图书网页的自动识别及书目信息抽取研究*

李湘东[1,2]　霍亚勇[1]　黄　莉[3]
[1](武汉大学信息管理学院　武汉 430072)
[2](武汉大学信息资源研究中心　武汉 430072)
[3](武汉大学图书馆　武汉 430072)

摘要: 【**目的**】以相关的图书类网页为对象, 研究图书网页的自动识别及书目信息抽取方法。【**方法**】在分析不同图书网页标签使用特征、布局结构以及书目信息表征的基础上, 通过定义通用规则及共现词和页面分析等技术建立图书网页自动识别及书目信息抽取模型。【**结果**】实验证明, 该模型针对来自一般性网站的图书网页识别率可以达到近 80%, 而针对各类图书网页书目信息的抽取准确率平均也达到 79%左右。【**局限**】该方法中阈值的设定综合考虑了多种类型图书网页信息特征, 但对于部分特征极其特殊的网页存在误判现象, 若进一步改进算法, 可能效果更好。【**结论**】此方法对于各种类型图书网页的自动识别和书目信息抽取均能取得比较理想的效果, 普适性较强, 同时也为图书网页信息组织管理和自动分类研究奠定了基础。

关键词: 图书网页　书目信息　自动识别　信息抽取
分类号: TP391

1　引　言

伴随着互联网的快速发展, 网络信息逐渐覆盖了政治、经济、文化等各个领域。网页文档本身作为一种信息传递的载体, 丰富人们信息来源的同时, 也给人们获得有用信息带来了极大的困难。面对浩瀚的网络信息资源, 如何有效地抽取网页信息, 帮助用户快速获得所需要的细粒度信息, 已经成为了迫切需要解决的问题。

图书网页是指含有图书的书名、作者、出版社、摘要及 ISBN 号等书目信息的网页; 不含书目信息或者不含有完整书目信息的任何网页, 本文统称为非图书网页。作为一种获取图书信息的重要渠道, 准确高效的信息抽取技术, 对图书网页的组织、管理和书目信息获取都起着至关重要的作用。对于图书馆的书目信息录入来说, 自动化的信息抽取技术, 不但可以弥补传统手工录入的不足, 节省大量人力和时间, 而且可以为读者提供更加优质的服务; 对于网络用户来说, 合理地将网络中异构的书目信息进行抽取和集成, 将有利于提高用户体验, 更好地实现图书信息资源共享; 对于出版发行机构、书店以及图书经销商等销售企业来说, 图书作为一种重要的文献类型, 已成为其主要的收集和处理对象, 高效准确的图书信息抽取技术对相关企业具有较强的商业价值。

为了能够准确高效地从网页中自动识别并抽取书目信息, 本文以相关的图书类网页为研究对象, 根据网页的标签使用特征、布局结构以及信息表征, 构建图书网页自动识别及书目信息抽取模型, 通过共现词技术和信息分析方法实现图书网页的自动识别, 并利用页面分析和启示性规则等过滤技术, 自动抽取图书网页中的书目信息。实验证明, 此方法针对来自图书专门性网站和各类一般性网站中的图书网页自动识别和抽取均能取得较为理想的效果。

收稿日期: 2013-12-18
收修改稿日期: 2014-01-11
*本文系湖北省高校图工委基金项目“传统分类体系下多种类型文献自动分类研究”(项目编号: 2012YB02)的研究成果之一。

2 相关研究

在网页中, 抽取信息经历了不断进化与完善的过程, 从网页信息抽取的对象角度分析, 其研究多集中在新闻类网页的标题或内容的抽取[1-7]。近些年来对于博客网页的信息抽取也逐渐成为热门的话题[8-11]。随着网络购物的快速崛起, 对于网页商品信息抽取工作的研究[12-14]也具有了时代意义。有效地实施信息抽取工作不仅对中文网页尤为重要, 同时对少数民族网页信息抽取的意义也是不言而喻的, 目前研究成果主要包括文献[15-17]。此外, 对于旅游、医药、农业等领域的网页信息抽取也有部分学者分别做过相关研究工作[18-20]。虽然目前对于网页信息抽取的研究涉及了诸多的对象或领域, 但尚未发现针对图书网页开展书目信息抽取的研究。

