

XML 的 DOM 树结构在 WEB 挖掘中的应用

卢远征^a, 叶晓彤^b

(四川理工学院 a. 自动化与电子信息学院; b. 网络管理中心, 四川 自贡 643000)

摘 要:面对飞速发展的信息时代,WEB 数据的挖掘日益重要,而传统的搜索引擎难以胜任对数据的挖掘处理。基于 XML 良好的结构性和层次性,提出了利用 DOM 树进行 WEB 挖掘的方法。首先利用 Tidy 工具库将 WEB 数据转换成良好结构的 XML 文件,简化生成 DOM 树,然后通过遍历解析 XML 的 DOM 树结构,提取需要的 WEB 信息,实现对 WEB 数据挖掘。实验表明,该方法能够方便地对数据进行结构化存储和信息处理。

关键词:WEB 挖掘; XML; Tidy; DOM 树

中图分类号:TP391.1

文献标志码:A

引 言

随着 WEB 信息技术的飞速发展,WEB 信息也以指数级日益增加,因此,如何从海量的数据中准确得到相关有效数据,成为了急需解决的问题。搜索引擎是从海量信息中获取指定数据的有效方式,但传统的基于 HTML 的搜索引擎仅仅能抓取和呈现孤立的数据本身,而难以对获取的数据进行有效的分类规整,更难以在此基础上进行智能化利用^[1]。因此,探索一种能将数据抓取和数据结构化存储融为一体的新的信息挖掘方式,将提升 WEB 信息的挖掘和利用效率。

但是,目前 WEB 信息的主要载体是 HTML 网页,这是一种半结构化的文本串集合,由各类信息和 HTML 标签组成,并且用户与服务器之间信息的传递主要依赖超文本传输协议(HTTP)。因此,要从网页中提取有用的数据,就要从中过滤各类 HTML 标记以及其它不相干信息。由于 HTML 语言的半结构化特性以及各开发人员在内容组织的巨大差异,甚至有的浏览器 HTML 语言在书写上要求不规范,给网页信息抽取带来了巨大困难和挑战,这使得传统基

于 HTML 的搜索引擎难以有效完成信息挖掘和智能化应用^[2]。

为解决上述问题,本文提出一种基于 XML(Extensive Markup Language)的 WEB 信息挖掘解决方案:通过 Tidy 工具包将 HTML 转换为结构良好的 XML 文件,再根据其 DOM(Document Object Model)树结构的特点,遍历解析每个节点得到需要的信息。这样的信息是结构化存储的,并能通过 XML 应用方便地实现对数据的智能化应用。

1 WEB 挖掘

WEB 挖掘指使用数据挖掘技术在万维网(World Wide Web)数据中发现潜在的、有用的信息。WEB 挖掘现在已经研究涉及到了多个研究领域,包括统计学、信息获取技术、数据库技术、人工智能(Artificial Intelligence)中的机器学习和神经网络等。WEB 挖掘一般可以分为三类^[3]:WEB 内容挖掘(Web Content Mining)、WEB 结构挖掘(Web Structure Mining)、WEB 用法挖掘(Web Usage Mining)。本文主要研究的是 WEB 的内容挖掘。

收稿日期:2013-03-27

基金项目:四川理工学院研究生创新基金项目(y2012007)

作者简介:卢远征(1988-),男,河南沈丘人,硕士生,主要从事 Web 信息挖掘和处理方面的研究,(E-mail) 664209898@qq.com

2 XML

2.1 XML 语言的特点

XML 是万维网上信息交换的新标准,它支持用户自定义标记,用有序的、嵌套的元素组织数据,是面向数据的,程序可读取这些标记并依据标记的语义处理数据。XML 具有 XML 文档的内容和结构完全分离、互操作性强、格式规范统一、支持多种编码、可扩展性等优点^[4]。

2.2 XML 和 HTML 的差异

XML 与 HTML,虽然同为 SGML(Standard Generalized Markup Language,标准通用标记语言)的一种,但是,HTML 的缺点比较突出,例如:(1)HTML 仅提供一种显示信息的方式;(2)标签是预先定义好的,不可扩展;(3)不具备任何高层次标记语言能力,不能理解文档概念,不具备结构化,不利于信息存储等。与之相对的是,XML 具有:(1)良好的扩展性;(2)XML 是面向数据而非面向显示的,其内容与形式相互分离、内容易识别、修改方便、数据显示灵活等优点。由于 XML 的优点突出,且能克服 HTML 的诸多缺点,人们开始研究用 XML 文件来存储组织管理 Internet 上庞大的信息资源。

2.3 XML 与 HTML 的互补

XML 还可以通过简单开放扩展的方式描述结构化的数据,XML 补充了 HTML,被广泛地用来描述使用者界面。HTML 描述数据的外观,而 XML 描述数据本身。由于数据显示与内容分开,XML 定义的数据允许指定不同的显示方式,使数据更合理地表现出来。本地的数据能够以客户配置、使用者选择或其他标准决定的方式动态地表现出来。

3 XML 的 DOM 结构在 WEB 挖掘中的应用

XML 已经成为正式的规范,开发人员能够用 XML 的格式标记和交换数据。XML 在三层架构(MVC)上为数据的处理提供了很好的方法。使用三层模型,XML 可以从数据中产生出来,形成的 XML 结构化数据可以使数据从商业规范和表现形式中分离出来,使 XML 在 WEB 挖掘中得到充分应用。

