基于网页 DOM 树节点路径相似度的正文抽取

潘心宇[1],陈长福[2],刘　蓉[1],王美清[1]
(1.福州大学 数学与计算机科学学院,福建 福州 350108;2.福建库易信息科技有限责任公司,福建 福州 350000)

摘　要:由于人工抽取网页信息效率低、成本高,因此根据对大量网页结构的观察,提出基于网页文档对象模型 DOM 树节点路径相似度的正文抽取方法。依据同网站下的网页结构相同的特点去除网页噪声得到网页的主题内容,然后结合正文节点在 DOM 树中的路径的相似度抽取正文。通过对不同类型的中文新闻网站上的 1 000 个网页进行实验,结果表明该方法对于 97.6% 的网页都能够去除大部分噪声并保持正文内容的完整性,正文抽取结果有 93.30% 的准确率和 95.59% 的召回率。所提算法对不同类型的网页都有较好的适应性。

关键词:DOM 树;信息抽取;HTML 标签;网页去噪;正文抽取
中图分类号:TP301.6　　**文献标识码**:A　　DOI: 10.19358/j.issn.1674-7720.2016.19.022
引用格式:潘心宇,陈长福,刘蓉,等. 基于网页 DOM 树节点路径相似度的正文抽取[J].微型机与应用,2016,35(19):74-77.

Content extraction based on the similarity of the Web pages' DOM tree nodes path

Pan Xinyu[1], Chen Changfu[2], Liu Rong[1], Wang Meiqing[1]
(1. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China;
2. Fujian Ecallcen Information Technology Company Limited, Fuzhou 350000, China)

Abstract: Due to the problem that the low efficiency and high cost of extracting information by human, according to the observation of large amount of Web pages' structure, the content extraction method based on the similarity of web pages' DOM tree node's path was proposed. It removed noise and got the main body of the Web page as the Web pages in the same website had the same structure, then combined the similarity of the path of content nodes in the DOM tree to extract content. Through the experiments of 1 000 Web pages from different Chinese news Websites, the results show that this method can remove most noise and maintain the integrity of the content for 97.6% of all Web pages, it has 93.30% precision rate and 95.59% recall rate, and it has good adaptability for different types of Web pages.

Key words: DOM tree; information extraction; HTML tag; Web denoising; content extraction

0　引言

随着互联网技术的快速发展,网页成为人们获取信息的重要来源之一。然而,网页上的数据是海量的,单纯依靠人工手段获取网页信息效率较低,因此需要借助软件对网页信息进行全部或部分地自动过滤和分类。目前常用的自动网页信息获取方法是正文内容抽取,该类方法是一种被广泛应用于互联网数据挖掘的技术,它的目标是从互联网庞大的数据中提取有意义的和有价值的信息,可以用于信息搜索、Web 文档分类、数据挖掘、机器翻译、文本摘要等。

常用的正文抽取方法可以分为以下 4 类:(1)传统的归纳总结正文抽取方法:根据一些信息模式,从特定的信息源中提取相关内容[1]。此方法效率较低、需要较多的手动操作,独立性以及适应性较差。(2)基于网页布局[2]和视觉[3-4]的正文抽取:该方法很大程度上依赖于网页的风格或者结构。当涉及到有更复杂的嵌套关系的网页时会出现偏差。(3)基于语义单元[5]或者数据挖掘、机器学习[6]的正文抽取:通过使用分词和文本分类,虽然准确率有所提高,但是解决方案比较复杂。(4)基于统计的正文抽取[7]:该方法简单而且具有更好的通用性,但是较低的精确度限制了它的进一步应用。此外,它不能处理短文本、表格文本以及有较长评论的文本。

