基于DOM模型扩展的Web信息提取

顾韵华 田 伟
(南京信息工程大学计算机与软件学院 南京 210044)

摘 要 提出了一种基于DOM模型扩展的Web信息提取方法。将Web页面表示为DOM树结构,对DOM树结点进行语义扩展并计算其影响度因子,依据结点的影响度因子进行剪枝,进而提取Web页面信息内容。该方法不要求对网页的结构有预先认识,具有自动和通用的特点。提取结果除可以直接用于Web浏览外,还可用于互联网数据挖掘、基于主题的搜索引擎等应用中。

关键词 文档对象模型,Web信息提取,影响度因子,DOM树扩展
中图法分类号 TP309.2 **文献标识码** A

Extraction of Information from Web Pages Based on Extended DOM Tree

GU Yun-hua TIAN Wei
(School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract A method of information extraction from Web pages was presented, and it is based on extended DOM tree. Web pages were firstly transformed to DOM tree, then the DOM tree was extended by adding semantic expression to node and influence degree was calculated for each node. According to influence degree of nodes, the DOM tree was pruned, and it can automatically extract the useful relevant content from Web pages. This approach is a universal method, which does not require to pre-know the structure of the Web page. The results of the information extraction are used not only for browsing but also for further Web information process, such as internet data mining, topic-based search engine.

Keywords DOM, Extraction of information from Web pages, Influence degree, Extended DOM tree

1 引言

Web已成为当今最庞大的信息库,基于Web的数据挖掘成为人们获取有效信息的重要途径,Web信息提取是当前信息领域的研究热点。但Web页面通常包含很多诸如广告链接、图片等与主题无关的内容,它们对页面主题来说是"噪声";而Web信息通常存在于半结构化的HTML文档中,这些都为Web信息提取带来了很大困难。

国内外对Web信息提取已经进行了大量的研究工作[1-3]。目前常用的方法有:(1)基于网页模板的方法,该方法主要针对"模板"型网页。同一网站的页面通常具有相同结构,网站页面中的相同部分即为"模板"。文献[4]提出了在模板基础上通过启发式规则进行信息提取的方法。这种方法的主要局限性在于需对网页结构有预先认识,缺乏通用性。文献[5]提出了通过机器学习技术形成模板库,从而提高信息提取的自动化程度的技术。虽然可以通过扩充模板库来一定程度地解决通用性问题,但互联网上网页结构千变万化,很难求全,并且模板库的维护开销也会增加。(2)基于页面表示模型的方法。所谓页面表示模型是指网页的结构化模型,它是网页分析的基础。常用页面表示模型有W3C提出的DOM(Document Object Model)和微软亚洲研究院提出的VIPS(VIsion-based Page Segmentation)。文献[6]的方法基于VIPS模型,把页面划分为视觉块,进而提取信息。由于视觉特征的复杂性,划分的启发式规则比较模糊,需用户干预,自动化程度不高。文献[7]的方法基于STU-DOM模型,将HTML转换为STU-DOM树,并引入局部相关度和上下文相关度对STU-DOM树进行结点过滤与剪枝,生成只含主题内容的HTML文档。该方法仅考虑了非链接字数和块内链接语义,有一定的局限性。文献[8]的方法基于后缀树模型,依赖于网页中的重复模式进行剪枝,也未能考虑标签结构及类别对主题的影响。

现有Web信息提取方法在处理中未考虑页面标签类别,以及标签属性对页面主题信息提取的影响,并且所提取的结果仅为原HTML文档的子集,而在提取过程中对文档分析所产生的中间结果,如信息块重要程度等有用信息无法为互联网数据挖掘、基于主题的搜索引擎等进一步的Web信息处理应用提供支持。针对这些问题,本文提出一种基于DOM模型扩展的Web信息提取方法,将Web页面表示为DOM树结构,考虑页面标签类型以及标签属性对主题的影响,对DOM树结点进行语义扩展并计算其影响度因子,依据结点的

