various web sites，classify the web pages，and extract the list structured information and content structured information in the web pages according to the web page classification. This technology plays an important role in improving the automation level and processing efficiency of website structured information collection and processing.

Key words： information extraction；structured information；classification model；adaptive

0 引言

在互联网大数据时代，互联网信息呈现爆炸式增长，其中蕴藏着很多有价值的重要信息需要处理与利用。通过智能化的大数据信息挖掘处理，可以从中分析把握技术发展的方向态势，迅速发现高价值的科技信息。

从关注的 Internet 网站源自动采集收集新发布的信息，并提取出其中的结构化信息，是建立互联网大数据系统的基础。通过网络爬虫系统可以从各类网站爬取大量的网页数据，但传统的网站信息整页采集方式信息混杂，无法直接进行大数据挖掘处理，而人工从网页中提取结构化的文本信息又存在成本高、工作效率低的问题。如何通过自动化的网页数据结构化信息采集技术实现自动从网页中提取结构化的信息，是进行互联网大数据挖掘处理的关键预处理技术。

本文研究针对传统的网站信息整页采集方式存在采集信息混杂、无法直接使用，而人工结构化采集方式成本高、工作效率低的问题，研究实现了一种基于 DOM 树的网页结构化信息提取方法（DOM based Web-page Structured Information Extraction，DWSIE），实现了一个网页结构化信息提取服务工具包，该工具包极大地提高了网站结构化信息采集处理的自动化水平和处理效率。

1 网页信息提取技术概述

网页结构化信息提取是指从网页中提取出结构化的文本数据信息。对于列表类、导航类网页提取的数据包括：标题、链接地址、发布日期、频道栏目名称等；对于内容类网页提取的数据包括：标题、发布日期、作者、正文、频道栏目名称、首图等。网页信息提取方法主要包括：基于统计的技术、基于视觉特性的技术、基于 DOM

树结构的技术和基于模板的技术等。

基于统计的方法通过统计网页各个区域中标签包含的信息量或链接文本与普通文本的比值等方法来获取网页的主题信息。文献[1]设计的 Crunch 系统利用区域中的 link/text(链接文本/普通文本)比值与既定经验阈值的大小关系来确定网页的正文区域;文献[2]提出了针对<table>标签布局的网页处理方法。

基于视觉特征抽取信息是通过扫描网页中<table>、<p>、<hr>、<ul>等分隔符,把页面分成若干视觉信息块,进而通过经验规则进行区域划分与信息提取。文献[3]在识别建立分割区域时充分利用了字体大小、背景颜色、空白区域等视觉特征,并总结出一些经验规则:(1)类似<hr>等标签通常用于分割不同的主题,因此如果一个区域内包含<hr>标签,那么倾向于分割这个区域;(2)如果一个区域的背景色与其内部子区域的背景色不同,则分割这个区域;(3)如果一个区域内大部分节点都是文本类型,则不再分割这个区域。由于视觉特征的复杂性和网页设计的多样性且存在许多不符合规范的页面,这种基于视觉的信息抽取技术在实际应用中会存在很多问题。

基于 DOM 树结构的技术是通过构建网页 DOM 树,进而基于对树的操作处理算法来提取网页数据信息。根据 HTML DOM 标准,HTML 网页文档中的所有内容都是节点:整个文档是一个文档节点,每个 HTML 元素是元素节点,HTML 元素内的文本是文本节点,每个 HTML 属性是属性节点,注释是注释节点。HTML 文档中的所有节点组成了一个文档节点树。在 Web 信息抽取中可以在网页树结构基础上通过一些针对树的操作方法总结归纳出待抽取部分的特征。基于 DOM 树结构的技术其操作过程相对于基于视觉的方法更加易于实现。在基于 DOM 树结构的抽取技术领域有许多成型的系统和经典算法。文献[4]提出了一种通过判断页面中 data-rich(数据密集)区域达到抽取页面主题信息目的的 DSE(Data-rich Section Extraction)算法,并定义页面内包含了主题信息的区域为 data-rich 区域。这种算法主要针对通过查询数据库动态生成的网页,认为同一网站的页面通常具有相似结构的信息组织方式。RoadRunner[5]系统中的信息抽取算法通过比较两棵同源网页的 DOM 树之间的匹配与不匹配部分来得到一个网页主题信息抽取程序。

