数字图书馆Web学术资源信息的分块采集研究

王兰成 朱建华
（南京政治学院上海校区军事信息管理系，上海 200433）

摘 要：在数字图书馆Web学术信息资源的优化采集中，有效结合网页空间特征、内容特征和标签信息对网页进行分块，研究对分块结果进行识别和合并，然后输出网页的主题文本和相关链接块集合，最后通过实验分析该方法能够进一步去除页面中噪音、准确地分析页面的主题相关性和提高Web主题信息采集的质量。

关键词：数字图书馆；Web学术资源；自动采集；信息系统
分类号：TP393 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2012.06.015

Research of Page Segmentation for Digital Library Based on Web Academic Resource Crawling

Wang Lancheng, Zhu Jianhua
(Department of Information Management, Shanghai Political College PLA, Shanghai 200433)

Abstract: Web academic resource crawling on digital library is an important research area. The effective integration of web space characteristics, content characteristics and label information on the web pages block are researched. The identification and the merger of results on Page Segmentation are studied. The subject of the final text page and related links block collection are output. It is fact that more accurate analysis of the topic pages and improve the quality of Web information collection subject.

Keywords: digital library, web academic resource, automation crawling, information system

1 引言

开展知识服务的先进技术之一是对Web学术资源信息的优化采集。据统计，目前最大的搜索引擎只能覆盖网络内容的30%~40%，通用搜索引擎提供的信息检索服务并不能满足人们日益增长的对个性化服务的需求，特别是随着信息多元化的增长，千篇一律地给所有用户同一个入口显然已经不能满足用户更深入的查询需求。尤其是在数字图书馆系统中，当需要特定专业或主题的资料时，单纯依靠通用搜索引擎难以满足专业信息服务的需求[1]，而主题搜索引擎能发现满足主题的网站或网页。数字图书馆系统中主题爬行系统和通用搜索引擎不能互相替代[2]，通用搜索引擎能满足一般读者的搜索需要，仍是Web通常学术资源的检索工具，主题搜索引擎则能满足专业读者和科研工作者的信息查询需求，检索查准率比较高。

基于主题学术资源的Web信息采集技术，可有效结合网页空间特征、内容特征和标签信息对网页进行分块，研究一种综合考虑词频信息、位置信息、页面结构以及分块特征的页面信息模型的构建方法，对分块结果进行识别合并，最终输出网页的主题文本和相关链接块集合。实验证明，该方法能最大限度地去除页面中噪音，更加准确地分析页面的主题相关性和提高Web主题信息采集的质量，从而实现数字图书馆系统中Web学术资源信息的优化采集。

第一作者简介：王兰成（1962- ），男，南京政治学院教授，博导，研究方向：计算机情报管理。
收稿日期：2012年7月29日。

2 主题搜索的分块采集

Web学术资源存在一些分析特征。按文件类型，目前许多Web学术网页仍采用HTML格式，尤其是Web2.0数字图书馆系统的社区学术资源等，这类网站存在噪声信息；按结构特征，一些著名的门户网站包含大量新创的学术资源，也是数字图书馆用户的重要信息源。提取和利用数字图书馆系统网站上这些集中型学术资源，还可进一步提高质量和效率。

2.1 结构特征的网页分块

Web学术网页除了主题内容之外，常包含如导航、广告、交互信息等不少与主题无关的内容。另外，Web学术网页可能包含多个不完全相关的主题，每个主题分布在网页的不同区域，且各部分的重要性不一样。主题爬虫以网页中的块为处理单位，能更好地获取网页中的信息，从而爬行到更多相关主题的网页。

目前，大多数Web学术网页采用HTML语言，文档包含页面本身文本和表示页面元素、结构、格式、链接HTML标签的两种信息。在数字图书馆系统中，用户浏览一个网页会把它视为一个不同的语义对象的集合，而不是一个简单的对象，如经常认为页面的重要主题内容在网页中间，导航链接在网页上方或是侧面等。许多图书馆网站的首页具有明显的分块特征，从视觉上可明显分出各主题资源、导航、Banner、Footer及广告等区域。这将帮助用户把页面分成不同的语义块，如果能通过对网页的结构进行分析，得到其页面分块规则，识别、提取该网页的主题文本内容，从网页中过滤掉噪音信息，区分出信息块的重要程度，特别是识别、提取出其中的相关链接块，必然能提高Web学术网页信息采集的质量。

2.2 网页分块的处理方法

近几年出现的一些分块算法，主要基于DOM树抽取网页内容分块方法（GDOM）[3]、利用网页页面布局分块的方法（YPM）[4]等。国内有学者研究新闻网站中结构复杂网页的分块算法TVPS[5]和基于网页分块的信息采集系统模型[6]等。这里给出一种面向Web学术网页主题爬行的分块方法，该方法充分利用网页的空间特征、内容特征和标签信息，并把这些信息有效地结合起来，对网页进行分块，最大限度地去除页面中噪音信息和噪音链接。