到稿日期:2009-01-03 返修日期:2009-09-24 本文受江苏省产业技术研究与开发基金项目(苏发改高技发[2006]1106号)资助。
顾韵华(1965—),女,副教授,CCF高级会员,主要研究方向为信息系统及安全,E-mail:yhgu@nuist.edu.cn;田 伟(1980—),男,讲师,主要研究方向为信息系统及安全。

影响度因子进行剪枝，进而提取 Web 页面信息内容，从而实现 Web 信息的自动提取。与已有方法相比，本方法除返回页面内容外，还同时返回内容的重要度量化参数——影响度因子，这可为互联网数据挖掘、基于主题的搜索引擎等应用提供了有益的参考。

2 DOM 模型及扩展

2.1 文档对象模型 DOM

DOM 即文档对象模型，是 W3C 制定的标准接口规范，是一种处理 HTML 和 XML 文件的标准 API。

DOM 提供了对整个文档的访问模型，将文档作为一个树形结构，树的每个结点表示了一个 HTML 标签或标签内的文本项。DOM 树结构精确地描述了 HTML 文档中标签间的相互关联性。将 HTML 或 XML 文档转化为 DOM 树的过程称为解析(parse)。HTML 文档被解析后，转化为 DOM 树，因此对 HTML 文档的处理可以通过对 DOM 树的操作实现。

DOM 模型不仅描述了文档的结构，还定义了结点对象的行为，利用对象的方法和属性，可以方便地访问、修改、添加和删除 DOM 树的结点和内容。

2.2 DOM 树扩展

根据 W3C 的定义，DOM 树结点的属性包括标记名(nodeName)、结点类型(nodeType，取值为 Tag|Txt)、结点内容(data)、父结点对象集合(parentNode)、子结点对象集合(firstChild，lastChild)、兄弟结点对象集合(previousSibling，nextSibling)等。DOM 树结点的这些属性给出了页面的基本内容和结构信息，但不能反映标签、属性以及内容等与主题的相关程度，因而缺乏主题提取所需的语义。

对 DOM 树扩展的总体思路为：考虑 HTML 页面标签的类别，以及标签属性值对页面主题信息的影响，将这种影响纳入对页面内容要素的计算中，对 DOM 树结点进行语义扩展，同时引入结点影响度因子来刻画该结点在树中的重要程度。

2.2.1 DOM 树结点语义扩展

为了增加 DOM 树结点与页面主题信息相关程度的语义信息，计算结点内容的重要度，将 HTML 标签的类别(Category)、非链接文字数(WordNum)、超链接数(LinkNum)、属性集(Attibution)和影响度因子(Influence)等属性添加到结点中，扩展其语义。HTML 标签依据其作用可分为 5 类：(1)描述标题及页面概要信息的标签：如〈title〉、〈meta〉等。(2)规划网页布局的标签：如〈table〉、〈tr〉、〈td〉、〈p〉、〈div〉等，其作用是描述网页内容的布局结构。(3)描述显示特点的标签：如〈b〉、〈I〉、〈strong〉、〈h1〉—〈h6〉等，其作用是强调重点内容，引起人们注意。(4)超链接相关的标签，表示网页间的内容相关性信息。(5)其他标签，如设置图像的标签〈img〉，在文本提取时将忽略这类标签。

根据 HTML 标签在刻画网页特征时的语义功能，将 DOM 树结点分为 6 种类别：标题类(TITLE)、正文类(CONTENT)、视觉类(VISION)、分块类(BLOCK)、超链类(LINK)和其他类(OTHER)，不同类的结点对 Web 信息提取的重要度不同。