基于模板的信息提取方法是基于很多网站网页生成与信息发布采用的“模板+数据”的方式,通过识别构建不同网站模板来提取相应的结构化数据。文献[6]根据网站 URL 树中在同一个目录节点下存在大量由同一模板生成的网页,由模板生成的网页布局基本一致等特征,作者定义了抽取模板 T={urlfix, tags prefix, tags suffix, label point}及模板生成方法,其中:urlfix 是指从根节点到目录节点的路径,tags prefix 是指从根节点到待抽取信息节点(不含待抽取信息节点)的标签序列,tags suffix 是指从待抽取节点(不含待抽取节点)到树的末尾的所有标签序列,label point 是网页待抽取信息块在网页 DOM 树中的路径。如果两个网页具有相同的 tags prefix 和 tags suffix,则认为这两个网页是由同一个模板生成的。在模板生成阶段,通过对一定数量的网页的对比操作总结出模板 T,将 T 存放到这些页面的目录节点中,并认为这个目录节点下的所有网页都可以按照这个模板来处理。文献[7]引入了 URL 模板匹配的概念,根据待抽取页面的 URL 与 URL 模板匹配库进行模板匹配,识别该页面是否可解析及确定该页面所用的解析模板。文献[8-10]也提出了类似的能够产生信息抽取模板的方法,只是对模板的定义形式略有不同。该方法对非模板型架构的网站效果较差。

2 基于 DOM 树的网页结构化信息提取方法——DWSIE

基于 HTML DOM 标准,可以通过为网页 DOM 处理方法建立节点间的父子关联关系,构建网页的 DOM 树。一个 HTML 网页可以构建生成如图 1 所示的 DOM 树图。

DWSIE 方法首先扫描网页构建网页的 DOM 树,然后通过节点分类模型对树中的节点进行分类识别,识别出树中的标题项节点、内容节点,进而识别出列表子树、内容子树,最后从网页 DOM 树图的列表项子图、内容子图中提取相应节点的结构化信息。

2.1 网页分类方法

网页分类首先对网页 DOM 树中每个节点进行分类:从节点的分类特征属性值,根据节点分类模型计算得到每个网页节点的类别;然后根据列表类网页、内容类网页的特征规则,对网页进行分类。

这里选取节点的分类特征属性集为:[IL, depth, SC, mean_IL, var_IL]。

其中:

(1)IL,即节点信息长度,指节点信息内容的字数(中文指字数,英文指单词数);

(2)depth,即节点所在层数;

(3)SC,即节点的子节点数。

还有子节点的统计类特征:

(1)mean_IL,表示节点其下属子节点信息长度的均值;

(2)var_IL,表示节点下属子节点信息长度的方差。

为了使分类器能够自适应不同网站网页的特征,这里采用神经网络分类器,利用人工标注的数据集对分类器进行训练学习。针对互联网大数据系统的业务领域,选择了 20 个典型网站,通过人工标注了 2 000 个网页,建立了网页分类标注数据库,并将数据库分为训练数据集和测试数据集。

根据节点分类目标,设计采用如图 2 所示的神经网络节点分类模型,输入为各信息节点的特征属项集:

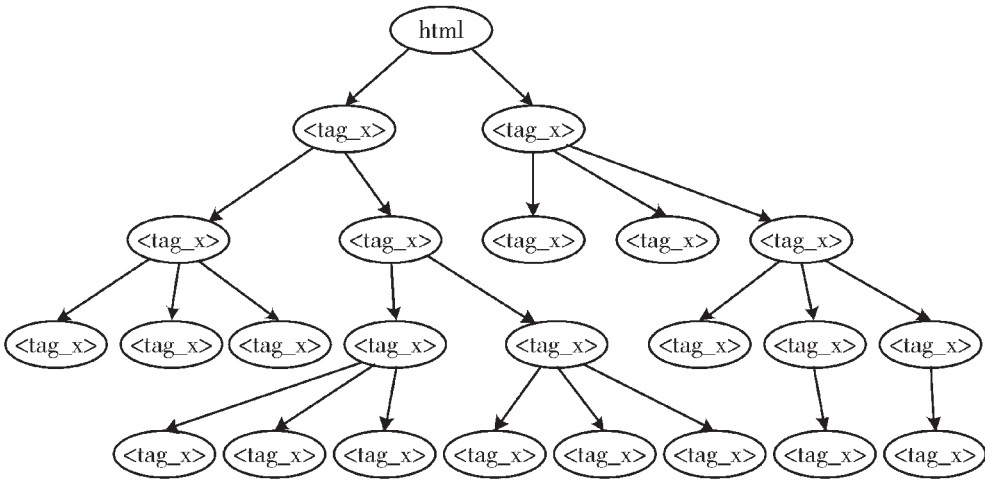


图 1 网页 DOM 树示意图(分类前)

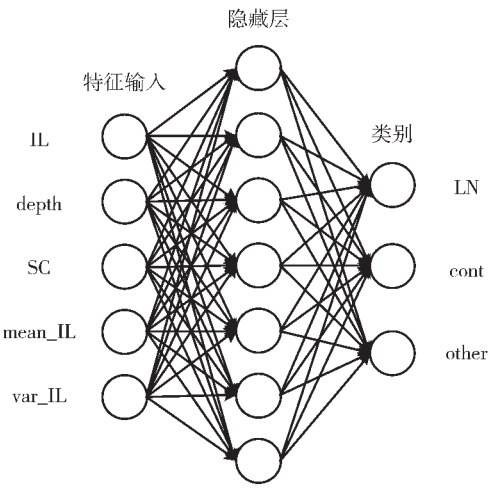


图 2 节点分类模型

$X=[\mathrm{IL},\mathrm{depth},\mathrm{SC},\mathrm{mean\_IL},\mathrm{var\_IL}]_i$，输出为节点类别 $Y=[\mathrm{LN},\mathrm{cont},\mathrm{other}]$。其中，LN 为列表类节点，cont 为内容类节点，other 为其他类节点。

节点分类模型的训练学习算法如下：

输入：标注数据集

输出：节点分类模型

步骤：

{

(1) 从标注数据集的训练数据集中随机采样 $N(N=300)$ 个网页，获取各网页的标注节点的特征属性标注数据：

X=[IL,depth,SC,mean_IL,var_IL]i

Y 为节点类别

(2) 对输入 X 归一化处理：

min_max_scaler=preprocessing.MinMaxScaler()

X_minMax=min_max_scaler.fit_transform(X)

(3) 训练神经网络：cnn.fit(X_minMax, Y)

(4) 保存训练后的模型参数；

}

建立了节点分类模型后，根据节点分类模型，可以通过节点分类算法进行网页 DOM 树的节点分类计算，识别出其中的 LN、cont 类节点。

DOM 树节点分类算法如下：

输入：网页 DOM 树

输出：含节点分类信息的网页 DOM 树

步骤：

{

(1) 对于网页 DOM 树中的每个节点，计算其特征属性向量：

Xi=[IL, depth, SC, mean_IL, var_IL]

(2) 进行网页特征属性数据的归一化处理：

X_minMax=in_max_scaler.fit_transform(X)

(3) 对于网页 DOM 树中的每个节点，通过神经网络分类器预测其类别：

c=cnn.predict(X_minMax[k])

}

列表类网页的 DOM 树一般含有若干个列表子树 LT(示意图见图 3(a))；内容类网页的 DOM 树一般含有一个内容子树 CT(示意图见图 3(b))。

$\mathrm{LT}=\{(\mathrm{LN})_i\}=\{(T,\mathrm{href},\mathrm{date},\mathrm{au})_i\}$