2.2.1 网页预处理

HTML文档有些是用工具辅助写的，有些则是网页编写者手工完成的，通常会出现一些错误，如无相应结束标记、标记嵌套不合理等，容易导致分块混乱。使用JTidy（http://jtidy.sourceforge.net）的类库，将其集成到系统中，页面会通过JTidy将源HTML文档转换成等价的XHTML文档，这样便于分块和内容提取。

清理页面后，需对整理好的规范网页构造DOM树。使用JAXP构建DOM树，主要是去掉DOM树中的无效节点，如不能表达页面布局结构的<SPAN>、<EM>、<STRONG>等标记。图1所示（b）是（a）经简化后的XHTML代码，代码量明显减少。

2.2.2 多特征的分块处理

（1）空间特征。网页分块的空间特征能够在一定程度上反映分块的重要性。通常情况下，网页作者将最重要的信息放置在网页的中间部分，并占用网页大幅版面，而将导航信息放在网页的上方或者左右一边，版权信息、联系方式等则被放置网页的最底部。网页分块的空间特征表示可分为绝对空间特征、相对空间特征、窗口空间特征。

绝对空间特征。网页被分割成若干矩形区域，每个矩形区域可以用块中心横坐标x和纵坐标y、矩

```
<TABLE>
  <TR>
    <TD><FONT face="黑体">
        <span style="font-weight:bold;">国家重点学科</span>
        </FONT>
    </TD>
    <TD><FONT face="黑体">批准时间:
  <span style="color:#FF6600;font-weight:bold;">2010 年 8 月</span>
        </FONT>
    </TD>
<TABLE>
```

（a）

```
<TABLE>
    <TR>
      <TD>国家重点学科</TD>
      <TD>批准时间: 2010 年 8 月></TD>
    <TR>
<TABLE>
```

（b）

图1 Web学术网页简化示例

形块宽度和高度特征值，其缺点是难以比较取自不同网页的分块，如一个较小网页中的大分块可能比一个较大网页中的小分块要小。

相对空间特征。用块中心横坐标x/网页宽度、块中心纵坐标y/网页高度、矩形块宽度/网页宽度以及矩形块高度/网页高度解决上述问题，但有可能导致一些重要的分块当作非重要分块来处理，如网页高度超过屏幕能够显示的高度时。

窗口空间特征。不再使用整篇网页的高度，而是使用一个固定高度的窗口（如浏览器窗口），表示为：块中心横坐标x/网页宽度、块中心纵坐标y/窗口高度、矩形块宽度/网页宽度及矩形块高度/窗口高度。

（2）内容特征。内容特征对重要度区分有很大的帮助。如网页主要部分通常带有图片、醒目标题和大段描述文字，而一条广告信息可能只包含若干图片却没有文字信息，导航区则通常会含很多链接。分块的内容特征可以概括为：包含的图片数量和大小、超链接数目和超链接文长度、带有<INPUT>和<SELECT>标签部分数目和大小、带有<FORM>标签部分数目和大小。

（3）标签信息。大多学术网页用<TABLE>和<DIV>标签来组织网页中不同主题的内容，在块的划分时没有考虑采用<TD>标签，原因是为了合理控制原子块的粒度大小，采用<TD>来划分块会造成过多的单元格块参与计算，这会影响本算法的效率，而且页面主要内容识别结果质量的提升并不是很明显；<HR>标签用来定义一水平线，在处理时区分两个分块；<P>标签用来划分段落，在网页分块中可以把不同的内容分到不同的内容块中，因为段落之间的主题联系非常紧密，所以在考虑<P>标签的时候还要参考其他标签。

2.2.3　网页信息的分块算法

Web学术网页信息的分块算法步骤描述如下。

步骤1：构造DOM树。读取HTML网页，对学术网页预处理后构造DOM树。

步骤2：简化DOM。DOM树中存在着大量标签，有部分标签是算法分析时需要用到的，但是有一些标签并没有用处，去掉DOM树中的无效节点能提高分析过程的效率。

步骤3：视觉信息分析。在得到了简化后的DOM节点树后，也就是在学术页面中块的基本元素获取后，对这些节点从背景色、字体颜色、字体特性（字体名称、大小、粗细、样式）等视觉方面作进一步的处理。首先读取当前页面链接的CSS文件，再获取页头内嵌CSS，构造一个以id，class为键的站点CSS列表；分析DOM树节点时根据标签的class和id属性，给DOM树中各节点附上视觉信息；因一个节点可能被重复定义CSS，按照浏览器解析的就近原则只取最近的CSS值。

