Computer Science and Application 计算机科学与应用, 2019, 9(2), 181-187
Published Online February 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.92022

Hans汉斯

Research on WEB Information Extraction Based on DOM Tree Statistics Keyword Path

Jianshi Zhao, Junqing Liang, Xiaolin Lv, Xuebo Huang, Yue Leng, Zaijun Zhang

School of Information and Control Engineering, Qingdao University of Technology, Qingdao Shandong
Email: hnzhaojianshi@163.com

Received: Jan. 10th, 2019; accepted: Jan. 22nd, 2019; published: Jan. 29th, 2019

Abstract

Extracting WEB information according to users' requirements plays an important role in WEB data mining. Aiming at extracting the list of products on the company website, this paper proposes a method based on DOM tree statistics keyword path to determine the critical path and extract content according to the critical path. It is mainly divided into the acquisition of key phrase candidates, the acquisition of the product page of the company's official website, the establishment of the DOM tree of the web page, and the determination of the key path and extraction. This paper proposes an information extraction method to solve the problem of large difference in web page structure on different companies' official websites. According to this method, the required product information is extracted from the official website of the auto parts manufacturer.

Keywords

DOM Tree, Critical Path, Key Phrase, Information Extraction

基于DOM树统计关键词路径的WEB信息提取研究

赵建视，梁俊卿，吕笑琳，黄学波，冷　悦，张在军

青岛理工大学信息与控制工程学院，山东　青岛
Email: hnzhaojianshi@163.com

收稿日期：2019年1月10日；录用日期：2019年1月22日；发布日期：2019年1月29日

摘　要

根据用户的需求提取WEB信息在WEB数据挖掘领域中起着重要的作用。本文以提取公司官网上的产品列

文章引用: 赵建视, 梁俊卿, 吕笑琳, 黄学波, 冷悦, 张在军. 基于 DOM 树统计关键词路径的 WEB 信息提取研究[J]. 计算机科学与应用, 2019, 9(2): 181-187. DOI: 10.12677/csa.2019.92022

表为目标，提出了一种基于DOM树统计关键词路径的方法来确定关键路径，根据关键路径来确定提取内容。过程主要分为候选关键词组的获取、企业公司官网产品页面的获取、web页面的DOM树建立以及确定关键路径及提取。本文提出了一种解决不同公司官网上网页结构差距较大问题的信息提取的方法。并且根据该方法实现了在汽车零部件生产厂商官网上提取需要的产品信息。

关键词

DOM树，关键路径，关键词组，信息提取

1. 引言

Web 信息提取是将 Web 作为信息源的一类信息抽取，就是从无结构或半结构的信息抽取中，识别出用户感兴趣的数据。随着互联网技术的发展，Web 成为全球企业与机构进行信息发布与应用部署的主要平台之一。大量 Web 网站和 Web 应用的出现使得 Web 的数据量急剧增长。Web 上的海量数据蕴含很多有价值的信息。为了获得并分析利用这些有价值的信息，通常需要从 Web 上获取精确有用的结构化数据，然后对这些结构化数据进一步分析处理[1]。目前市场上已有很多信息抽取系统，比如，八爪鱼采集器就是其中之一。八爪鱼采集器主要是通过建立模版进行采集，首先选择模式和相应的网站模版，接着预览模版的采集字段，最后设置参数，运行完成数据采集。主要原理是通过正则表达式与 Xpath 原理来获取网页数据。该采集器主要的缺点是需要事先知道网页结构并且根据网页结构写出规则来确定 Xpath。对于未知网页结构的网页采集效果较差。