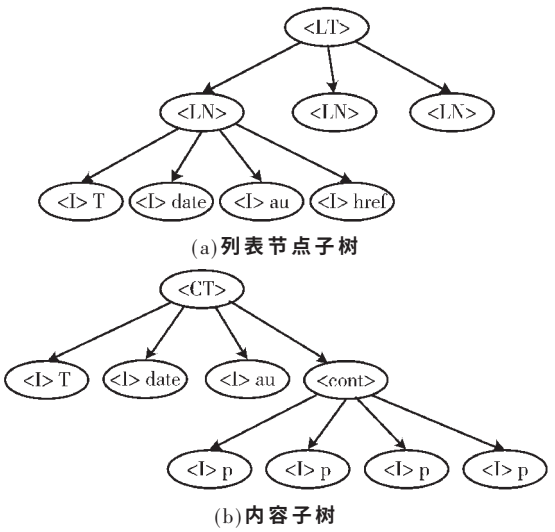
$\mathrm{CT}=\{T,\mathrm{date},\mathrm{au},\mathrm{cont}\}$

其中，LN 表示列表节点，T 表示标题，href 表示标题的内容链接地址，au 表示作者，date 表示发布时间，cont 表示网页正文内容。

利用节点分类模型对树中的节点进行分类，识别出树中的标题类节点 LN、内容节点 cont，进而根据经验规则识别出列表子树 LT、内容子树 CT，节点分类后的 DOM 树如图 4 所示。

对网页 DOM 树节点分类后，可以通过网页分类算法进行网页的分类，将网页分类为列表类网页 LP 或内容类网页 CP。网页分类算法如下：

输入：节点分类后的 DOM 树



(a)列表节点子树

(b)内容子树

图 3 信息子树图

输出：网页类别

步骤：

```
{
    if    DOM 树含有内容项节点 then
          网页类别为内容类网页：PT=CP
    elseif    DOM 树含有列表子树节点 then
          网页类别为列表类网页：PT=LP
    else
          网页类别未知：PT=Unknown
}
```

2.2 网页信息提取方法

网页分类后即可根据不同网页提取相应的结构化网页信息。网页结构化信息提取算法如下：

输入：网页分类后的网页 DOM 树

输出：该网页的结构化信息

步骤：

```
    {
  if 网页类别为列表类网页 LP then
   {
          对每个列表子树 LTi:
          { 对 LTi 的每个列表节点 LNj：
             {
                  从节点信息中提取超链接 href；
                  从其子节点中提取标题信息 T；
                  从其子节点中提取日期信息 date；
                  从其子节点中提取作者信息 au；
             }
          }
   }elseif 网页类别为内容类网页 CP then
   {
                  从内容节点提取内容信息 cont；
                  从内容子树中提取标题信息 T；
                  从内容子树中提取日期信息 date；
                  从内容子树中提取作者信息 au；
   }
}
```

2.3 网页信息提取处理流程

网页信息提取的处理流程如图 5 所示。

网页信息提取结果采用 JSON 格式描述，示例如下：

```
{
    "url": a_url,                          //网页的 URL 地址
    "page":"页面源码",
    "page_type": a_PT,                      // TP,CP,Unknown
    "titles": [
        {
            "title": "标题 1",
            "href": "链接地址 1",
            "au": "作者姓名",
            "date": "2018-9-1"
        },
        {
            "title": "标题 2",
            "href": "链接地址 2",
            "au": "作者姓名"
        },
```