步骤4：综合学术网页的内容特征，判断网页类型，如果网页中链接文字数与非链接文字数的比值小于某个阈值，则认为该网页是主题型网页。

步骤5：结合空间特征，利用标签信息迭代分块。首先定义一个有序标签集{<TABLE>/<DIV>，<TR>，<P>，<HR>，<UL>，<DIR>，<DL>}，基于列表中的第一个标签来划分一个Web页面从而识别各种块，然后基于第二个标签再来划分已识别出的块，以此类推。连续划分直到已划分出的块集中的任意块中不再有划分标签集中的标签，这就确保了块的原子性，并且在它们之上不会再有更进一步的划分。

步骤6：进入结束状态。

考虑到互联网中大多数的网页会使用<TABLE>或<DIV>标签组织网页的内容，以下给出在算法中控制分块粒度的一些启发规则。

规则1：如果该节点的所有子节点都具有相同的视觉特性（包括背景颜色，字体的颜色、大小等），且和该节点的视觉特性相同，则去掉该节点的所有子节点，也就是说该节点合并了该节点所有子节点；如果不相同，则可以根据实际情况进行合并或者不合并。

规则2：节点的尺寸限制。如果一个节点的尺寸过小认为是非布局节点，当节点的高度和宽度小于一定的阈值则不往下进行分块。

规则3：如果节点的子节点为纯文本、超链接列表、图片等则不再往下细划，一般为导航或列表块；如果节点的内容为flash或图片链接，则一般为广告块也不细分。

3　主题型采集内容块的识别与合并

通过上面的分块算法可以得到内容块的集合，这些内容块在大小、形式和重要性上是不一样的。主题型分块的识别方法较为直观，如果分块包含的是一个或若干个正文段落，信息块中包含了大量的文本信息和一些链接信息，且分块处于网页视觉的

中心位置，系统则认为是主题型分块。在主题型分块的识别过程中，本文具体研究以下两个问题。

（1）包含标题信息块的识别。这样的信息块往往是一个主题信息的开端部分，比如一篇报道分了好几个段落，而这个标题块则是起始段落，其识别借助于HTML标签<hl>，<h2>，…，<h6>，<b>等，在网页设计这些标签显示为不同程度的标题，也考虑其他的视觉效果和位置因素，如标题标签虽然没有被使用，但通过CSS视觉分析发现某分块中文字区别于周围其他文本且处于视觉中心位置，而且该分块之后还有大量文本，也将该分块作为标题块。在划分算法中标题块不仅可被认为是一个大信息块的开始，同样可作为前一信息块的结束。

（2）一些小信息块的合并。划分算法使得一些大块划分过于细致而产生了不少信息小块，这些信息块文本信息比较少，往往处于一些主题型分块的周围，其识别方法主要根据信息块的文本量和分块所处的位置。

对于链接型分块，则根据它们的空间特征区别对待。这些分块经常代表着导航栏、广告栏、版本信息，其中导航栏和广告栏一般富含链接和图片，而且位置相对比较偏，不会分布在网页中心位置；版本信息一般处于网页的最低端，有固定的启发信息，如在版本信息块中经常出现的一些关键字有“copyright”等，根据这些信息“较”或“很”容易判断出版本信息块。另外，很多页面块中的信息是用图像表示的，文字不出现或者仅对图像解释，这种信息块的内容无法提取，如果它的位置不处于网页中心位置则被视为噪音信息块。

对于单纯的链接块，要区分是导航链接、相关链接，或是噪音链接。如果一个链接块中所有链接的锚文本集合与该网页主题相关，那么该链接块被称为相关链接块，其链接指向的网页一般与当前网页的主题是相关的，而且它的空间位置一般是紧随着网页的主要信息块；如果一个链接块中的链接只是为了方便读者浏览和网站组织，则该链接块被称为导航链接块，其链接指向的网页可能与当前网页的主题相同，也可能不同，导航链接一般位置相对较偏，处于网页的上方或者左右两侧；如果一个链接块中链接指向的网页与当前网页主题根本没有关系，则该链接块是噪音链接块，其在正式的Web学术资源网页上较少。

从内容块识别的过程看出，页面主要信息包含主题型分块和一些小信息块，这些小信息块需要合并。同时，标题块中很可能提供了整个内容块的主题，但由于字数特别少而难以把握具体关键字。另外，对一些明显是噪音的分块则还要进行清理。

内容块的合并过程如下。

步骤1：首先进入起始状态，读取栈顶的分块类型。

步骤2：如是标题块则是一个大信息块的起始，且之前块已结束，则先将页面块临时存放区的内容块输出（输出合并以后的信息块），然后将新标题块放入页面临时存放区，之后回到读取栈顶数据的初始状态。

步骤3：如是主题型，即这一块是信息的主要块，则将它与页面临时存放区的内容合并后，回到初始状态重新读取标识；如是小信息块执行同样操作，则将临时存放区内的内容块和当前块合并，之后回到初始状态。