(1)标题类(TITLE)：指 HTML 文档中标题标签的专有类别。

(2)正文类(CONTENT)：指包含网页正文内容的标签类别，如包含文字的〈td〉标签。

(3)视觉类(VISION)：指描述页面显示特性的标签类别，如〈b〉、〈strong〉等。

(4)分块类(BLOCK)：指用于网页内容分块的标签类别，如〈table〉、〈tr〉等。

(5)超链类(LINK)：指包含超链接的标签类别，如〈a〉。

(6)其他类(OTHER)：指不属于以上 5 种类别的标签类型。

以上 6 类结点对页面主题的重要度依次降低。扩展后的 DOM 树结点结构如图 1 所示。

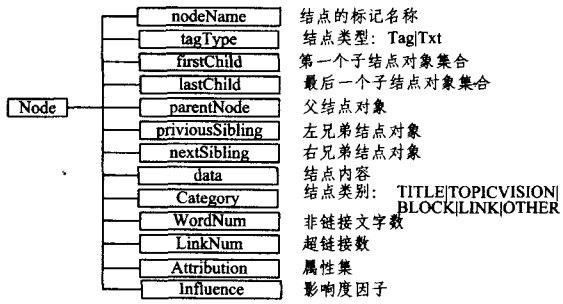

图 1 扩展的 DOM 树结点结构

2.2.2 结点影响度因子

Web 页面的有效内容大多存在 DOM 树的叶结点中，DOM 树中的其余结点主要用于表示内容分块及页面的外观特性。在已有的页面信息提取方法中，对这些结点往往只考虑内容分块作用，而忽略了视觉结点对页面内容的影响。实际上，网页设计者通常会利用显示标签以及标签属性强调重点内容，不妨称其为强调标签和标签强调属性，例如〈b〉标签，或〈font〉标签的 size 属性。此外，不同类别结点对其子孙结点内容块的影响也是不同的。例如，以标题类结点为祖先结点的内容块，其重要程度应更高。

为了评判 DOM 树中结点对内容的影响程度，定义了结点影响度因子。

定义 1(DOM 树结点影响度因子) 表示结点对内容影响的相对程度，用 $Influence(node)$ 表示，$Influence(node) \in [0,1]$。该值越大，表明影响程度越高。

结点影响度因子的确定要综合考虑结点类别和标签强调属性，其初值按 TITLE，CONTENT，VISION，BLOCK，LINK，OTHER 类别降序排列。可构造影响度因子初值向量 Init_vlale。同时结点影响度因子具有传递性，即某结点的影响度因子值应向其子结点传递。因此，叶结点的影响度因子可由下式计算：

$$Influence(leaf) = \sum_{i=1}^{k} Influence(Ancestor_i)$$

其中，$Ancestor_i$ 是叶结点的祖先结点，k 为祖先结点数。

3 Web 信息提取系统结构及相关算法

3.1 Web 信息提取系统结构

本文提出的 Web 信息提取系统框架如图 2 所示，分为 4

个部分：HTML 解析器、DOM 树扩展器、DOM 剪枝器、Web 信息提取器。

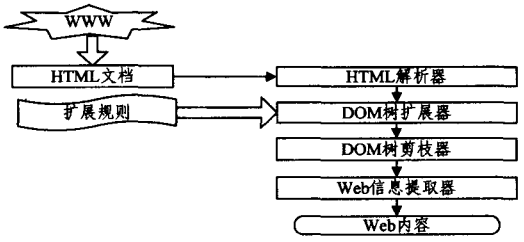

图 2　Web 信息提取系统框架

HTML 解析器(HTML Parser)将 HTML 文档转换为 DOM 树，本系统采用 HTMLParser。HTMLParser 是在 sourceForge. net 上的一个开源项目，是一个对 HTML 进行分析的快速实时的解析器。DOM 树扩展器(DOM Expander)根据树扩展算法对 DOM 进行结点类型标识，并计算结点影响度因子。DOM 树剪枝器(DOM Pruner)根据剪枝算法除去 DOM 树中与主题无关的结点。最后 Web 信息提取器对 DOM 树保留的结点进行遍历，输出 Web 内容及其影响度因子。