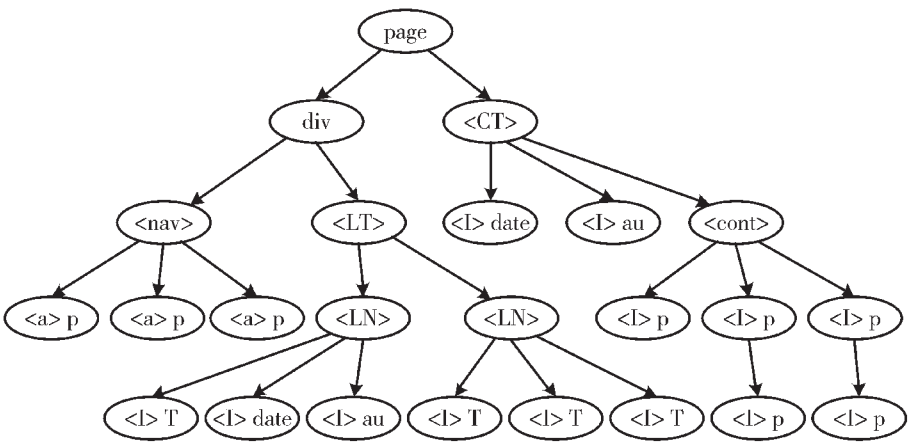


图 4 网页 DOM 树示意图(分类后)

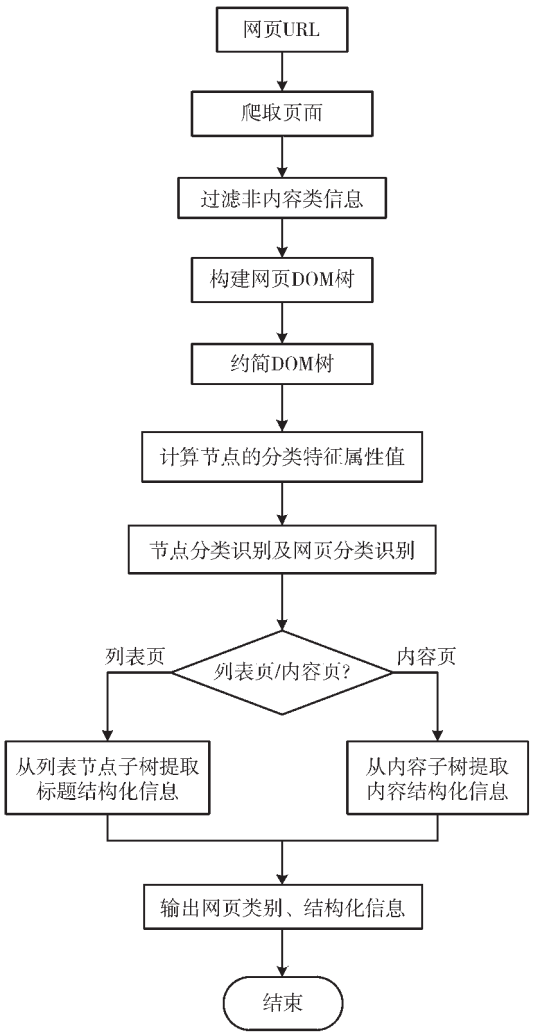


图 5 网页结构化信息提取处理流程

```
        ...
    ],
    "doc": {
        "title": "内容页标题",
        "column": "栏目",
        "au": "发布者",
        "date": "发布日期",
        "main_img": "首图",
        "cont": "内容",
        "keywords": "关键词 1,关键词 2,..."
    }
}
```

2.4 网页信息提取服务

本文基于上述的 DWSIE 网页结构化信息提取方法，研制了一个网页结构化提取应用服务，实现了 JSON-RPC 协议、SOAP 协议的网页分类 detect_pagetye、网页结构化信息提取 extract_pageinfo 服务接口，如图 6 所示。该服务已应用于互联网信息挖掘系统中作为其网页数据采集的关键预处理模块。

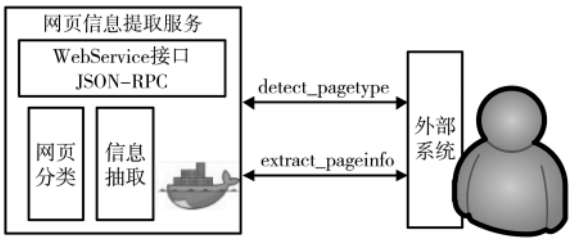


图 6 DWSIE 网页信息提取服务

3 算法性能测试分析

测试环境：1 台主流中端性能服务器；硬件基本配置为：2xE5 CPU、64 GB 内存、4 TB 硬盘、1 000 Mb/s 网卡；软件基本配置：Ubuntu 16.04，Docker 17.12，Python3.6；测试数据为标注数据库中的测试数据集。

