2017年第17届IEEE通信技术国际会议

**基于DOM树和统计的Web内容信息提取信息**

***蓬勃发展的抽象web页面包含大量信息，而它们包含的内容很少，并且包含很多不相关的噪声信息，比如脚本代码、链接、广告等等。*这些不相关的噪声信息占用了大量的空间，不适合过渡到小型移动设备、数据挖掘和信息检索。 因此，web信息提取技术变得越来越重要。 然而，大多数提取方法不能适应各种异构的web结构，具有较差的通用性和提取效率。 本文提出了一种能够适应网页异构性和多变性的方法，具有较高的精度和查全率。 我们的方法是基于DOM结构，将一个web页面划分为几个块，用统计信息提取内容块，而不是机器学习重复训练和手工标注，这在精度、回忆和F1上都有很好的表现。**

***关键词内容信息提取; DOM结构; 内容块; 统计信息***

介绍。

随着网络的发展，大量的信息以非常快的速度占据了网页。 正如我们所看到的，网络已经成为世界上最大的数据源之一，其中包含了大量有价值的信息。 同时,无关紧要的信息,如广告和网站导航越来越甚至占据了一半的网页,这使得它很难定位和提取内容信息,和web页面的格式不同在成千上万的方面,并不断变化,如何保证萃取的多功能性和适应性技术是一个巨大的挑战。

Web信息提取技术主要包括基于包装器的提取技术、DOM树结构、页面分割和统计信息。 基于包装器的技术需要大量的训练样本和手工标注，当网页结构发生变化时，应标注新的训练集和手工标注，使自动化程度不高。 基于页面分割的内容提取技术旨在将web页面划分为几个块[1-6]。 现在的页面分割技术主要基于视觉信息[1,2]和特定的标签[3-5]。 基于视觉信息的方法需要下载和解析大量的样式

Zhengping金

国家联网与交换重点实验室

技术

北京邮电大学

中国,北京电子邮件:zhpjin@bupt.edu.cn

信息，这是一种耗费时间的行为，视觉启发式规则的使用比较模糊，人们需要继续总结规则。 基于特定标签的方法不能在所有网页上使用，通用性较差。 基于包装器的提取主要是提取特定或类似的网页[7-9]。 他们需要大量的训练样本和手工标注，当网页结构发生变化时，应该标注新的训练集和手工标注，使自动化程度不高。 基于统计信息的内容提取技术是对web页面中列出的节点中包含的信息进行计数，提取内容信息。 Debnath et al.[10]选取具有一定特征的文本、图像和脚本代码的内容块，总结出FE (Feature Extractor)算法。 Gottron等人提出了CCB (Content Code blur)算法，即通过内容块选择相同格式的源代码字符。 Weninger等人[12]提出了CETR算法(通过标记比率提取内容)，该算法采用HTML文档行的文本标签比率和相邻行的变化率。 也有一些方法以文本密度为测量内容特征的基础[13-16]。 Song等[15]提出的文本密度方法CETD(以文本密度提取)，将文本的数量和标签的数量作为文本的密度，但只将文本和链接文本信息作为度量，忽略包含主题特征的层次结构。

为了适应web页面的异质性和可变性，提高提取的准确性和通用性，本文提出了一种基于DOM结构和统计信息的CEDS方法，以保证提取的通用性和有效性。 首先，我们根据各元素之间的相关和相似度的DOM结构关系，将网页划分为几个块，然后根据主题特征统计信息区分信息块，去除噪声信息块，离开内容信息块。 最后，我们基于一些结构信息去除内容信息块中的噪声信息。 我们的方法召回率为97.90%，精确度为98.54%，F1值为98.21%，高于现有方法。

978 - 1 - 5090 - 5090 - 1308 9/17 /©2017年IEEE 31.00美元

二世。 方法

目前，新闻网页分为两类:专题新闻页面和图片新闻页面。 主题新闻页面有大量的文字，只有少量的图片作为辅助作用，相反，在图片新闻页面中，图片作为主要显示，文字只说明图片的描述。 在本文中，我们的目标是主题新闻页面。

*a . DOM*

文档对象模型(Document Object Model, DOM)是一个与语言无关的平台和接口，它允许程序和脚本动态地访问和更新文档的内容、结构和样式，这是W3C组织[17]推荐的。