从网页信息抽取方法角度分析, 目前的研究主要围绕以下几个方面开展:

(1) 基于网页DOM树分析, 如文献[3,17,18,20]将解析后的树形结构网页根据特定的抽取规则或算法完成信息抽取。

(2) 基于网页结构和视觉特征方法, 如文献[1]充分利用HTML标记和新闻特征, 从发布者对新闻内容的修饰角度出发, 提出了内容分割的抽取算法; 文献[8]针对博客网页的特点与规律, 提出一种根据网页结构和关键字计算相似度的方法识别博客网页。

(3) 基于统计方法, 文献[6]提出了一种基于统计的方法实现抽取所有的正文信息都放在一个table中的特定情况, 文献[7]提出的基于标记窗(Tag Window)的网页正文获取方法扩大了成果[6]使用范围, 有效地弥补了它不能处理非table结构的网页正文提取的缺点。

(4) 结合复杂语义特征和网页结构的方法, 文献[10]充分利用页面中“首页”等指示性短语的特点, 提出利用具有明确语义和功能指示作用的功能语义单元来抽取评论信息的技术, 有效抽取博客评论, 文献[14]充分利用了商品的语义特征和网页结构的表现形式, 提出了一种基于语义熵的节点聚集度判别算法, 生成DOM树; 通过计算语义熵来获取商品的属性标签和其对应的属性值。文献[19]提出了基于多维语义的互联网药品信息提取方法, 通过使用从多个维度描述互联网药品领域语义的语义词典, 发现来自不同网页内容中隐藏的共性信息, 从而获取药品信息。

总之, 对于网页信息抽取的研究已经涉及到诸多领域, 但是这些方法都存在一些不足, 传统的依据特定的数据源或者训练集生成抽取模板来完成信息抽取的方法, 效率较低, 通用性较差; 单独基于视觉特征或统计的方法, 又很难排除网页噪声的干扰; 基于DOM树的方法准确率有了保证, 但是实现过程较为复杂, 时间效率不佳; 结合语义特征和网页结构的方法, 能够准确定位抽取元素, 但是构建词义规则的方法相当复杂。而本文应用的共现词等相关技术具有简单、高效等特点, 能够充分根据图书网页的特点弥补上述方法中的不足之处, 应用于图书网页识别及抽取的相关研究中, 具有较强的实用价值。

图书网页上的信息主要包括两类: 显性知识, 即浏览器直接呈现给用户的可视化书目信息; 隐性知识, 即对信息进行修饰的各类网页标签。因此, 本文充分利用图书网页自身特征, 将显性知识和隐性知识相结合, 提出了一种图书网页识别和信息抽取的方法, 有效地解决了由于网页差异化而难以高效、准确抽取有关图书的书目信息的难题。

3 基于规则的图书网页书目信息抽取

3.1 模型主体框架

本文研究目的在于图书网页的自动识别和书目信息抽取, 因此, 针对来自图书专门性网站和各类一般性网站等不同来源的图书网页, 首先使用信息分析和共现词技术区分图书网页和非图书网页; 然后对图书网页进行预处理, 过滤无用标签及字符; 最后根据信息抽取规则完成图书网页的书目信息抽取。基本模型框架如图 1 所示。

3.2 图书网页的识别

通过 Google 的布尔逻辑与高级检索, 选取能够代表图书网页独有特征的共现词进行检索, 并且认为“图书、ISBN、出版社、价格”共现一次, 则所得网页极有可能为图书网页, 将此类网页进行保留, 以备二次判别, 因此使用此 4 项关键词获得来自各种图书专门性网站以及各类一般性网站、混合有图书网页和非图书网页的网站上的相关网页数千篇。对其中的部分网页内容进行统计和分析发现, 同时含有该 4 项关键词的网页中, 70%以上为图书网页。由此可见, 该 4 项