FINN A 等[8]提出正文抽取(Body Text Extrac tion, BTE)算法,将网页中的文字和标签作为序列,抽取序列中文字最多和标签最少的连续的内容。PINTO D 等[9]提出文档斜率曲线(Document Slope Curves, DSC)算法,在 FINN 的方法的基础上使用窗口方法实现多正文抽取。MANTRATZIS C 等[10]提出链接定额过滤(Link Quota Filters, LQE)算法,通过网页结构分析,分离正文和导航目录等超链接。DEBNATH S 等[11]提出特征提取器(Feature Extractor, FE)算法,选择包含有一定特征的文本、图像而且重复出现次数较少的内容块。GOTTRON T 等[12]提出正文代码模糊(Content Code Blur-ring, CCB)算法,选择相同格式的长文本作为网页的正文。刘利等[13]提出基于多

特征融合的网页正文信息抽取，从网页的多个特征和设计习惯入手定位正文位置。王利等[14]提出基于内容相似度的正文抽取，根据树节点中文本内容与各级标题的相似度判定小块文本信息的有效性，由此进行网页清洗和正文抽取。

分析网页信息会发现，网页中包含大量与网页主题无关的噪声内容，如广告链接、导航栏、版权信息等。在正文抽取过程中，这些网页噪声会影响抽取效果，因此需要通过去噪方式对网页进行预处理。常用的网页去噪方法有：

YI L 等[15]提出用风格树（Style Tree，ST）来表达网页的结构和内容特征，出现相同特征次数多的部分更有可能是噪声数据。GIBSON D 等[16]提出 Shingle 和模板 Hash 方法。这两种算法的缺点是计算量较大。WANG J Y 等[17]提出的主题数据提取（Data-rich Section Extraction，DSE）算法，该算法通过从上到下比较两棵相同模板的文档对象模型（Document Object Model，DOM）树，去除树中相同的部分，剩下的部分作为网页的主题内容。

根据对现有方法的总结以及对网页特征的分析，本文提出基于 DOM 树节点路径相似度的正文抽取方法，对于不同结构的网页都有较好的适应性，对来源于新浪、网易、搜狐、腾讯等大型门户网站以及多家各类型网站的 1 000 个网页进行了抽取实验，实验结果表明本文方法有较好的抽取准确度。

1 网页去噪

目前，大部分网页的源代码是以超文本标记语言（Hyper Text Markup Language，HTML）的形式存在的。对于同一网站下的不同网页，它们由同一个模板生成，因此这些网页具有相似的结构，而这些网页中相同的部分就是噪声内容，它们与网页所要表达的主题没有关系。本文在 DSE 算法的基础上，首先将与网页无关的标签及相关代码删除，然后通过将某个网页与同一网站下的 2 个或多个网页进行对比去除相同部分，从而达到去除噪声的目的。

1.1 删除无关的标签

网页源代码包含了以不同的标签括起来的各段代码。例如，网页标题和一些修饰性代码主要嵌在标签 < head > 和 < /head > 的内部，网页主题内容包含在 < body > 和 < /body > 标签之间，客户端脚本则包含在 < script > 和 < /script > 标签之间。通过对大量 HTML 文本的研究和分析，发现以下几类标签与网页主题内容的相关性很低，在对比网页之前可以将这部分内容过滤掉以提高后续的对比速度。

< head > 与 < /head > 标签以及它们之间的内容。

< script > < /script > 标签。该标签中内容的主要功能是定义客户端脚本，与网页所要表达的内容关系不大，也可以将其删除，类似地，< noScript > < /noScript > 也可删除。

大部分网页通过层叠样式表（Cascading Style Sheets，CSS）来调整页面的布局，< style > < /style > 标签用于定义 HTML 文档的样式信息，同样可以删除。

