基于 DIV 标签树的网页主题信息抽取方法①

欧阳柳波 杨 柱 易 显 (湖南大学 软件学院 湖南 长沙 410082)

摘 要： 随着 CSS+DIV 布局方式逐渐成为网页结构布局的主流，对此类网页进行高效的主题信息抽取已成为专业搜索引擎的迫切任务之一。提出一种基于 DIV 标签树的网页主题信息抽取方法，首先根据 DIV 标签把 HTML 文档解析成 DIV 森林，然后过滤掉 DIV 标签树中的噪声结点并且建立 STU-DIV 模型树，最后通过主题相关度分析和剪枝算法，剪掉与主题信息无关的 DIV 标签树。通过对多个新闻网站的网页进行分析处理,实验证明此方法能够有效地抽取新闻网页的主题信息。

关键词： 主题信息抽取；DIV 标签树；STU-DIV 模型树；主题相关度；剪枝算法

A New Way of Extracting the Topic Information in Web Pages Based on DIV Tag-tree

OU YANG Liu-Bo, YANG Zhu, YI Xian (Software College, Hunan University, Changsha 410082, China)

Abstract: Since CSS+DIV Topological Mode has become the major trend of the structural layout of web pages, the efficient extraction of the topic information in these web pages has become one of the urgent tasks for all professional surfing engines. This paper puts forward a new way of extracting the topic information in web pages based on the DIV tag-tree. It divides HTML files into DIV-forest with the help of DIV-tag. Then it filters the noise nodes in DIV tag-trees and sets up STU-DIV model-trees. Finally, it crops the DIV tag-trees irrelevant to the topic information by Topic Corelation Analysis and Cut-Tree Algorithm. It proves that this method can efficiently extract the topic information in web pages by analyzing several news web pages .

Keywords: extraction of topic information; DIV tag-tree; STU-DIV model-tree; topic corelation; Cut-Tree algorithm

1 引言

搜索引擎自其出现以来就得到了迅猛的发展，并日渐成熟，然而网页构建技术的革新使得传统的专业搜索引擎不得不采用新技术和新方法来应对。随着 CSS+DIV 布局方式逐渐成为网页结构布局的主流，对此类网页进行高效的主题信息抽取已成为各类专业搜索引擎的迫切任务之一。尤其在新闻网站和博客网站中，正文信息和相关链接存放着该网页的主题信息，是用户获取感兴趣信息的主要区域，对这些主题信息准确而高效地提取，是新闻搜索引擎和博客搜索引擎的重要基础。

近年来国内外学者提出了许多网页信息抽取方法[1]，根据抽取原理和抽取方式的不同，分为以下几类：Soderland S [2]提出的基于自然语言处理的方法，它的主要问题在于对各领域、各样式的 Web 页面给出一个有效的小规模训练样本很困难。文献[3,4]提出的基于包装器的信息抽取是从几个不同信息源中抽取信息，需要一系列的包装器程序库。由于一个包装器只能处理一种特定的信息源，当出现一类新的 Web 页面或者旧的页面结构发生变化,原来的包装器就会失效，无法从数据源中获得数据或得到错误的数据。基于 Ontology 的方法[5]需要构造一个完整的 Ontology 库，而这要由专家花费很长时间。Gatterbauer W,Bohunsky P.[6,7]提出的基于 Table 结构进行信息抽取的方法以及常见的基于 DOM 树[8,9]的信息抽取方法，其主要问题是网页中的同一主题信息往往分布在不同的 Table 标签结构中，使主题信息比较分散，为信息抽取带来很大困难。在 CSS+DIV 布局的网页中，

① 基金项目:国家自然科学基金(60970098，60803024)
收稿时间:2009-11-08;收到修改稿时间:2009-12-09

主题信息集中分布在 DIV 标签对里，单一地从 Table 标签结构或 DOM 树中提取信息，已经不适用。为解决这种情况，本文利用 DIV 标签树结构，提出构建 STU-DIV 模型树，结合主题相关度分析和剪枝算法来进行网页主题信息抽取的方法。

2 主题信息抽取

2.1 基本定义

定义 1. DIV 标签树 :在 CSS+DIV 布局的网页中，每一个 … 标签对里面存在若干 HTML 的标签，将这些标签看成一个个的结点，多个这样的结点构成 DIV 标签树。