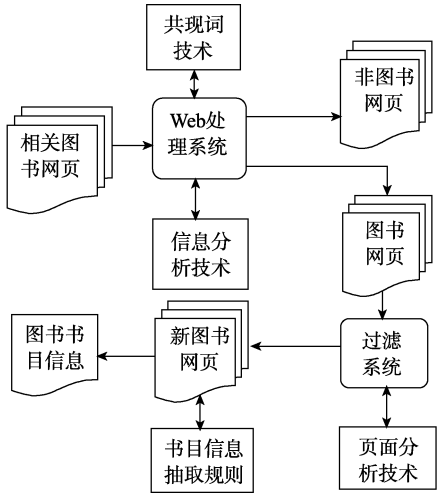

图 1　图书网页识别及书目信息抽取框架

关键词是采用共现词技术实施判别图书网页的基本要素。但为进一步提高图书网页的识别率, 本文同时使用信息分析技术对 Google 检索出的网页中的显性知识分析发现, 在非图书网页中, 虽然也含有“图书、ISBN、出版社、价格”等 4 项关键词, 但其相关信息并不完整, 如“ISBN:”、“出版社:”等, 其内容为空值; 在这些非图书网页中出现此类现象的比例高达近 80%。而与之相比, 图书网页中仅有不到 8%的网页出现了上述现象。除此之外, 它们在网页标签使用特征等方面基本无差别。因此, 利用这一显著特点, 与共现词技术相结合, 可以有效区分图书网页和非图书网页, 达到图书网页自动识别的目的, 在本研究中已达到 78%的识别率, 详见实验 1。

3.3　图书网页的书目信息抽取

(1) 预处理阶段

本文首先针对网页标签进行过滤处理, 删除对图书网页书目信息抽取没有实质意义的标签, 排除噪声干扰, 缩小抽取范围, 为下一步书目信息抽取奠定基础。通过对各类网页结构布局和标签使用特征比较分析发现, 网页标签大致分为以下几类:

类型 1: 主体标签, 主要由<HEAD>和<BODY>两部分构成, 大量描述网页的其他标签可以嵌套在其中。

类型 2: 布局标签, 主要是为合理布局网页而使用的部分标签, 包括<TABLE></TABLE>、<DIV></DIV>、<TR></TR>、<TD></TD>等。

类型 3: 视觉标注标签, 主要是方便浏览者阅读, 突出重点信息而使用的一类标签, 包括<SPAN></SPAN>、<B></B>、<STRONG></STRONG>、<IMAGE>等。分析发现, 大量的视觉标注标签中基本不会含有主要信息, 这些标签很可能对信息抽取造成干扰, 因此, 可以首先采用顺序遍历的方法, 过滤掉这部分标签; 而对于布局标签, 虽然含有部分网页噪声, 但是绝大部分的主要信息包含在其中, 所以将其保留。

(2) 书目信息抽取阶段

图书网页的信息主要是用来描述图书内容的, 往往具有较长的文字表达篇幅。HTML 的主体标签中, <TITLE>、<KEYWORDS>、<DESCRIPTION>, 这些标签中的内容高度概括了网页的主要内容, 在图书网页中表现为: 介绍图书名称、作者、出版社等相关书目信息。通过对识别出的图书网页统计发现, 其中 85%左右的网页使用<KEYWORDS>和< DESCRIPTION >标签准确地描述了图书的基本书目信息, 而且与<TITLE>标签内容具有高度相关性, 为图书网页书目信息的抽取提供了便利。针对第一步预处理后的结果, 图书网页信息抽取步骤归纳如下:

①使用正则表达式抽取<TITLE>、<KEYWORDS>、<DESCRIPTION>标签中的信息。

②对网页源文件按顺序分段截取“>”和“<”中间的内容, 并输出结果。

③过滤步骤②中截取得到的特殊字符, 得到仅含有标点符号的字符串。

④删除步骤③结果中的非汉字文本内容, 排除对书目信息抽取可能造成的干扰, 保证了下一步计算字符串长度时, 所统计的为中文字符个数。

⑤设定特定阈值 L, 根据字符串长度, 获得此范围内的段落; 对各段及步骤①所获得图书名称进行预处理, 将每个文本看作是由一组词(t1,t2,…,tn)构成, 采用向量空间模型对其进行量化, 使用最常用的 TF*IDF 方法, 为每个特征词赋权值 wi, 如公式(1)所示, 因此将(t1,t2,…,tn)分解得到的正交词条矢量组就构成了一个文本的向量空间,使用公式(2)分别计算每段文本与书名的相似度, 抽取最相似段落作为图书网页书目信息。

$$w(t_j,\ d_i)=\frac{tf_i\times\log(\frac{|D|}{df_i}+1)}{\sqrt{\sum_{i=1}^{n}(tf_i\times\log(\frac{|D|}{df_i}+1))^2}} \qquad (1)$$

其中, tf_i 表示特征词 t_j 在当前文本 d_i 中出现的频率, |D|

是文本总数，df_i 是在所有文本中特征词 t_j 出现的频率，其中为了消除文档长度的影响，对所有的特征项权重进行规一化处理。

$$sim(U,V)=cos(U,V)=\frac{\sum_{i=1}^{n} W_{ui}\times W_{vi}}{\sqrt{\sum_{i=1}^{n} W_{ui}^2}\times\sqrt{\sum_{i=1}^{n} W_{vi}^2}} \quad (2)$$

⑥采用综合评定等级方法评价实验结果。

4 实验与结果分析

4.1 实验数据集

本文力图建立一种具有普适性的图书网页识别和书目信息抽取模型，因此选取了来自各种图书专门性网站以及各类一般性网站中的网页进行实验。目前，自动抓取各类网页的手段繁多，既可借助火狐浏览器的 DataScraper 和 MetaStudio 等辅助插件来完成，也可以使用各种开源网络爬虫软件获得。但为确保实验的透明性和准确性，本研究在图书网页信息抽取过程中使用的实验材料主要采用手动下载的方式获取。这些网页既包括国内知名的图书专门性网站，如当当网(http://book.dangdang.com/)、孔夫子网(http://www.kongfz.com/)上的网页,也包括利用 Google 搜索引擎以“图书 ISBN 出版社 价格”等 4 个关键词和布尔逻辑与方式检索得到的任意网页(检索时间: 2013 年 4 月 15 日 10 时 05 分)。为方便比较和表述，以下各种材料以 100 篇为单位统一进行统计分析，如表 1 所示:

表 1 实验材料来源

网页来源	样本数量(篇)
Google 网页	100
当当网	100
孔夫子网	100

4.2 实验结果与分析

(1) 实验 1: 图书网页识别

采用人工标注方法对来自 Google 检索的 100 篇实验材料进行处理，之后由系统自动识别，实验结果如表 2 所示。

(2) 实验 2: 确定阈值 L 范围

考虑到实验样本间的均衡性，图书网页书目信息抽取过程使用的实验材料构成如下所述: 实验 1 识别出的图书网页、来自于当当网和孔夫子网等图书专门性网站的图书网页各 100 篇。为获得最佳 L 值，通过设定不同取值区间，分别对以上三个不同来源的图书网页进行实验，用信息保留概率，即: 图书网页中摘要、简介等信息包含在过滤后的字符串内的概率，来确定最佳的 L 区间。