(1)测试 1：方法准确性测试

从测试数据集中随机抽样 200 条数据进行测试，进行 10 次测试，分别计算记录每次测试的列表类网页分类 F1 值 F1(LP)、内容类网页分类 F1 值 F1(CP)、列表信息提取 F1(T)、内容信息提取值 F1(C)。测试结果如图 7 所示。

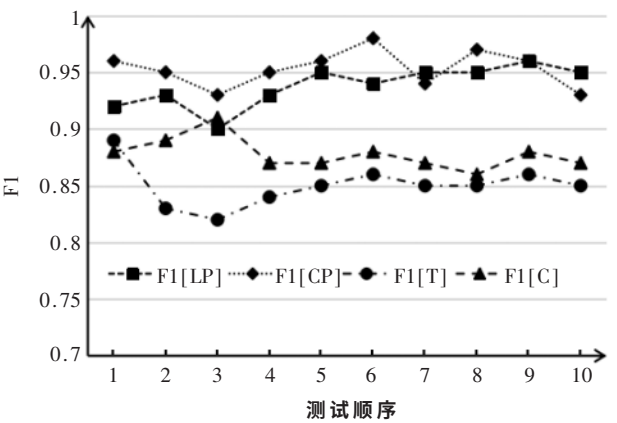


图 7 随机 10 次抽样测试 F1 值

测试数据表明 DBWSE 算法的准确度指标为：网页分类 F1>0.9，结构化信息提取 F1>0.8，能够满足工程应用的技术指标要求。

(2)测试 2：对不同网站的自适应性测试

选择 8 个网站，分别从各网站随机抽取 50 条数据计算各网站数据的分类和信息提取的 F1 值，测试数据结果如图 8 所示。

测试数据表明针对不同的网站，DBWSE 算法能够适应不同的网站进行网页分类与结构化信息提取，F1 值能够满足技术指标要求，网页分类 F1>0.9，结构化信息提取 F1>0.8。

4 结论

本文面向网站网页数据的结构化信息采集技术需求，研究提出了基于网页 DOM 树分析的自适应网页分类与结构化信息提取方法 DWSIE，主要技术特点如下：

(1)自适应网页分类模型通过构建典型网站的网页

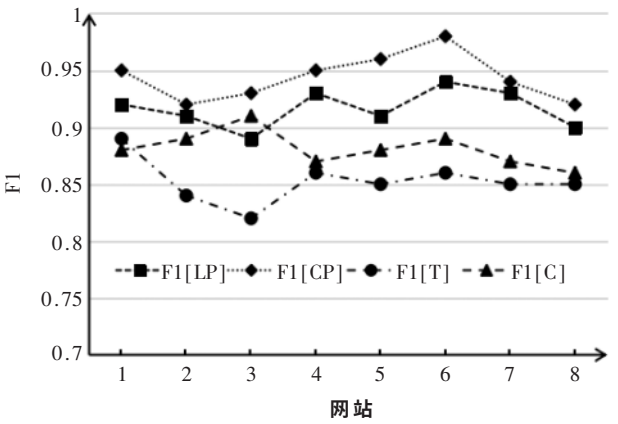


图 8 不同网站测试 F1 值

分类标注数据库，进行分类模型的训练学习，使得分类模型能够具有较高的普遍性，能够自适应不同的网站样式的差异；

(2)基于子图上下文的结构化信息提取方法，通过列表子树、内容子树来提取相应的网页标题项、内容项结构化信息，并通过日期提取、实体提取进行信息验证，从而保证了较高的结构化信息提取准确率。

参考文献

[1] GUPTA S, KAISER G. DOM-based content extraction of HTML documents[C]. Proceedings of the 12th World Wide Web Conference. New York: ACM Presss, 2003: 207-214.

[2] 孙承杰，关毅. 基于统计的网页正文信息抽取方法的研究[J]. 中文信息学报，2004，18(5)：17-22.