若网络爬虫抓取的网页是一个包含多个 DIV 标签对的 HTML 文档,，则可以把整个 HTML 文档转化为由多个 DIV 标签树构成的森林。

定义 2. STU 树：STU 即语义文本单元，DIV 标签树中每一个结点可以包含两个 STU 结点。其中一个 STU 结点代表这个 DIV 标签树中的结点有多少个链接文字数(即锚文本字数)L，另一个 STU 结点代表着这个 DIV 标签树中的结点含有多少个非链接文字数 C 由这三个结点构成一棵 STU 树。如下图：

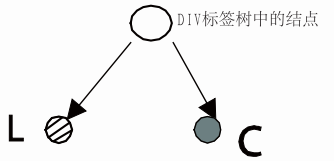

图 1　STU 树

定义 3. STU-DIV 模型树：STU-DIV 模型树是向 DIV 标签树中结点添加 STU 结点，使 DIV 标签树中的每个结点变成一棵 STU 树，从而形成 STU-DIV 模型树。

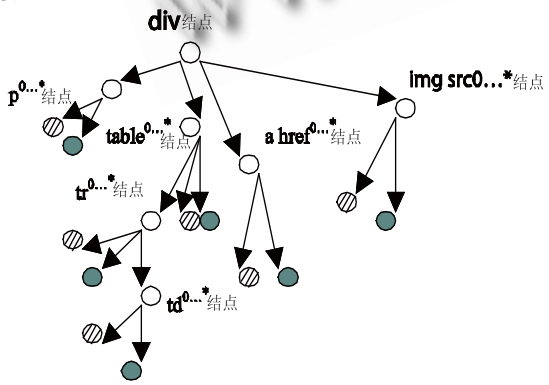

图 2　STU-DIV 模型树

2.2 主题信息抽取过程

基于 DIV 标签树(树形结构的)信息抽取有如下四个过程：HTML 解析，噪声过滤，主题相关度分析，剪枝，如图 3 所示。

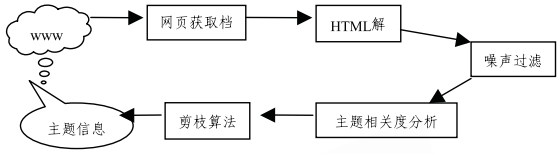

图 3　主题信息抽取过程

2.2.1 HTML 解析

传统的 HTML 解析算法只是通过读取网页源代码形成 HTML 文档，没有对 HTML 文档进一步处理。本文提出在获取 HTML 文档后，从文档中抽取每一个 DIV 标签对，DIV 标签对可以嵌套。由于每一个 DIV 标签对对应着一棵 DIV 标签树，因此将嵌套的 DIV 标签树抽取出来作为被嵌套的 DIV 标签树的子树，从而将 HTML 文档转换成 DIV 森林。这样就容易看出两棵 DIV 树有没有共同的祖先。 下例是两个嵌套的 DIV 标签对，它对应着 DIV 森林中两棵嵌套的 DIV 树。

9 月 4 日下午，温家宝在北京市第三十五中学主持召开北京市教师代表座谈会前，和出席座谈会的教师代表握手。新华社记者李学仁 摄

其中第二个 DIV 标签对原来嵌在第一个标签对里面，抽取出来后，形成了第一个 DIV 标签对对应的 DIV 标签树的子树。

2.2.2 噪声过滤

噪声过滤：从树的根结点开始，删除所有无关的噪声结点。噪声结点通常是图片(img) 、脚本(script)，form , style ,iframe 等。如果某个噪声结点中还嵌套了 DIV 标签对，例如 form 表单标签中就有可能嵌套 DIV 标签，因为在第一步的解析过程中，已经将 DIV 标签对提出，形成了这棵被嵌套的 DIV 树的子树，所以这些噪声结点可以直接删除。

删除这些噪声结点的原因是这些节点一般不包含主题信息。包含主题信息的结点称为主题相关结点，一般有<title>， ，<font>，<a href>, <h>，<p>，<td>，因此只需要从这些结点中提取信息，将其它不包含主题信息的噪声结点过滤掉。