表 2 图书网页识别结果

材料来源	网页总数(个)	正确结果(个)	错误结果(个)	准确率(%)
Google 网页	100	78	22	78

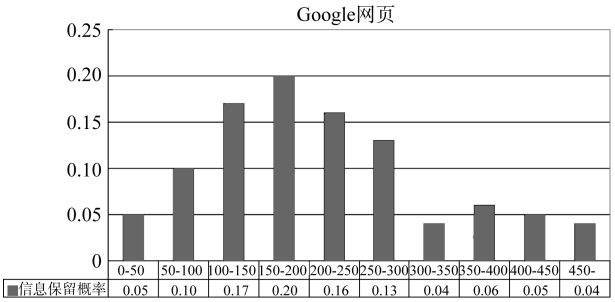

图 2 Google 网页确定阈值 L 的实验结果

由图 2 的实验结果可以得出: 当阈值 L 在 0-50 和 300 以上的时候，信息保留概率均在 0.1 以下，最高概率仅为 350-400 时的 0.06，未达到平均水平；而 L 在区间 50-300 时的平均水平均在 0.1 以上。因此对于来自 Google 检索的图书网页，L 在 50-300 时，信息可以被有效保存下来，且概率总和为 0.76。

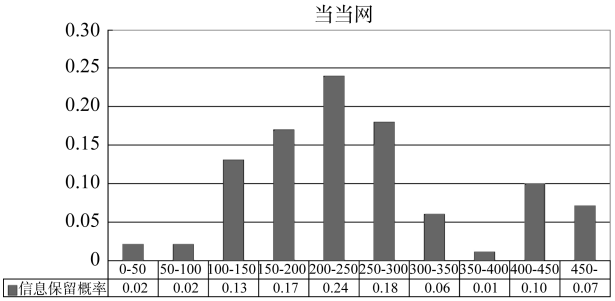

图 3 当当网确定阈值 L 的实验结果

由图 3 的实验结果可以得出: 当阈值 L 在 0-100、300-400 和 450 以上的时候信息保留概率均在 0.1 以下，最高概率仅为 0.07；虽然在区间 400-450 时信息保留概率达到 0.1，但比较孤立，不具有代表性。因此认为平均概率在 0.1 以上多集中在区间 100-300，对于来自当当网的图书网页，L 在 100-300 时，信息可以被有效保存下来，且概率总和为 0.72。

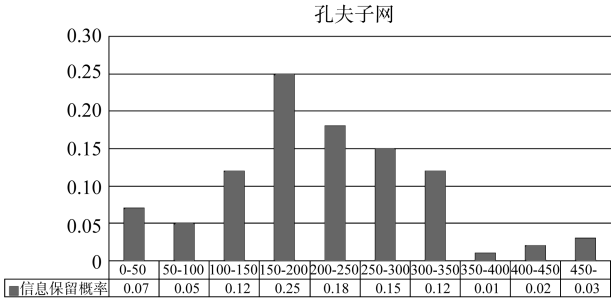

图 4　孔夫子网确定阈值 L 的实验结果

由图 4 的实验结果可以得出: 当阈值 L 在 0-100 和 350 以上的时候信息保留概率均在 0.1 以下, 最高概率仅为区间 0-50 上的 0.07。因此, 对于来自孔夫子网的图书网页, 认为 L 在 100-350 时, 信息可以被有效保存下来, 且概率总和达到 0.82。

综上所述, 针对三组不同实验材料, L 均有各自最佳的取值范围。如图 5 所示, L 在不同取值范围内, Google 网页、当当网和孔夫子网的图书网页信息保留概率大致服从正态分布, 因此可知, 将 L 阈值界定在 100-300 范围内, 效果比较理想。