步骤4：如是链接和图片型，则根据不同空间位置分别处理，即广告块和版权块是噪音块；如是单纯的链接块，则区分是导航链接还是相关链接，或者是噪音链接；导航链接和相关链接块放入页面临时存放区，噪音块删除，之后回到初始状态。

步骤5：当栈为空时，将临时存放区的内容作为最后一个内容块输出，进入结束状态。

经过划分与合并以后的网页，变成一组信息块的集合。这样的信息块不仅主题单一，区分了文本内容信息块和链接信息块，而且最大限度地去除了页面中噪音内容，能更好地指导主题爬虫在主题块中跟踪新的链接。

4 实验分析与结论

实验从分块过滤后的效果和主题爬行两个方面进行，借用网易和新浪等门户网站以及北大天网提供的大规模中文Web测试集CWT200G数据。为比较不同网页分块算法的性能，本文实现了几种前人分块的方法，包括Y.Chen的YPM方法和Gupta的GDOM方法以及一种基准方法NOM（去除HTML不可见标签并保留所有文本结点及锚结点内容）。定义评测标准如下：

噪音文本去除率=（正确去除的噪音文本长度/总的噪音文本长度）×100%；

噪音链接去除率=（正确去除的噪音链接数/总的噪音链接数）×100%；

表1　分块识别和合并后的输出结果

方法、指标	噪音文本去除率	噪音链接去除率	内容提取率	链接提取率	内容误去率	链接误去率
NOM	0%	0%	100%	100%	0%	0%
YPM	80.2%	78.3%	92.3%	91.7%	7.7%	8.3%
GDOM	88.3%	77.7%	93.7%	94.1%	6.3%	5.9%
SPCOLA	91.5%	93.1%	96.3%	97.1%	3.7%	2.9%

内容提取率=（正确提取的主题文本长度/总的主题文本长度）×100%；

链接提取率=（正确提取的相关链接数/总的相关链接数）×100%；

内容误去率=（错误去除的主题文本长度/总的主题文本长度）×100%；

链接误去率=（错误去除的相关链接数/总的相关链接数）×100%。

通过随机抽取1000张网页，采用人工观察其分块识别和合并后的内容得到输出结果（表1）。

可见，NOM方法的内容提取率和链接提取率是最高的，但它的噪音去除率也是最低的，这主要是由于该方法提取了网页中所有的文本信息和链接信息；YPM的评测指标不理想，主要因为它只是粗略对网页进行划分，不能很好地区分主题和噪音信息，所以内容和链接的误去率较高；GDOM优于YPM，它的缺点在于只是简单地利用链接树和文字数的比例来识别链接块，因此噪音链接的去除率较低。本文提出的方法对大多Web学术资源网页的分块与识别效果比较理想，产生一些结果不理想的网页的原因是：一些网页大量使用动态脚本语言生成页面框架，并不遵从Web标准规范；本文算法预设的一些阈值对于有些网页不是最佳；有些网页的信息分块方式比较特殊，如用一个空白图片条放在两个块之间未能检测出这些方式。

以上内容证明了本文提出的综合网页空间、内容和标签信息特征分块方法具有先进性和一定的普适性。将该方法应用于主题爬行，加入分块算法不但过滤掉了大量的噪音信息和噪音链接，而且以网页中的分块为最小处理单元，如果网页中的内容块与主题相关则将内容块中的链接抽取出来，否则将该内容块丢弃。这种方法通过对网页的主题文本和相关链接块的准确识别、提取，能更准确地计算网页的相关度，更有效地发现相关链接，从而能显著提高主题爬行算法的性能。

实现数字图书馆系统中Web学术资源信息的优化采集，有效地对网页进行分块，对后续研究起着重要作用。实现分块算法后，将对每个网页进行分块并保存块信息到数据库，为基于分块做链接分析和主题判定奠定基础。

参考文献

[1] 张先祥.中国互联网行业发展综述[C]//四川省通信学会2006年学术年会论文集.成都,2006:415−417.

[2] Menczer F. Complementing Search Engines with Online Web Mining Agents[J].Decision Support Systems, 2003, 35(2):195–212.

[3] Gupta S, Kaiser G, Neistadt D, et al. DOM Based Content Extraction of HTML Documents[C]//The Proceedings of the 12th World Wide Web Conference (WWW 2003). Budapest, Hungary, May 2003.

[4] Chen Y, Xie X, Ma W Y, et al. Adapting Web Pages for Small−screen Devices[J]. Internet Computing, IEEE, 2005, 9(1):50−56.

[5] 于满泉,陈铁睿,许洪波.基于分块的网页信息解析器的研究与设计[J].计算机应用,2005(4):974−976.

[6] 徐薇.Web信息采集中页面分块技术的研究[J].武汉科技学院学报,2007(5):43−45.