设定阈值 k,如果某个链接标签上的文字小于 k 个，也应将该标签过滤掉，一方面因为它们大多数导航到网站的其他主题页面，一般分布在网页的顶部；另一方面，链接上的文字要描述一个主题信息在绝大多数情况下不会少于 k 个字。采取过滤算法还一个重要原因是不必为 DIV 标签树中每个结点去分析它的主题相关度，节约时间的开销。

2.2.3 主题相关度分析

一棵过滤后的 DIV 标签树有四种情况(1)只包含非链接文字；(2)只包含链接文字；(3)既包含非链接文字也包含链接文字；(4)链接文字和非链接文字都不包含。

在情况(1)和(3)下，DIV 标签树的主题相关度(Topic_correlativity)，由下面的公式来计算：

$$TC(DIV_i)=\frac{\sum_{j}^{N} Cij}{\sum_{j}^{N} Lij}$$

其中 $TC(DIV_i)$ 表示第 i 棵 DIV 标签树的主题相关度($1\le i\le$DIV 标签树的总数)，$\sum_{j}^{N} C_{ij}$ 表示第 i 棵 DIV 标签树中所有有关结点包含的非链接文字总数($0\le j\le$第 i 棵 DIV 标签树中有关结点数 N)；$\sum_{j}^{N} L_{ij}$ 表示第 i 棵 DIV 标签树中所有有关结点包含的链接文字总数。

在这两种情况下将主题相关度的阈值设为 m，如果 $TC(DIV_i)\ge$m，那么称第 i 棵 DIV 标签树主题相关，且这种情况下的 DIV 标签树包含着正文信息。

遍历所有包含正文信息的 DIV 标签树以后，接下来到相关链接。因为相关链接绝大多数情况下分布在正文的下方，所以从最后一个包含正文信息的 DIV 标签树下面开始遍历。在情况(2)或者 00$\le TC(DIV_i)<$ m 的 DIV 标签树满足以下三种情况时，

1)设参数 d,离最后一个包含正文信息的 DIV 标签树最近且距离不超过 d 棵 DIV 标签树；

2)与最后一个包含正文信息的 DIV 标签树有共同祖先；

3)设定参数 s,链接数>=s；

DIV 标签树视为主题相关，DIV 标签树中的链接为相关链接。

在情况(4)下的 DIV 标签树，因为没有包含任何信息，在下面介绍的剪枝算法中，将直接剪掉。

2.2.4 剪枝算法

设定一个阀值 n，规定一棵 DIV 标签树的非链接文字总和即 $0\le \sum_{j}^{N} C_{ij} <n$,且 $\sum_{j}^{N} L_{ij}=0$ 时，那么称这棵 DIV 标签树为空。上文主题相关度分析中的情况(4)的 DIV 标签树就是为空。

剪枝算法：深度遍历整个 DIV 森林，首先剪掉最后两棵 DIV 标签树，因为这两棵树绝大多数情况是描述网站的版权信息。如果 DIV 标签树为空，也应直接剪掉；然后通过上文介绍的主题相关度分析先提取所有包含正文信息的 DIV 标签树然后再找到包含相关链接的 DIV 标签树，将其它不相关的 DIV 标签树剪掉。

整个剪枝算法如下：

```
/**
 * 找到包含正文信息的 DIV 标签树
*/
int find_TC(DIVi) {
    while(遍历 DIV 森林){
        if(DIVi 为最后两棵 DIV 标签树){
            删除 DIVi;//去掉版权信息
        }
        If(DIVi 为空) {
        DIVi,主题无关;
        删除 DIVi;
          continue;
    }else {
        If(DIVi 的非链接文字总数!=0) {

        if(DIVi 的主题相关度> 阀值 m) {
            DIVi 主题相关
            id = i;
        }
     }
    }
```

```
    返回最后一个包含正文信息的 DIV 标签树的 id
}

/**
 * 找到包含相关链接的 DIV 标签树
*/
int link_TC(find_TC 函数的返回值 id){
    for(int j=id+1; j<= DIV 标签树的总个数; j++){

        if(DIVj 的非链接文字总数!=0&& DIVj 的链接个数 >= s && DIVj. 与
    DIVid  有共同的祖先)
        {
            DIVj 主题相关;
            id=j;
            返回 DIVj 的 id;
        }
    }
}
```

3 实验

通过调查可知现在知名新闻网站和博客网站大多数是采用 CSS+DIV 布局。本文用 Java 在 Eclipse 开发平台下，实现适用于新闻网站的基于 DIV 标签树的信息抽取算法。具体实验步骤如下：