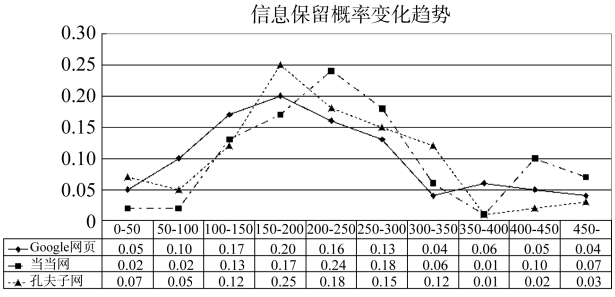

图 5　各类网页不同 L 取值下信息保留概率变化趋势

(3) 实验 3: 图书网页书目信息抽取

本文对图书网页书目信息抽取结果采用综合评定等级的方法, 将抽取结果与正确的书目信息从内容完整性和相关性两个方面进行衡量, 等级共分为 5 级。第 1 级: 能够 100%正确完整地抽取信息; 第 2 级: 能够 90%以上完整抽取信息, 但有小部分信息丢失; 第 3 级: 能够 70%-90%完整抽取信息, 但掺杂了部分无关信息; 第 4 级: 仅能很少部分抽取出书目信息, 噪声很大; 第 5 级: 抽取的信息与书目信息无关, 属于纯噪声信息。

从表 3 可以得出: 来自于当当网的 100 篇实验材料在抽取信息时, 其完整性和相关性方面要明显好于另外两个材料来源, 准确率达到 84%, 主要是因为当当网的图书网页布局界限比较分明, 标签使用更加规范, 图书摘要和相关简介信息与抽取模型的设计更加匹配。来自孔夫子网站的图书网页, 虽然抽取信息效果与当当网的有一定差距, 但准确率也达到 79%; 其主要原因是来自孔夫子网的图书网页, 图书介绍远不如当当网规范, 其中部分网页甚至省略或者使用较少文字描述图书内容, 这就导致抽取过程一些纯噪声信息的出现。虽然通过 Google 检索获取的图书网页抽取书目信息效果不及当当网和孔夫子网, 但是准确率也达到了 75%左右, 分析其主要原因是因为 Google 检索出的网页分别来自以非图书网页为主的各类一般性网站, 无论是在网页设计布局方面, 还是网页信息纯度方面都存在很大差异性, 一些以盈利为目的的商业网站, 经常在主要信息附近添加广告内容, 这就给信息提取造成了很大干扰, 部分网页在设计时为了追求个性化, 经常将主要信息使用网页标签进行分割处理, 很难将其辨别抽取出来。即使存在诸多问题, 本文提出的图书网页书目信息抽取模型仍然取得了较为理想的效果, 平均准确率达到 79%左右, 具有较强的实用价值。

表 3　各类图书网页书目信息抽取结果

来源 \ 等级	1 级	2 级	3 级	4 级	5 级	合计(篇)	准确率(%)
Google 网页	38	22	15	4	21	100	75
当当网	54	19	11	10	6	100	84
孔夫子网	45	15	19	6	15	100	79

5　结　论

本文提出的模型主要针对各类图书网页进行网页自动识别及书目信息抽取。实验证明, 该方法实现简洁, 复杂度低, 并且对于当前网页多样化、差异化、复杂化等特点具有很强的普适性, 效果比较理想; 但是, 对于极特殊的部分网页, 还是存在误判和抽取失败等问题, 下一步的工作重点在于完善抽取方法, 进一步提高抽取准确率。

参考文献:

[1] 罗永莲, 秦振吉. 新闻网页主题内容提取方法研究[J]. 微计算机应用, 2007, 28(5): 556-560. (Luo Yonglian, Qin Zhenji. Research on Extracting Topic Content from News

Web Pages [J].Microcomputer Applications, 2007, 28(5): 556-560.)

[2] 施洋, 张奇, 黄萱菁. 含有语义特征的网页新闻自动抽取[J]. 计算机工程, 2010, 36(7): 173-178. (Shi Yang,Zhang Qi, Huang Xuanjing. Automatic Web News Extraction with Semantic Features[J]. Computer Engineering, 2010, 36(7): 173-178.)