*b页面分割*

我们将参考网页的DOM结构规则，将页面划分为几个信息块。 在分割操作开始之前，我们需要做一些预处理工作。

*1)为每个节点添加分层属性，我们称之为level。*层次信息可以丰富节点结构信息，与新闻网页的内容信息有很多关系。 我们可以在图1中看到分层信息。 DOM结构越深，节点级别就越大。

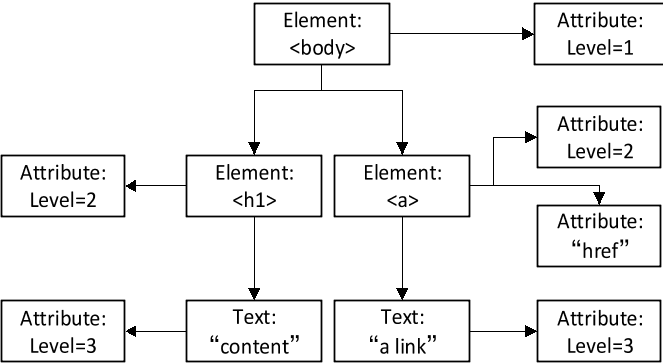


图1所示。 具有级别的DOM树示例。

*2)遍历DOM树，得到叶子节点作为基本对象集，根据图1中的DOM树结构，叶子节点是页面的显示原子对象，文本信息都存储在叶子节点上。*

*3)获取提取的真实节点。 简单的叶子节点有时不能完全反映DOM结构。*例如，叶子节点被封装在一些修饰层中，它的父节点只有一个子节点，这意味着这些修饰只用于修饰叶子节点。 因此，这个父节点可以表示这个叶节点。 如图2所示，

· 节点可以表示文本叶子节点。 我们将该方法称为GetRealNode(node)，我们可以在算法1中看到它。

· 1。 < ul水平=“17”>

· 2。 <李水平=“18”> <水平= " 19 " > Bagehot的笔记本< / > < /李>

· 3所示。 <li  level="18"><a level="19">Buttonwood的notebook

· < ! ——省略与上面结构相同的节点——>

5。 <李水平=“18”> <水平= " 19 " >演讲角< / > < /李>

6。 <li level="18"><a level="19">经济学家解释

< / ul >

图2。 一段简短的HTML代码

*4)基于最少共同祖先的融合。 我们发现，内容在DOM结构中具有相似性，因此结构相似的节点应该分配给相同的块。*如图2所示，

· 节点在结构上相似，并且在内容级别上相互链接，因此它们应该被分配到一个块中。

· **算法1:fusebycommonancestry (leafNodeList, root)**

· /\*isCommonLevel (N1, N2)是确定N1和N2是否具有相同的级别，如果是，则返回true，否则返回false\*/

· /\* isBrother(N1, N2)是用来确定N1和N2是否是滑动节点，如果是，返回true，则返回false\*/

· /\* getLCA (N1, N2)是为N1和N2\*/获取最不常见的祖先节点

· 1。 输入:LeafNodeList和Root Node 2。 输出:节点块

· 3所示。 开始

· 4所示。 对于每一个相邻的叶节点N1 N2，在叶子中做5。 N1←GetRealNode(N1)

· 6。 N2←GetRealNode(N2)

· 7所示。 isCommonLevel←isCommonLevel(N1、N2)8。 isBrother←isBrother(N1、N2)

· 9。 父母←getLCA(N1、N2)

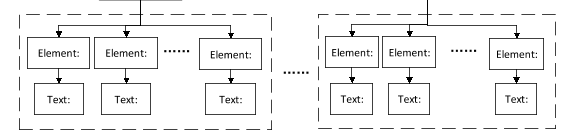
· 10。 如果是普通和&是兄弟和父母!=根11。 addIntoBlock(父)

· 12。 结束13。 结束

· 14。 结束

· DOM结构上的叶节点在页面显示,视觉连续性和与内容相关的节点在DOM结构相似,如节点级别相同,兄弟节点关系,内面有共同的祖先节点,如果相邻叶节点满足上述三个条件,他们被融合成一块。 详情请参见算法1。

|  |  |  |  |
| --- | --- | --- | --- |
| · | 元素: |  |  |
|  |  |  |  |
|  |  | 文本: |  |
| 文本: |  |  |  |
|  | 元素: |  | 元素: |