Step1: 提取网页源代码，并形成 HTML 文档。

Step2: 对HTML文档进行预处理,首先去掉HTML文档中的 HTML 注释部分以及 Javascript 中的注释。再提取<title>标签内容，形成网页主题信息的标题。

Step3：通过解析预处理后的 HTML 文档，形成 DIV 森林。

Step4：再通过过滤过程，过滤掉每一棵 DIV 标签树中信息无关的结点。

Step5：判断 DIV 标签树是否为空，为空的话直接删掉，然后再深度遍历每一棵 DIV 标签树，计算它们的主题相关度，直到找到所有包含正文信息的 DIV 标签树。最后再找到网页的相关链接。

Step6：通过剪枝算法，剪掉信息无关 DIV 标签树，只剩下 Step5 中找到的主题相关的 DIV 标签树。

实验从以下三个指标分析了本文提出的主题信息抽取算法的性能。

正文信息保留的完整性(TRF)：正文信息保留完整的结果网页数占来源网页数的百分比。

相关链接保留的完整性(LRF)：相关链接保留完整的结果网页数占来源网页数的百分比。

无关链接删除的彻底性(ULQ)：无关链接删除彻底的结果网页数占来源网页数的百分比。

表 1　抽取结果表

来源网站	网页数目	TRF (%)	LRF (%)	ULQ(%)
新浪	100	95	94	94
网易	100	99	93	92
中国新闻网	100	99	98	92
中华网	100	98	95	97
腾讯网	100	97	93	91

实验结果是在链接上文字个数的阀值 k 取值为 6；主题相关度阀值 m 取值为 0.5；判断 DIV 为空的阀值 n 取值为 50；以及参数 d，s 分别为 5，2 情况下得到的。主题信息保留的完整性，无关链接删除的彻底性都达到了 90%以上。由于 CSS+DIV 布局的网页中 table 标签几乎不用于存储正文信息，所以基于 table 标签的主题信息抽取在这种布局的网页中，几乎抽取不到主题信息，特别是正文信息。

4 小结

本文针对 CSS+DIV 布局的网页中 DIV 标签树这一特殊结构，提出了一种新的主题信息抽取方法，改变了以前单一地从 TABLE 结点提取信息的模式，并通过实验证明此方式的好处。以后的工作中，可以进一步提高阀值自动调节性能，使抽取的主题信息更加准确。抽取后的主题信息，可以用于提高网页分类和信息检索的效率和准确性。

参考文献

1 Laender AHF, Ribeiro-Neto BA，da Silva AS, Teixeira JS. A Brief Survey of Web Data Extraction Tools. SIGMOD Records, 2002,31(2):1－3.

2 Soderland S. Learning information extraction rules for semi-structured and Free Tex. Machine Learning, 1999,34(1-3):233－272.

3 Muslea I, Minton S, Knolock C. Hierarchical wrapper induction for semistructured information sources. Autonomous Agents and Multi-Agent Systems, 2001,4(1/2):93－114.

(下接第 139 页)

(上接第 196 页)

4 Craig A, Knoblock Kristna L, et al. Accurately and reliably extracting data from the web: A machine leanring approach. Data Engineering Bulletin, 2000, 23(4):33－41.

5 Benslimane SM , Malki M, Rahmouni MK, D Benslimane. Extracting Personalised Ontology from Data-Intensive Web Application: an HTML Forms-Based Reverse Engineering Approach. Informatica, 2007,18(4):511－534.

6 Gatterbauer W, Bohunsky P. Table Extraction Using Spatial Reasoing on the CSS2 Visual Box Model. In: Proc. of the 21st National Conference on Artificial Intelligence (AAAI 2006),Washington: AAAI Press, 2006.1313－1318.

7 Gatterbauer W, Bohunsky P, Herzog M, et al. Towards Domain Independent Information Extraction from Web Tables. Proc. of the 16th International World Wide Web Conference(WWW 2007), 2007.71－80.

8 李效东,顾毓清.基于 DOM 的 Web 信息提取.计算机学报, 2002,25(5):526－533.

9 杨俊, 李志蜀.基于 DOM 的 WEB 主题信息抽取.四川大学学报, 2008,45(5):1077－1080.