[3] 孔胜, 王宇. 一种基于正文特征的新闻网页抽取方法[J]. 情报杂志, 2010, 29(8): 122-125. (Kong Sheng,Wang Yu.A News Page Information Extraction Based on Web Feature [J]. Journal of Intelligence, 2010, 29(8): 122-125.)

[4] 刘伟, 严华梁. 一种统一的 Web 新闻对象自动抽取方法[J]. 计算机工程, 2012, 38(11): 167-169. (Liu Wei, Yan Hualiang. A Unified and Automatic Web News Object Extraction Approach [J]. Computer Engineering, 2012, 38(11): 167-169.)

[5] 朱红灿, 龙朝阳. 基于熵的新闻网页抽取方法的研究[J]. 现代图书情报技术, 2007(4): 48-51. (Zhu Hongcan,Long Chaoyang. An Entropy-Based Approach for News Article Extraction from Web Page [J]. New Technology of Library and Information Service, 2007(4): 48-51.)

[6] 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.(Sun Chengjie, Guan Yi. A Statistical Approach for Content Extraction from Web Page[J]. Journal of Chinese Information Processing, 2004, 18(5): 17-22.)

[7] 赵欣欣, 索红光, 刘玉树. 基于标记窗的网页正文信息提取方法[J]. 计算机应用研究, 2007, 24(3): 144-148. (Zhao Xinxin, Suo Hongguang, Liu Yushu. Web Content Information Extraction Method Based on Tag Window[J]. Application Research of Computer, 2007, 24(3): 144-148.)

[8] Zheng S Y, Song R H, Wen J R. Template-independent News Extraction Based on Visual Consistency[C]. In: Proceedings of the AAAI'07, Vancouver, Canada. 2007.

[9] 郑德权, 张迪, 赵铁军, 等. Blog 网页分类与识别技术研究[J]. 通信学报, 2007, 28(12): 156-160. (Zheng Dequan,Zhang Di, Zhao Tiejun. Study on the Classification and Identification of Blog Pages[J]. Journal of Communication, 2007, 28(12): 156-160.)

[10] 范纯龙, 夏佳, 肖昕, 等. 基于功能语义单元的博客评论抽取技术[J]. 计算机应用, 2011, 31(9): 17-23. (Fan Chunlong, Xia Jia, Xiao Xin, et al. Extraction Technology of Blog Comments Based on Functional Semantic Units[J]. Journal of Computer Application, 2011, 31(9): 17-23.)

[11] 曹冬林, 廖祥文, 许洪波, 等. 基于网页格式信息量的博客文章和评论抽取模型[J]. 软件学报, 2009, 20(5): 1282-1291. (Cao Donglin, Liao Xiangwen, Xu Hongbo, et al. Extraction Model Based on Web Format Information Quantity in Blog Post and Comment Extraction[J]. Journal of Software, 2009, 20(5): 1282-1291.)

[12] 唐伟, 洪宇, 冯艳卉, 等. 网页中商品"属性-值"关系的自动抽取方法研究[J]. 中文信息学报, 2012, 27(1): 21-29. (Tang Wei, Hong Yu, Feng Yanhui, et al.Automatic Extraction of the Product "Attribute-Value" Pair from the Web Pages[J]. Journal of Chinese Information Processing, 2012, 27(1): 21-29.)

[13] 杨舟, 卓林, 赵朋朋, 等. 一种针对商品数据记录的自动抽取方法[J]. 计算机工程, 2010, 36(23): 262-265. (Yang Zhou, Zhuo Lin, Zhao Pengpeng, et al. Automatic Extraction Method for Product Data Records[J]. Computer Engineering, 2010, 36(23): 262-265.)