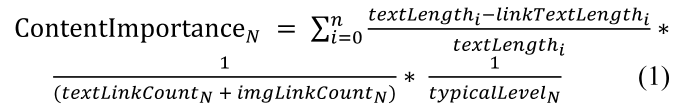
· 

· 图3。 页面分割后的DOM树

· 在上述操作之后，DOM结构中的叶子节点被划分为一些信息块，这些信息块被图3中的虚线分割。 研究对象已从许多叶节点变为更少的中间节点。

· *c统计信息*

· 我们选取新浪新闻网(http://news.sina.com.cn/)上的一篇文章为例，使用上面的方法将文章分成块。 通过大量的案例研究，我们发现内容信息与噪声信息之间存在一定的差异。 内容有更多的纯文本，更少的超链接和更少的修改，噪音有更少的纯文本，更多的超链接和更多的修改。

· 

i:块中每个节点的顺序

ℎ:节点中的文本的长度

ℎ:超链接文本的长度

节点。

:超链接图片的数量

我的节点

:可以表示

节点N的层次结构信息

ContentImportance

180年

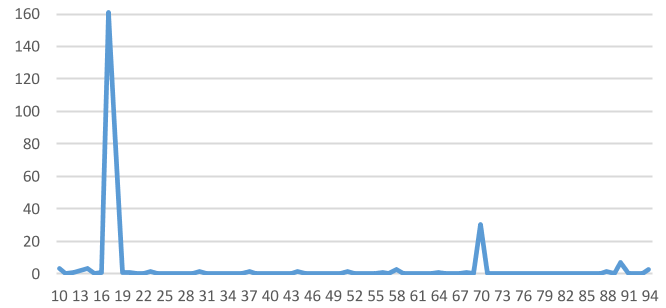


图4。 ContentImportance的每一块

每个节点都具有上面部分中标记的level属性。 对于每个信息块，它都有几个节点，我们需要获得该块最具代表性的级别。 经过大量的分析，我们发现:节点级别的最大数量是最常见的级别，当有多个节点级别时，我们选择最大的级别。

通过图4可以发现，内容信息和噪声信息之间的差异可以有效地扩展内容信息和内容信息之间的差异。

*d .阈值*

本文中阈值的目的是将信息块划分为两个类。 方差是描述类之间差异的指标

方差越大，类之间的差异越大。 经过以上分析，我们以方差为基本原则。

∑

  ∗

∑(−)

=最大

本文的目的是将信息块划分为两个类。 方差是描述类之间差异的指标，方差越大，类之间的差异就越大。 经过以上分析，我们以方差为基本原则。

我们发现τ以迭代的方式,使最大的两个类之间的方差。 方程2-4表示平均值、方差和阈值。 N是类的数量,是每个少女ContentImportance值,μ是平均值,方差,τ是阈值,阈值系数ε,分析之后,范围是[1,3],迭代步骤是0.1,标准差是最大化的阈值迭代法。 公式如下。

获得阈值后，确定大于阈值的信息块为内容块，其余为噪声块。

*E.从内容块中去噪*

阈值选择后的主题信息块也可能包含一些噪声信息，比如在新闻内容信息块中会捕获一些注释。 经过大量的分析，得到了去噪算法。 如果相邻节点之间的距离比四大,相邻叶节点的祖先节点没有根的内容信息块和这些叶节点的水平并不是常见的水平,这些节点噪声存在的节点内容的信息块。

三世。 评价

本文使用的数据集来自最近通过爬虫类随机获取的主题新闻页面集合。 通过手工获取比较和评价的依据。

=

(,)。

                        (6)

*R = LCS A, B，长度*

*B.length*

                        (7)

F1 =                                                                   (8)

我们的内容抽取系统的性能基于三个评价指标:召回率、精度和F1。 回忆等于系统的正确结果与所有可能的正确结果的比率。 精度等于系统正确提取结果与提取结果之和的比值。 F1用于平均召回率和精度性能。

在方程6 - 8,是内容的序列信息通过我们使用本文的方法,B是通过手动方式获得的内容信息,LCS(A,B)是最长公共子序列的两个信息集,我们可以看到实验结果表中的我。

表一实验结果(P为精度，R为召回)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据源 | 网站 | P | R | F1 |
| 《经济学人》 | http://www.econo mist.com/ | 98.12% | 96.34% | 96.98% |
| 新浪新闻 | http://news.sina。c om.cn/ | 99.29% | 98.16% | 98.49% |
| 全球新闻 | http://www.huanq iu.com/ | 97.33% | 98.41% | 97.49% |
| 搜狐新闻 | http://www.sohu。com/ | 96.58% | 99.12% | 97.47% |
| Firefox的新闻 | http://www.firefo xchina.cn/ | 97.17% | 99.82% | 97.98% |