[3] Cai Deng, Yu Shipeng, Wen Jirong, et al. VIPS: a vision based age segmentation algorithm[R/OL]. (2003-11)[2020-03-03]. http://research. microsoft.com/apps/pubs/defanlt.aspx?id=70027.pdf.

[4] Wang Jiying, LOCHOVSKY F H. Data-rich section extraction from HTML pages[C]. Proceedings of the 3rd International Conference on Web Informations Systems Engineering. Washington DC: IEEE Computer Society, 2002: 2313-2322.

[5] CRESCENZI V, MECCA G. RoadRunner: towards automatic data extraction from large Web sites[C]. Proceedings of the 27th VLDB Conference. San Francisco: Morgan Kaufmann Publishers, 2001: 109-118.

[6] 欧建文，董首斌，蔡斌. 模板化网页主题信息提取方法[J]. 清华大学学报(自然科学版)，2005，45(9)：1743-1747.

[7] 张彦超，刘方，李勇，等. 基于自动生成模板的 Web 信息提取技术[J]. 北京交通大学学报(自然科学版)，2009，33(5)：40-45.

[8] 郑长松，傅彦，余莉. 基于模板的 Web 信息自动抽取方法[J]. 计算机应用研究，2009，26(2)：570-582.

[9] 陈治昂，周知予，李大学. 一种基于模板的快速网页文本自动抽取算法[J]. 计算机应用研究，2009，26(7)：2646-2649.

[10] 杨少华，林海略，韩燕波. 针对模板生成网页的一种数据自动抽取方法[J]. 软件学报，2008，19(2)：209-223.

(收稿日期：2020-03-03)

作者简介：

淮晓永(1973-)，男，博士，高级工程师，主要研究方向：智能软件工程、云计算。

韩晓东(1994-)，男，硕士，工程师，主要研究方向：计算机软件、人工智能。

高若辰(1996-)，女，硕士研究生，主要研究方向：智能信息处理。

(上接第 96 页)

进一步提高和完善人体骨架姿势的正确性。

参考文献

[1] SCHWARZ L A, MKHITARYAN A, MATEUS D, et al. Human skeleton tracking from depth data using geodesic distances and optical flow[J]. Image and Vision Computing, 2012, 30(3): 217-226.

[2] Zheng Xiao, Fu Mengyin, Yang Yi, et al. 3D Humanpostures recognition using Kinect[C]. 2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHM-SC), 2012: 344-347.

[3] HU R Z L, HARTFIEL A, TUNG J, et al. 3D pose tracking of walker users' lower limb with a structured-light camera on a moving platform[C]. Computer Vision and Pattern Recognition Workshops, 2011: 29-36.

[4] 钱堃，马旭东，戴先中. 基于抽象隐马尔科夫模型的运动行为识别方法[J]. 模式识别与人工智能，2009(3)：433-439.

[5] 孙冰岩，曹琦，王星. 基于 Hausdorff 距离的目标跟踪方法研究[J]. 现代防御技术，2010(5)：127-130.

[6] 李昕迪. 基于 Kinect 的人体姿势识别方法在舞蹈训练中的应用[D]. 哈尔滨：黑龙江大学，2015.

[7] 胡小华，李向攀，祁洋阳. 可穿戴式人体姿态检测系统设计[J]. 电子技术应用，2017，43(9)：13-16.

[8] 王晓琳. 基于计算机视觉的手势识别人机交互技术[D]. 杭州：浙江工业大学，2010.

[9] 张璘，杨丰墒. 基于深度学习的图像分类搜索系统[J]. 电子技术应用，2019，45(12)：51-55.

[10] 冷晶晶. 基于 Kinect 骨架信息的人体动作识别[J]. 数字技术与应用，2014(9)：80.

[11] 战荫伟，于芝枝，蔡俊. 基于 Kinect 角度测量的姿势识别算法[J]. 传感器与微系统，2014，33(7)：129-132.

(收稿日期：2020-02-25)

作者简介：

杨海清(1971-)，男，副教授，主要研究方向：无线传感和网络控制技术及应用。

钱涛(1994-)，男，硕士研究生，主要研究方向：无线传感技术。

版权声明