3.2　DOM 树扩展算法

DOM 树扩展的过程分为两个阶段：第一阶段，遍历由 HTML 解析器生成的 DOM 树，对每个结点标识其类型，添加有效标签属性，并按影响度因子初值向量 Init_vlale 对结点影响度因子初始化；第二阶段，计算结点影响度因子，构造包含内容的叶结点向量 LEAFS，对 LEAFS 中每个叶结点自下向上查找其祖先结点，累加其祖先结点的影响度因子。

算法 1　DOM 树扩展算法

```
DOMExpand(T)
  for each node∈T do
    node. Category=
      {TITLE|TOPICVISION|BLOCK
      |LINK|OTHER}
      node. Attribution=[name,value]
      node. Influence=Init_value(node)
    endfor
    construct(LEAFS)
  for each leafnode∈LEAFS do
      计算其祖先结点 Influence 累加和 sum
      leafnode. Influence=sum
    endfor
end // DOMExpand
```

3.3　剪枝算法

由于对结点计算了影响度因子，因此可设定被剪枝结点的影响度因子阈值。

定义 2(结点影响度因子阈值 λ)　如果 $Influence(node_i)\geqslant\lambda$，则称结点 $node_i$ 与主题相关，否则为无关。

剪枝器对 DOM 树中结点进行剪枝的基本原则是判断结点是否为主题无关结点，即影响度因子小于 λ 的结点应该被剪枝。

算法 2　DOM 树剪枝算法

```
DOMPruner (T)
  for each node∈T do
    if node. Influence<λ
      delete(node)
    endif
  end for
end // DOMPruner
```

根据剪枝算法，可以删除 DOM 树中与主题无关的链接列表和没有内容的块。

4　实验与分析

本文所提出的方法是针对某单一 Web 页面，剔除与主题信息不相关内容，得到该页的主要信息。该类 Web 信息提取方法可用完整性和压缩率来衡量提取工作的效果。

定义 3(完整性)　指主题内容完整的结果网页数占来源网页数的百分比。

定义 4(压缩率)　指结果网页的文件大小占来源网页文件大小的百分比。

对本文的提取方法进行了测试，实验测试对象包括新闻、体育、娱乐和电子商务多个领域，结构差别大，既有主题内容分散和结构繁杂的网页(如娱乐型网页)，又有主题内容集中的网页(如新闻型网页)，这有助于验证算法性能。表 1 显示了测试结果。其中完整性是将原始网页与提取后的网页进行人工分析的结果。

由表 1 所示的实验结果可见，本方法平均完整性为 90.5%，能够较完整地保留主题内容；平均压缩率为 27.2%，信息提取效果明显。

表 1　实验结果

网页类型	数量	完整性	压缩率
新闻	30	96.2%	28.1%
体育	18	86.3%	26.8%
娱乐	36	88.5%	25.2%
电子商务	22	91.1%	28.7%

经实验分析发现，剪枝结点的阈值 λ 对提取结果有重要影响，λ 值选取不当可能导致提取效果不理想，如删除主题链接或保留无关链接。可根据实际应用调整 λ 值大小，增大 λ 值可以删除更多无关链接，提高主题相关度，但可能会降低完整性。实验中选取 λ 为 $\sqrt{avg(node)}$，这是经过实验得到的值。

结束语　本文提出了一种基于 DOM 模型扩展的 Web 信息提取方法：通过对 DOM 树进行语义扩展并计算结点的影响度因子，再依据结点的影响度因子进行剪枝，进而提取 Web 页面信息内容。实验表明，本方法有较高的完整性和压缩率。该方法还可进一步改进，如针对不同类型的网页采用不同的结点影响度因子阈值进行剪枝。

与已有方法相比，本文提出的方法具有自动和通用的特点，不要求对网页的结构有预先认识，同时本方法适用面广，所提取的 Web 信息内容除可以直接用于 Web 浏览外，还可用于互联网数据挖掘、基于主题的搜索引擎等应用中。

参 考 文 献

[1]　Freitag D. Machine learning for information extraction in information domains[J]. Machine Learning, 2000, 39(2/3): 169-202