目前，很多学者对 Web 信息提取做了一些研究并提供了设计方案。王一州提出一种基于网页聚类的正文信息提取方法，提取信息时主要根据网页的结构特征对网页进行聚类，利用相似网页集合的正文位置生产特征[2]。但该方法只适合提取来自同一网站的网页。孙景春在基于改进网页视觉特征分块算法 VIPS 基础上，通过归纳 Web 网页视觉特征及视觉块特征信息，提出了基于视觉块识别的网页元数据的 Web 页面信息提取算法[3]。主要用于提取主题型网页及 BBS 型网页的信息。马金娜提出了一种基于 DOM 树节点重要度的 Web 信息提取方法。首先将 Web 页面表示为 DOM 树，然后对 DOM 树结点重要度进行定义，最后基于 DOM 树节点重要度进行 Web 信息内容提取[4]。高峰等人结合有监督广度优先搜索策略提出了一种通用垂直的 WEB 信息提取方法。首先自动识别目标主题和目录页面 URL，并利用 URL 聚类生成 URL 正则表达式过滤器，然后利用正则表达式过滤器和解析路径模板以及有监督的广度优先与网页赋权搜索策略进行 WEB 信息提取[5]。该方法过度依赖 URL 格式，当网站中不同专题的目标页面 URL 格式相近时，提取效果不佳。赵朗从深度学习的角度出发，在循环神经网络算法的基础上，提出了基于双层神经网络的信息提取算法[6]。该方法需要用户提交关键词，用户参与度较高，对用户提出的查询进行了统一规定，缺乏个性化服务。王健提出了基于 Hadoop 的 web 页面正文抽取的方法，抽取流程主要为页面预处理、网页分割、正文块识别以及正文语句合并[7]。该方法在对列表型页面进行抽取时显示出很强的局限性。

从以上研究中可以发现，目前学者们的研究和市场上的抽取方法都是针对某一特定类型的网站，大部分都是针对电商平台的网页信息提取。而对于网页结构及设计模式差距较大的企业官网没有提出比较好的解决方案。本文在基于 DOM 模型结合关键词路径的基础上提出统计关键词路径出现的次数，确定

关键路径，最后确定提取内容的方法。解决了网页结构及设计模块较大的一些企业公司官网的信息提取的问题，实现了从企业公司官网提取产品信息。

2. 基于 DOM 树的统计关键词路径方法

互联网上的同一网站一般采取相同的模版产生同类主题，而不同的网站主题的组织形式也有很多相似之处，例如电子商务类型网站。与提取电子商务网站不同，即使在某一特定领域，不同企业公司官网的主题的组织形式往往有很大的差距。为了解决企业公司官网组织形式的差异性，本文提出了一种基于 DOM 树的统计关键词路径的方法。基于 DOM 树的统计关键词路径的信息提取主要由以下几个模块组成，分别为候选关键词组的获取、企业官网产品页面的获取、网页解析成 DOM 树、确定关键路径并提取信息。如图 1 所示。

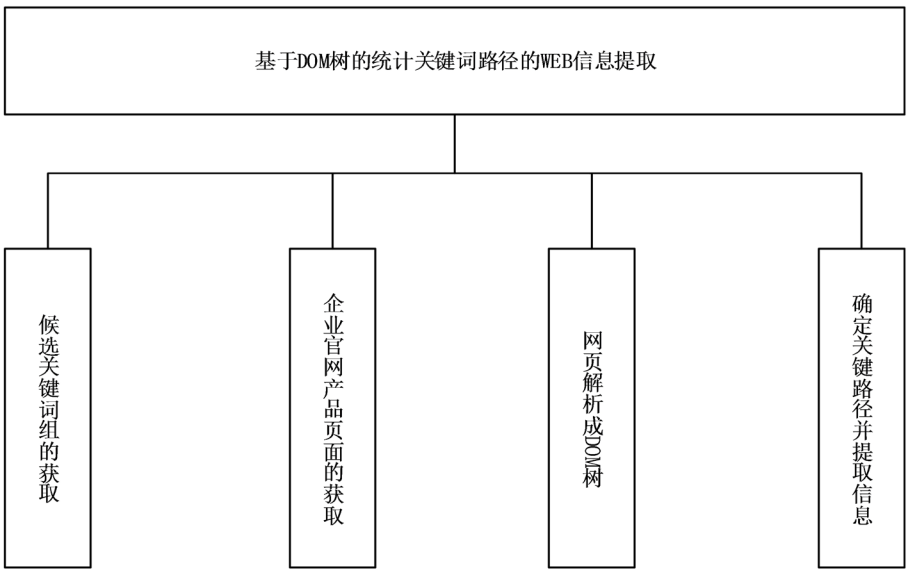

Figure 1. Information extraction of statistical keyword path based on DOM Tree
图 1. 基于 DOM 树的统计关键词路径的信息提取