在本文中，我们将现有的方法与表二中名为CEDS的方法进行了比较。

表二。 实验结果与其他方法比较

|  |  |  |  |
| --- | --- | --- | --- |
| 方法 | 精度 | 回忆 | F1 |
| 菲 | 74.19% | 60.21% | 66.47% |
| 建行 | 81.27% | 90.34% | 85.57% |
| CETR | 86.98% | 93.70% | 90.22% |
| CETD | 94.49% | 94.58% | 94.53% |
| 清洁能源 | 97.90% | 98.54% | 98.21% |

通过对比，我们的方法召回率为97.90%，精确度为98.54%，F1值为98.21%，高于其他方法，证明了其有效性。

本文旨在适应网页的异质性和可变性，提高提取的准确性和通用性。 我们基于DOM结

四。结论

构将一个原始web页面划分为多个信息块，并利用统计原理提取内容信息块。 最后，我们从内容块中去噪以完成提取工作。 我们的方法不需要人工标注，没有机器学习重复训练，我们的方法证明了它在精度、回忆和F1值上都有很好的效果。

承认

这项工作得到了国家自然科学基金委(批准号61502044)和中央大学基础研究基金(批准号2015RC23)的支持。

参考文献

蔡德，余s，文杰，等，VIPS:基于视觉的页面分割算法。 技术报告，微软技术报告，MSR-TR-2003-79, 2003。

许士杰等。利用格式塔规则对网页中的语义块进行分组识别[J].北京:科技大学出版社。 万维网，2015:1-22。

发现web文档的信息性内容块。 :学报SIGKDD 02。 纽约，纽约，美国，2002:588-593。

b0 Carey H J, Manic M. HTML网页内容提取使用段落标签[C]//工业电子(ISIE)， 2016 IEEE第25届国际研讨会。 IEEE,2016:1099 - 1105。

b0 Shen W, Zou X.一种基于DOM结构的Web文自动提取算法[J]。 《混合信息技术国际期刊》，2015,8(3):243-254。

b0 Bing L, Guo R, Lam W，等。基于结构化预测的网页分割及其在网页分类中的应用[C]//第37届国际ACM SIGIR信息检索研究与开发会议论文集。 ACM,2014:767 - 776。

b0 Annam M, Sajeev G P. Entropy based informationcontent density approach for efficient web content extraction[C]//计算、通信和信息学(ICACCI)的进展，2016国际会议。 IEEE,2016:118 - 124。

[8] Gibson J, Wellner B, Lubar S. Adaptive Web页面内容标识/ Proc第九届ACM Web信息与数据管理国际研讨会。 葡萄牙里斯本,2007:105 - 112。

[9]Ziegler C N, Skubacz M.从新闻页面中提取的内容，利用粒子群算法在IEEE / WIC / ACM国际网络智能会议上的语言和结构特点//Proc上进行了优化。 弗里蒙特,美国,2007:242 - 249。

[10]Debnath S, Mitra P, Giles C L.从应用计算研讨会的Webpages / / Proc中自动提取信息块。 圣菲，美国，2005:1722-1726。

内容代码模糊:第19届数据库和专家系统应用国际会议内容提取/ / / Proc的新方法。 意大利都灵,2008:29-33。

温宁格T，徐文华H，韩佳伟。 第19届万维网国际会议通过标签比率/ / Proc提取cetr -内容。 美国罗利,2010:971 - 980。

2008年第19届国际数据库和专家系统应用研讨会，“以文本-标签比率从网络中抽取文本”。 IEEE,2008,pp。。

[14] Sun F, Song D, Liao L. DOM基于文本密度的内容提取。 会议记录:第34届国际会议记录。 中国,北京,2011:245 - 254。

[15] Song D, Sun F, Liao L.一种文本密度和DOM节点视觉重要性的混合内容提取方法[J]。 知识与信息系统，2015,42(1):75-96。

b0 Annam M, Sajeev G P. Entropy based informationcontent density approach for efficient web content extraction[C]//计算、通信和信息学(ICACCI)的进展，2016国际会议。 IEEE,2016:118 - 124。

[17]W3C文档对象模型(2012)网站。 https://www.w3.org/DOM/。