[2]　Gupta, Kaiser G, Neistadt D, et al. DOM-based content extraction of HTML documents[C]// Proc. of the 12th Int'l World Wide Web Conf. New York: ACM Press, 2003: 207-214

(下转第 289 页)

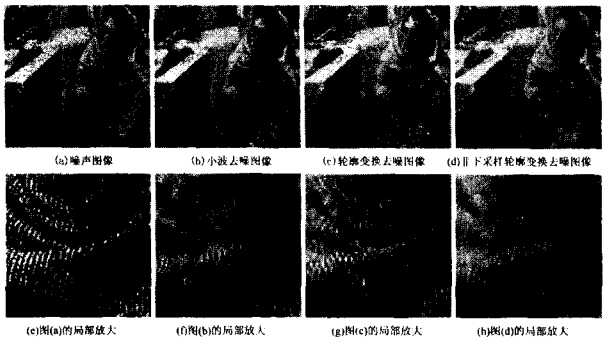

图5 标准灰度图像 Barbara 的去噪(噪声 $\sigma=40$)结果及局部放大

表1 含噪声图像与去噪声图像的峰值信噪比对照(3种方法)

图像	噪声标准差 σ	PSNR/dB			
		噪声图像	小波去噪	轮廓变换去噪	非下采样轮廓变换去噪
Lena	20	22.18	28.13	29.41	30.94
	30	18.17	26.84	27.48	29.33
	40	16.37	25.97	26.52	27.72
	50	14.60	25.27	25.49	26.30
Barbara	20	22.22	23.92	24.73	25.51
	30	18.82	22.38	23.32	24.78
	40	16.50	21.40	22.58	23.79
	50	14.80	20.07	21.14	22.60
Mandrill	20	22.17	22.57	23.04	23.85
	30	28.67	20.53	20.99	21.22
	40	16.30	20.22	20.57	20.76
	50	14.55	19.96	20.23	20.35

分析对比上述实验结果,不难得出如下结论:

(1)本文算法对平滑图像(如 Lena)和边缘图像(如 Barbara)的去噪效果良好,明显优于小波域去噪方法和轮廓变换域去噪方法。例如,对于标准灰度图像 Lena 和 Barbara,本文算法(即非下采样轮廓变换域去噪方法)的峰值信噪比(PSNR)比小波域去噪高 2～2.5dB,比轮廓变换域去噪高 1～1.5dB。

(2)本文算法对纹理图像(如 Mandrill)的去噪效果一般,但仍好于小波域去噪方法和轮廓变换域去噪方法。这是因为非下采样轮廓变换所提取的图像方向性是有限的,并不能提取任意方向性,以至于对部分复杂纹理无法获得很好的去噪声效果。

(3)本文算法能够很好恢复图像的细节轮廓信息(如 Lena 帽檐、Barbara 裤子纹理等),较好地克服了小波去噪、轮廓变换去噪伪吉布斯(Gibbs)现象所带来的视觉失真。

结束语 图像去噪是整个图像处理过程的关键步骤之一。本文以同时具有平移不变性、频率选择性、正则性等优良性能的非下采样轮廓变换理论为基础,提出了一种新的图像去噪方法。方法首先对图像进行非下采样轮廓变换,以得到不同尺度、不同方向上的变换系数;然后结合噪声分布特点确定多尺度阈值,并依此阈值对高频系数进行去噪处理;最后对去噪处理后的变换系数进行反变换,以得到去噪图像。实验结果表明,本方法不仅拥有较强的抑制噪声的能力,而且具有较好的边缘保护能力,同时能够消除图像边缘附近的伪吉布斯(Gibbs)现象。

参考文献

[1] Mallat S, Hwang W L. Singularity detection and processing with wavelets[J]. IEEE Trans. on Information Theory, 1992, 38(2): 617-643

[2] Xu Y, Weaver J, Healy M. Wavelet transform domain filters: A spatially selective noise filtration technique[J]. IEEE Trans. on Image Processing, 1994, 3(6): 747-758