[14] 吴晓彦, 郑骁庆, 顾轶灵, 等. 基于结构语义熵的网上商品信息提取系统[J]. 计算机应用与软件, 2010, 27(9): 49-53. (Wu Xiaoyan, Zheng Xiaoqing, Gu Yiling, et al. Extraction Algorithm of Merchandise Information on Networks Based on Structured-Semantic Entropy[J]. Computer Application and Software, 2010, 27(9): 49-53.)

[15] 李文博. 基于 XML 的藏文网页的信息抽取与转存技术研究[D]. 兰州: 西北民族大学, 2006. (Li Wenbo. The Research of XML-Based Tibet Web Page Information Extraction and Conversion Storage[D]. Lanzhou: Northwest University for Nationalities, 2006.)

[16] 蔡李, 单艳, 薛化建. 维吾尔文网页正文抽取系统的研究与实现[J]. 计算机工程与设计, 2012, 33(2): 551-555. (Cai Li, Shan Yan, Xue Huajian. Research and Implementation of Uyghur Web Content Extraction System[J].Computer Engineering and Design, 2012, 33(2): 551-555.)

[17] 王瑞, 周喜, 李晓. 基于正文相关度的维吾尔网页正文提取[J]. 计算机工程, 2012, 38(21): 153-160. (Wang Rui, Zhou Xi, Li Xiao. Content Extraction of Uighur Web Based on Content Correlativity[J]. Computer Engineering, 2012, 38(21): 153-160.)

[18] 王爽. 面向数字旅游网页的 Web 信息抽取技术研究[D]. 西安: 西安电子科技大学, 2012. (Wang Shuang. Research of Web Information Extraction Technology Oriented to Digital Tourism Website[D]. Xi'an: Xidian University, 2012.)

[19] 顾轶灵. 基于多维语义的互联网药品信息提取方法[J]. 计算机系统应用, 2011, 20(11): 50-54. (Gu Yiling. Multidimensional-Semantics-Based Web Medicine Information Extraction[J]. Computer Systems and Applications, 2011, 20(11):

50-54.)
[20] 王文生, 谢能付. 基于 Web 的农业信息自动抽取方法研究[C]. 见: 全国农业信息分析理论与方法学术研讨会. 2007: 77-83. (Wang Wensheng, Xie Nengfu. Research on Web-based Agriculture Information Extraction[C]. In: National Seminar on Agricultural Information Analysis Theory and Method. 2007: 77-83.)

作者贡献声明:

李湘东: 提出研究方向和思路, 介绍相关技术的应用;
霍亚勇: 实验流程设计, 实验材料采集及分析, 进行实验;
黄莉: 实验数据分析;
霍亚勇: 论文起草;
李湘东: 最终版本修订。

(通讯作者: 霍亚勇 E-mail: 413261403@qq.com)

Study of Book Pages Automatic Identification and Bibliographic Information Extraction

Li Xiangdong[1, 2]　Huo Yayong[1]　Huang Li[3]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for the Studies of Information Resources, Wuhan University, Wuhan 430072, China)
[3](Wuhan University Library, Wuhan 430072, China)

Abstract: [**Objective**] The article studies the book pages automatic identification and the thematic information extraction method, which sets relevant book pages as the objects. [**Methods**] Based on the analysis of the features usage of different book pages labels, layout structure and theme information representation, the article establishes a book pages automatic identification and thematic information extraction model through defining general rules, using co-occurrence words and pages analysis, etc. [**Results**] The result shows that the book pages identification rates from the general Web sites of the model can reach nearly 80%, and the average abstraction rates of the thematic information about kinds of book pages can reach nearly 79%. [**Limitations**] The method of threshold setting comprehensively considerates various types of books characteristics of Web information, but for some features extremely special webpages exists misjudgment phenomenon, if the algorithm is further improved, it may be better. [**Conclusions**] The method for automatic identification of all kinds of book pages and thematic information extraction can obtain ideal result, it has a strong universality, at the same time, it also has laid the foundation for the book Web page information organization management and automatic classification research.

Keywords: Book pages　Bibliographic information　Automatic identification　Information extraction