候选关键词组的获取：与电商网站的抽取方法不同，如：基于标签的网页信息抽取方法[8]。同一个企业官网上不同网页的主题信息存在差异，不能通过分析相同的网页特征，产生关键词组。本文的关键词组的获取通过第三方网站来获取关键词组，并且与企业的代码一一对应。

企业公司官网产品页面的获取：在以往的信息提取中，网页的导航一般都是被视为噪音处理的。为了从企业官网首页获取产品信息的页面，导航栏成为了必须要提取的一部分。目前关于网页导航提取的研究相对较少。本文中使用关键词与正则表达式提取网页导航中的产品网页的链接。

网页解析成 DOM 树：一个页面可以用 DOM 树表示，DOM 即文档对象模型，定义了 HTML 文档对象和 XML 文档的逻辑结构，给出了一种访问和处理 HTML 文档的 XML 文档的方法，可以根据 HTML 文档和 XML 文档结构形成一棵对象节点树，称为 DOM 树[9]。本文借鉴基于 Jsoup 的通用页面采集系统的设计与实现[10]中的方法利用 Jsoup 解析 DOM 树。

确定关键路径并提取信息：从选定的关键词组中，使用产品网页中的文本信息进行校正与筛选，主要运用词语相似度比较算法，进行筛选关键词组，确定最终的关键词组。从解析的 DOM 树中提取关键词路径，提取算法主要借鉴逆序解析 DOM 树及网页正文抽取[11]。获取每个关键词的路径，统计关键词路径。统计路径次数，路径出现最多的确定为关键路径。根据关键路径进行信息提取。

3. 相关方法介绍

3.1. 候选关键词组的获取

与以往的研究不同，企业公司官网上的每个网页的主题信息都不同，无法使用聚类的方法来确定关键词，而我们仅仅只需要提取产品页面中的产品信息。而不是需要提取每个网页的正文内容。所以候选关键词组的确定就要从第三方网站获取。本文主要从同花顺网站上获取候选关键词。同花顺官网上的产品信息如图 2 所示。

Figure 2. Company information introduction of Straight Flush Website
图 2. 同花顺网站公司信息介绍

由图 2 可以分析出同花顺网站上企业介绍部分中会涉及到产品信息，根据这一特点我们可以根据我们的需求爬取产品名称以确定关键词组。首先我们需要爬取这些公司的代码。由于同花顺网站上的公司介绍的链接除了上市企业股票代码不同，而其他地方均相同，所以根据企业股票代码列表爬取同花顺上每个公司的介绍页面。然后根据 Jsoup 的选择器 select 选择产品名称所在的位置。以爬取汽车零部件企业官网信息为例，具体流程如图 3 所示：