注释标签 < ! -- 注释内容 -- > 、< ! 注释内容 > 只是为网站编辑提供说明，并不会在浏览器中显示，也可删除。

在预处理过程中利用正则表达式删除以上噪声代码。正则表达式通过使用单个字符串来描述、匹配一系列符合某个句法规则的网页源代码。符合匹配规则的源代码将被删除。

删除完无关标签后，再删除空白行，这样完成了去噪的第一步。

1.2 通过网页对比去除噪声

网页对比可以通过对比它们的 DOM 树来实现。DOM 是文档中数据和结构的一个树形表示，它定义了表示和修改文档所需的对象、这些对象的行为和属性以及这些对象之间的关系。DOM 实际上是以面向对象方式描述的文档模型。它可以以一种独立于平台和语言的方式访问和修改一个文档的内容和结构。图 1 给出了一个文档的 DOM 树的结构图。

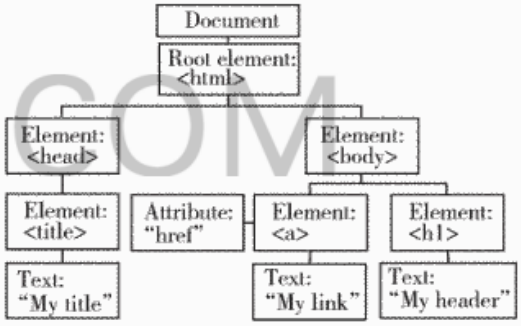

图 1 DOM 树结构图

通过 HTML 解析（如使用解析器 htmlcxx）可以将 HTML 文档转换为 DOM 树结构。假设要处理的是某网站的网页 URL1，随机选取该网站下的另外两个网页 URL2 和 URL3，获得它们的 DOM 树。然后分别对比 DOM1 \DOM2 以及 DOM1 \DOM3，输出不同的节点。

对比算法的基本思路是：按深度遍历 3 棵树的节点，为每个节点设置深度、路径、文本内容、是否为 tag（HTML 标签）。以第 1 个网页作为目标与另外两个网页进行对比，如果 3 个节点深度相同，则判断节点的文本内容是否相同，相同的加入模板集合中，不同的加入网页内容集合中；如果 3 个节点深度不同，则根据不同情况对相应的节点进行处理，其中网页 1 的节点加入到网页内容集合中。直到 3 个网页都遍历到 end 节点为止。最后得到的就是网页 1 的主题内容，过滤了噪声部分。

算法伪代码如下：

```
for(i = begin1 ; end1; j = begin2 ; end2; k = begin3 ; end3)
{
```

```
if(depth1 == depth2 == depth3)
if(i->text() == j->text() == k->text())
i 加入模板集合;
else
i 加入内容集合;
else
{
while(depth1 > depth2 || depth1 > depth3)
{
i 加入内容集合;
i++;
}
while(depth1 < depth2)
j++;
while(depth1 < depth3)
k++;
}
}
```

2 正文抽取

HTML 文档转换成 DOM 树以后，每个节点都有唯一确定的路径。网页中不同内容块的节点在 DOM 树中的公共路径较少，而同一内容块的节点的公共路径很长。本文以这些路径之间的相似度作为不同节点是否属于同一内容块的依据。所有的主题内容都在叶子节点上，记所有叶子节点的路径为：

$P=\{P_A,P_B,\cdots\}$，$P_A=\{T_{A_1},T_{A_2},\cdots,T_{A_n}\}$

其中 T_{A_i} 为文本节点内容。

例如：

```
<html>
  <body>
    <div>
      <p>This is the first block.</p>
      <p>This is the second block.</p>
      <p>This is the third block.</p>
    </div>
    <div>
      <p>test1</p>
    </div>
  </body>
</html>
```

这段网页源代码中的“This is the first block”节点的路径为：

$P_1=\{$<html>, <body>, , <p>, This is the first block$\}$

“This is the second block”节点的路径为：

$P_2=\{$<html>, <body>, , <p>, This is the second block$\}$

记深度相同的节点 A、B 的相似度为 $\mathrm{sim}(T_A,T_B)=\begin{cases}\frac{1}{2^{\mathrm{depth}}} & T_A=T_B\\ 0 & T_A\neq T_B\end{cases}$，depth 为节点的深度，则任意两个节点 A、B 的路径的相似度可以定义为：$\mathrm{sim}(P_A,P_B)=\sum_{\substack{i\in(1,n_A),\\ j\in(1,n_B)}}\mathrm{sim}(T_{A_i},T_{B_j})=\begin{cases}\sum_{i\in(1,n_A)}\mathrm{sim}(T_{A_i},T_{B_i}) & n_A\leqslant n_B\\ \sum_{i\in(1,n_B)}\mathrm{sim}(T_{A_i},T_{B_i}) & n_A>n_B\end{cases}$