[3] Donoho D L, Johnstone I M. Ideal spatial adaptation via wavelet shrinkage[J]. Biometrika, 1994, 81(3): 425-455

[4] Gao H, Bruce A. WaveShrink with firm shrinkage[J]. Statistics. Sinica, July 1997: 855-874

[5] Chang S G, Yu B, Vetterli M. Adaptive wavelet thresholding for image denoising and compression[J]. IEEE Trans. on Image Processing, 2000, 9(9): 1532-1546

[6] Chen G Y, Bui T D, Krzyzak A. Multiwavelets image denoising using neighboring coefficients[J]. IEEE Trans. on Image Processing Letters, 2003, 10(7): 211-214

[7] Stein C M. Estimation of the mean of a multivariate normal distribution[J]. Annals of Statistics, 1981, 9(6): 1135-1151

[8] Huang H C, Lee C M T. Data adaptive median filters for signal and image denoising using a generalized SURE criterion[J]. IEEE Trans. on Image Processing Letters, 2006, 13(9): 561-564

[9] 曲天书,戴逸松,王树勋. 基于 SURE 无偏估计的自适应小波阈值去噪[J]. 电子学报, 2002, 30(2): 266-268

[10] Luisier F, Blu T, Unser M. Sure-based wavelet thresholding integrating inter-scale dependencies[C]// Proceedings of the 2006 IEEE International Conference on Image Processing (ICIP'06). Atlanta GA, USA, October 2006: 1457-1460

[11] Do M N, Vetterli M. The finite ridgelet transform for image representation[J]. IEEE Trans Image Processing, 2003, 12(1): 16-28

[12] Starck J L, Candes E J, Donoho D L. The curvelet transform for image de-noising[J]. IEEE Trans. on Image Processing, 2002, 6(11): 670-684

[13] Do M N, Vetterli M. The Contourlet transform: an efficient directional multiresolution image representation[J]. IEEE Trans. on Image Processing, 2005, 14(12): 2091-2106

[14] Do M N, Vetterli M. Framing pyramids[J]. IEEE Trans. on Signal Processing, 2003, 51(9): 2329-2342

[15] 戴维,于盛林,孙栓. 基于 Contourlet 变换自适应阈值的图像去噪算法[J]. 电子学报, 2007, 35(10): 1939-1943

[16] Arthur L, Cunha D, et al. The nonsubsampled Contourlet transform: theory, design, and applications[J]. IEEE Trans. on Image Processing, 2006, 10(15): 3089-3101

[17] Shensa M J. The discrete wavelet transform: Wedding the átrous and Mallat algorithms[J]. IEEE Trans. on Signal Processing, 1992, 40(1): 2464-2482

[18] Bamberger R H, Smith M J. A filter bank for the directional decomposition of image: theory and design[J]. IEEE Trans. on Signal Proceeding, 1992, 40(4): 882-893

[19] Donoho D L. De-noising by soft-thresholding[J]. IEEE Trans. on Information Theory, 1995, 41(3): 613-627

(上接第 237 页)

[3] Gupta S, Kaiser G E, Grimm P, et al. Automating Content Extraction of HTML Documents [J]. World Wide Web Journal

[4] 张志刚,陈静,李晓明. 一种 HTML 网页净化方法[J]. 情报学报, 2004(4): 387-393

[5] 欧健文,董守斌,蔡斌. 模板化网页主题信息的提取方法[J]. 清华大学学报:自然科学版, 2005, 45(1): 1743-1747

[6] Deng C, Yu S P, Wen J R, et al. VIPS: a Vision Based Page Segmentation algorithm[R]. MSR-TR-2003-79. 2003

[7] 王琦,唐世渭,杨冬青,等. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展, 2004, 41(10): 1786-1791

[8] 高强,张敬之,耿桦,等. 基于重复模式的信息抽取[J]. 计算机科学, 2007, 34(4): 210-212

[9] 冯艳为,王成良. 基于 Web 部件的个性化网站创建技术. 重庆工学院学报:自然科学版, 2008, 22(2): 121-126