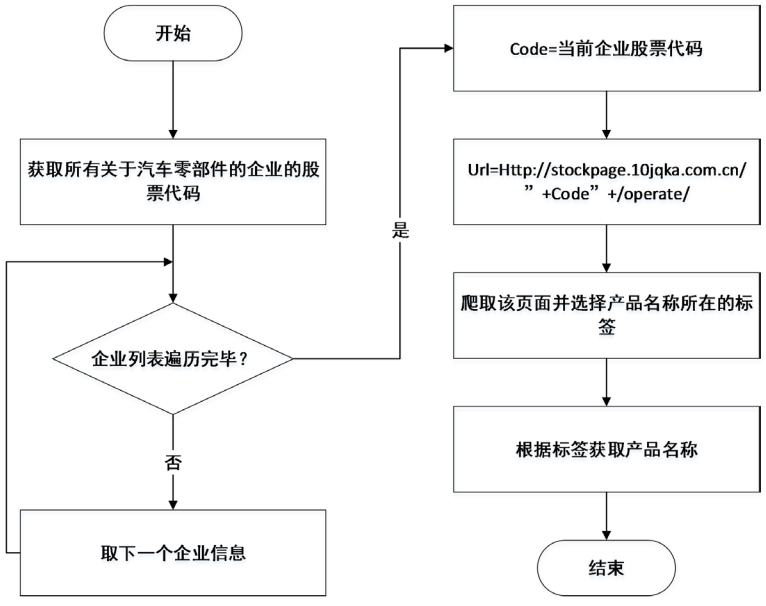

Figure 3. Crawl the product information on the company information page of Straight Flush Website
图 3. 爬取同花顺企业信息页面中的产品信息

3.2. 企业官网产品页面的获取

本文主要从同花顺网站上获取公司的官方网址，但是公司官网首页往往包含许多信息。我们根据需求，需要得到产品链接的网址。产品链接一般出现在网页的导航部分。目前研究都是提取网页的信息，对导航的信息提取的研究少之又少。本文就汽车零部件行业的企业官网进行研究，使用基于关键字与正则表达式的方法进行导航确定及产品链接的获取。流程如图 4 所示。

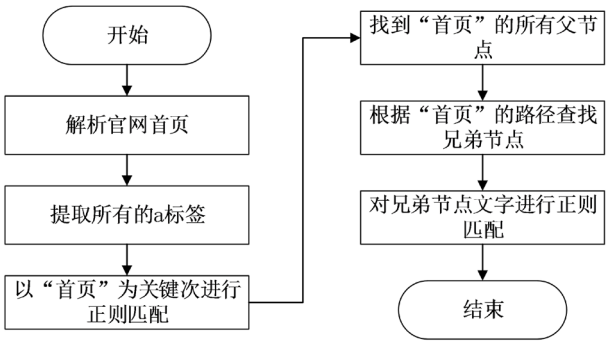

Figure 4. Flow chart for extracting product pages
图 4. 提取产品页面的流程图

1、首先解析官方首页，代码如下：

Document doc = null;

doc = Jsoup.*connect*(compan_url).get();

2、选择所有的 a 标签，代码如下：

Elements links = body.select(a);

3、对首页关键词进行正则匹配，代码如下：

String pattern = ”.*首.*页*.”;

提取 a 标签的文字内容并找到和正则表达式匹配的文字内容。

4、对于找到首页的标签，寻找父节点顺应至 Body 标签，把标签按顺序存入 Linked Hash Map，存入 Linked Hash Map 中而不存入 hashmaip 中是因为 Linked Hash Map 存在顺序性。最终获取到带有首页文字的路径。

5、通过得到首页的路径，开始查找与首页路径相同节点的文字内容，使用 jsoup 的选择器。select(“a.contains(产品)”).first()；找到产品的链接与文字内容并把文字内容和链接存入数据库中。

3.3. 企业官网产品页面的获取

网页解析成 Dom 树就是把 HTML 文档转化成 DOM 树的过程，将 HTML 文档中的元素映射成 DOM 树中的节点。DOM 树结构如图 5 所示。

3.4. 确定关键路径并提取信息

根据从同花顺网站上得到的候选关键词组，然后从官网产品首页解析的 DOM 树，转化成 xhtml 代码：doc.outputSettings().syntax(Document.OutputSettings.Syntax.xml).escapeMode(Entities.EscapeMode.xhtml)；去除 JS 代码。然后使用 NodeVisitor 去遍历文本节点，这里采用深度优先遍历。把文本节点与候选关键词组进行比较，比较规则如下：

1、$\exists X_i \subseteq Y_K \vee \exists Y_k \subseteq X_i$，则 $\left(A-\{X_i\}\right)\cup\{Y_k\}$。设 $A-\{X_i\}$ 为 C，$\left(A-\{X_i\}\right)\cup\{Y_k\}$ 设为 D。