其中 n_A、n_B 分别表示节点 A、B 的深度。

通过对大量网页的研究发现，正文内容节点大都拥有共同的父节点或者祖父节点，取阈值 $\mathrm{Th}=1-\frac{1}{2^{\mathrm{depth}(\max l)-2}}$，其中，max$l$ 为 P 中字符最多的节点；depth 为节点深度，即路径 P_i 中的元素个数。记集合 P 中字符最多的节点为 L，与 P 中其他节点计算相似度，大于阈值的作为正文内容。

3 实验结果分析

本文从新浪、网易、搜狐、腾讯等大型门户网站以及多家各类型网站中抽取了 1 000 个网页作为测试数据，采用基于网页 DOM 树节点路径相似度的正文抽取方法进行实验，去噪结果和正文抽取结果如表 1 所示。

表 1 本文方法的正文抽取实验结果

测试网页来源	网页总数	取得全部正文内容网页个数	正文抽取正确/错误	准确率/%	召回率/%
新浪	100	100	100/0	100.00	100.00
网易	100	93	90/3	90.00	96.77
搜狐	100	96	92/4	92.00	95.83
腾讯	100	99	97/2	97.00	97.98
中新网	100	96	92/4	92.00	95.83
环球网	100	98	95/3	95.00	96.94
其他网站	400	394	367/51	91.75	93.15
合计	1 000	976	933/67	93.30	95.59

从表 1 的统计结果可以看出，有 97.6% 的网页清洗掉了大部分的噪声并且完整保留了网页中的有效信息；对于新浪、网易等门户网站的抽取结果较好，都有 90% 以上的准确率和 95% 以上的召回率；对于其他不同结构的网站，本文的正文抽取方法也都能适用，很好地实现了网页正文抽取的工作，并且有着较高的准确率和召回率。

为了验证本文方法的有效性，以上述的 1 000 个网页作为样本，将本文方法与 BTE、DSC、FE、LQF、CCB 等算法进行对比实验，实验结果如表 2 所示。

表 2 各种算法对比结果

算法	平均准确率/%	平均召回率/%
BTE	75.19	89.16
DSC	68.27	92.31
FE	77.24	64.20
LQF	85.34	91.47
CCB	81.52	90.06
本文方法	93.30	95.59

由表 2 可以看出，本文提出的方法相对于现有的统计

方法有更好的准确率和召回率。

互联网的发展为用户带来了一个包含丰富信息的巨型数据库，但是如何识别其中的有效数据是应用的关键。本文的正文抽取方法利用网页 DOM 树节点路径相似的特点实现正文抽取，为之后的数据分类、分析等工作奠定了基础。

4 结论

本文根据新闻正文内容在网页中相对集中且同网站的新闻页面有相同模板的特点，提出基于网页 DOM 树节点路径相似度的正文抽取方法，先用正则表达式删除网页源代码中与正文内容无关的代码，然后将得到的网页转换为 DOM 树，再将目标网页的 DOM 树与另外两个网页的 DOM 树进行对比去除噪声，最后，根据节点路径相似度来抽取正文内容。该方法对来自不同网站的数据能够快速、准确地抽取正文内容，适用于结构变化不大的网页，但是对正文内容较少的网页抽取效果仍有待提高。下一步主要工作是加入内容节点与标题节点的路径之间的距离判断节点是否为正文，以提高算法的准确度。

参考文献

[1] KUSHMERICK N, WELD D S, DOORENBOS R. Wrapper induction for information extraction[C]. IJCAI 1997: Proceedings of the 1997 International Joint Conference on Artificial Intelligence, 1997: 729-737.