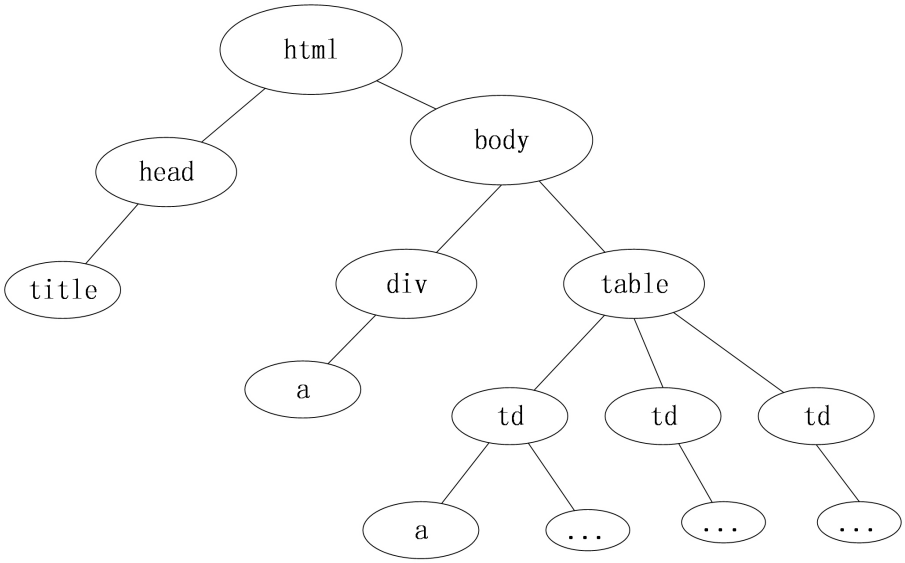

Figure5. DOM Tree Structure
图 5. DOM 树结构图

2、用 C 中的元素与 Y_K 进行词语相似度比较，若相似度 ＞90%则替换。

其中候选关键词组设为 A，网页文本设为 B。设 $A = \{X_i|x_i$ 为关键词$\}$，$B = \{Y_k|y_k$ 为网页文本内容$\}$。

根据比较规则得到被官网网页替换的关键词组，把关键词存入列表中。确定为关键词组。其中词语相似度比较工具用的是 WordSimilarity.下一步需要得到这些关键词的路径，这里采用基于 DOM 树逆序解析的方法，遍历 DOM 树，找到关键词所在的最终子节点(一般情况是 a 节点)，然后逆序找到其所有的父节点，直至到 body 结束。然后把这些节点标签和属性对应的存入 LinkedHashMap 中。把所有关键词的路径都存入到一个 LinkedHashMap 中，然后将 LinkedHashMap 转化成树。通过比较数组相同元素的个数，记录相同路径出现的次数并统计出现次数。根据统计出的次数，把次数最多的路径确定为关键路径，通过关键路径用 jsoup 的选择器进行提取。

4. 实验结果与分析

为了验证本文提出的方法的有效性，分析了公司类型为汽车零部件的 103 个官方网站，其中 91 个官网上有产品的介绍。根据本文提出的导航栏提取规则，提取出 55 个官网上的产品链接信息。其中 36 个网站的链接未被提取出来，主要原因分为以下几点：导航为图片、导航无首页关键字、官网的首页需要进一步点击。

根据提取出的 55 个公司官网的产品页面，使用本文提出的基于 DOM 树的统计关键词路径的 WEB 信息提取方法进行信息提取。其中正确抽取的占比 53%，抽取错误的占比 35%，未抽取出来的占比 7%，补充错误的占比 5%。如图 6 所示。抽取错误的原因是确定关键词路径时发生错误，即出现最多的那条路径不是产品信息所在的路径。未抽取出来的原因是没有找到关键路径。补充错误的原因是公司官网的一些其他无关介绍和产品介绍的关键路径相同故全部提取，从而产生补充错误。

实验结果表明，本文能够有效的提取企业官网中的结构差距较大、没有具体特征规律可寻的网站上的信息，解决了网站结构差距较大提取困难的问题。由于目前大部分文献都是研究较为特定的一些网站，比如电商网站，所以本文方法在一定程度上填补了提取网页结构差别较大网页上信息的空白，给出了一种解决海量网页信息提取的方法。

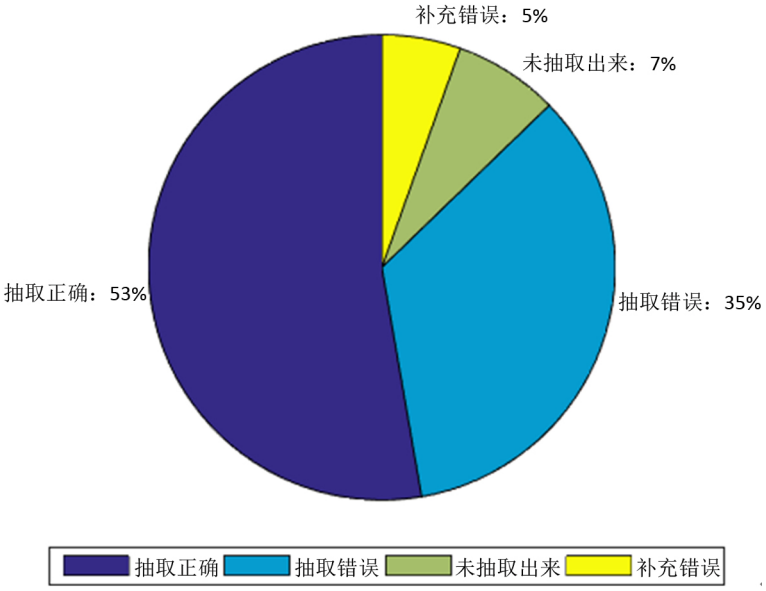

Figure 6. The proportion of statistical keyword information extraction results based on DOM tree
图 6. 基于 DOM 树的统计关键词信息提取结果占比图

5. 总结

基于 DOM 树的统计关键路径的方法对企业官网的信息进行提取，解决了企业官网结构差距比较大而带来的信息提取比较困难的问题。本文通过该方法，实现了在汽车零部件行业官网上根据需求提取产品信息。实验结果表明，本文提出的方法能够有效解决企业官网由于结构相差较大而带来的信息提取问题。

基金项目

本研究由山东省优秀中青年科学家科研奖励基金(ZR2016FB21)提供支持。

参考文献

[1] 施生生. 精确 Web 信息抽取关键技术与系统研究[D]: [博士学位论文]. 南京: 南京大学, 2017.
[2] 王一洲, 陈星, 戴远飞. 基于网页聚类的正文信息提取方法[J]. 小型微型计算机系统, 2018, 39(1): 111-115.
[3] 孙景春. 基于视觉块识别的网页元数据提取方法[D]: [硕士学位论文]. 南京: 东南大学, 2017.
[4] 马金娜. 基于 DOM 树节点重要度的 WEB 主题信息提取研究[D]: [硕士学位论文]. 重庆: 西南大学, 2016.
[5] 高峰, 刘震, 高辉. 结合有监督广度优先搜索策略的通用垂直爬虫方法[J]. 计算机工程, 2018, 44(11): 289-299.
[6] 赵朗. 基于深度学习的 Web 信息抽取研究与实现[D]: [硕士学位论文]. 杭州: 浙江大学, 2017.
[7] 王健. 基于 Hadoop 的 Web 页面正文抽取技术的研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2017.
[8] 鲁雷. 基于标签的网页信息抽取方法研究[D]: [硕士学位论文]. 青岛: 中国石油大学(华东), 2016.
[9] 寇月, 李冬, 申德荣, 于戈, 聂铁铮. 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.
[10] 毛凯. 基于 Jsoup 的通用网页采集系统的设计与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2015.
[11] 张瑞雪, 宋明秋, 公衍磊. 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, 225.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org