[2] FU L, MENG Y, XIA Y J, et al. Web content extraction based on webpage layout analysis[C]. ITCS 2010: Proceedings of the 2010 Second International Conference on Information Technology and Computer Science, 2010: 40-43.

[3] CAI D, YU S P, WEN J R, et al. VIPS: a vision based on page segmentation algorithm[R]. Microsoft Co., Tech. Report, 2003.

[4] WANG J Q, CHEN Q C, WANG X L, et al. Basic semantic units based web page content extraction[C]. SMC 2008: Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetics, Piscataway, NJ: IEEE Press, 2008: 1489-1494.

[5] UZUN E, AGUN H V, YERLIKAYA T. Web content extraction by using decision tree learning[C]. SIU 2012: Signal Processing and Communications Applications Conference, 2012: 1-4.

[6] PAN D H, QIUS G, YIN D W. Web page content extraction method based on link density and statistic[C]. WiCOM 2008: Wireless Communications, Networking and Mobile Computing, Dalian, China, IEEE Press, 2008: 1-4.

[7] REIS D C, GOLGHER P B. Automatic web news extraction using tree edit distance[C]. Proc. WWW 2004: The 13th International Conference on World Wide Web, New York: ACM, 2004: 502-511.

[8] FINN A, KUSHMERICK N, SMYTH B. Fact or fiction: Content classification for digital libraries[C]. Proc of the 2nd DELOS Network of Excellence Workshop on Personalization and Recommender Systems in Digital Libraries. Dublin, Ireland, 2001: 1-6.

[9] PINTO D, BRANSTEIN M, COLEMAN R, et al. QuASM: A system for question answering using semi-structured data[C]. Proc of the 2nd ACM/ IEEE-CS Joint Conference on Digital Libraries. Portland, USA, 2002: 46-55.

[10] MANTRATZIS C, ORGUN M, CASSIDY S. Separating XHTML content from navigation clutter using DOM-structure block analysis[C]. Proc of the 16th ACM Conference on Hypertext and Hypermedia, Salzburg, Austria, 2005: 145-147.

[11] DEBNATH S, MITRA P, GILES C L. Automatic extraction of informative blocks from webpages[C]. Proc of the ACM Symposium on Applied Computing, SantaFe, USA, 2005: 1722-1726.

[12] GOTTRON T. Content code blurring: A new approach to content extraction[C]. Proc of the 19th International Conference on Database and Expert Systems Applications, Turin, Italy, 2008: 29-33.

[13] 刘利，戴齐，尹红风，等. 基于多特征融合的网页正文信息抽取[J]. 计算机应用与软件，2014，31(7)：47-49.

[14] 王利，刘宗田，王燕华，等. 基于内容相似度的网页正文提取[J]. 计算机工程，2010，36(6)：102-104.

[15] YI L, LIU B, LI X. Eliminating noise information in web pages for data mining[C]. SIGKDD 2003: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, 2003: 296-305.

[16] GIBSON D, PUNERA K, TOMKINS A. The volume and evolution of web page templates[C]. Proc. WWW 2005: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, New York: ACM, 2005: 830-839.

[17] WANG J Y, LOCHOVSKY F H. Data-rich section extraction from HTML pages[C]. WISE 2002: Proceedings of the 3rd International Conference on Web Information Systems Engineering (Workshops), Los Alamitos, CA: IEEE Computer Society, 2002: 313-322.

（收稿日期：2016-05-13）

作者简介：

潘心宇（1992－），男，硕士研究生，主要研究方向：数据挖掘、模式识别。

陈长福（1974－），男，学士，主要研究方向：网络信息挖掘、信息分类。

刘蓉（1972－），通信作者，女，硕士，讲师，主要研究方向：数值计算。E-mail：liu_r@fzu.edu.cn。