3.1 HTML 的转换

HTML Tidy 是一个 W3C 提供的一个免费工具,它可以很容易将普通网页文件转换为规范 XHTML 格式或者 XML 格式的文档,并且把网页中不符合规范的语法和格式修正过来,这样就可以像处理 XML 文档一样处理 HTML 文件^[5]。从 HTML 到 XML 转换的模型图^[6]如图 1 所示。

Tidy 实现转换的核心代码如下:

```
public static void main( String args[ ] ) {
.....
try {
.....
Document html = XMLHelper. tidyHTML ( URL );
Document xsl = XMLHelper. outputXMLToFile( html,
" XML" + File. separator + " result. xml" );
} catch ( XMLHelperException xmle ) {
.....
}
}
```

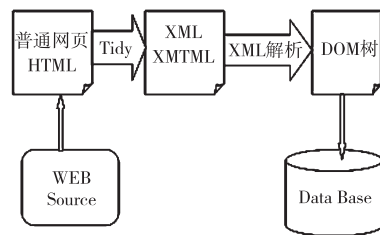


图 1 XML 的转化模型

3.2 网页的 DOM 树结构

根据网站 WEB 网页的结构特征,把网页转换成 XML,可以很清楚的看见其 DOM 结构(图 2)。DOM 是 HTML 和 XML 文档的一个应用程序接口(API),它提供了一种结构化的文档表示方法,使用户可以修改文档的内容以及最终的表现形式。

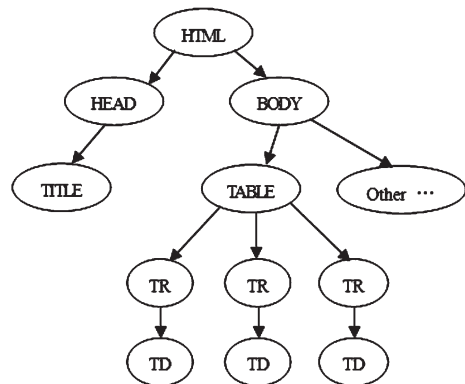


图 2 网页结构的 DOM 树结构

如图 2 所示,WEB 页面 DOM 树是根据 WEB 页面中的标签创建而成,反映了 HTML 文档的内容和组织结构,即组成 HTML 文档的各个元素以及各个元素之间的构成关系。WEB 页面的传统 DOM 树虽然能够反映文档的组织结构,却没有体现页面的视觉特性和语义信息。页面的视觉特征一般体现在字体、背景颜色、段落

划分等方面;语义信息一般表现为页面内容的类型,如文本、多媒体或超链接等。如果充分利用 WEB 页面的视觉特性和语义信息,那么将会提高信息抽取的正确率^[7-8]。图 2 对应的 HTML 的基本框架结构为:

```

<html >
<head >
Head 区域块内容
</head >
<body >
Body 区域块内容
</body >
</html >

```

3.3 网页标题和内容的获取

由于网页的半结构化的特性, TITLE 很容易的通过标签分析得到。但要得到完整的正文却很复杂。首先,网页中没有明显的标签标识出正文;其次,正文可能分散在多个 HTML 标签中,问题在于如何组合出完整的正文。

在视觉上,网页是由若干内容块组成的。当 HTML 网页被解析为 DOM 树后,内容块的信息会保存在某个节点中,该节点的标签属性称之为容器标签。经常用到的容器标签有: <table >、<tr >、<td >、<p >、<div >、 等。在制作网页的时候,这些容器标签大部分是允许被嵌套使用的,即如一个表格(内容块)被嵌套在另一个表格(内容块)之中,可以依据嵌套在内容块中的各个容器类标签把网页划分为更小的块,直到无法再拆分为止,而网页正文信息的抽取是以内容块为单位进行保留和删除的。在转换成 DOM 树之后,需要将 HTML 文档中的一些仅仅为了增强网页的交互性或美观性而基本不可能包含网页正文信息的标签以及标签之间的内容删除掉。例如:注释标签(<!-- -->)、样式标签(<style >)、脚本标签(<script >)等标签,其正则表达式见表 1。

表 1 标签的正则表达式

标 签	正则表达式
内嵌图片	< \s * img ([^] * ?) >
HTML 注释	< ! - [\s \S] * ? - - >
Script 脚本	< \s * script ([^ >] * ?) > [\s \S] * ? < \s * / \s * script \s * >
Css 样式表	< \s * style ([^ >] * ?) > [\s \S] * ? < \s * / \s * style \s * >
Meta 标签	< \s * meta ([^ >]) * ? >
链接标签	< \s * link ([^ >]) * ? >
换行符	< \s * hr ([^ >]) * ? >
.....

正则表达式的核心代码:

```

Pattern pattern = Pattern.compile(RegExpName);
//RegExpName 存放正则表达式
Matcher matcher
= pattern.matcher(sb);
while( matcher.find() ) {
//调用 Matcher 类的 find() 方法来查看是否匹配
.....
}

```

在信息过滤玩后再根据 DOM 树的遍历取得各个节点的信息内容。

- Node getFirstChild(): 得到第一个孩子的节点;
 - Node getLastChild(): 得到最后一个孩子的节点;
 - Node getNextSibling(): 得到下一个兄弟的节点;
 - Node getPreviousSibling(): 得到上一个兄弟的节点;
 - Node getParentNode(): 得到父亲的节点;
 - Node getFirstChildNodes(): 得到所有孩子的节点;
- 其正则过滤的核心代码:

```

public static void main( String [ ] args ) throws Exception {
DOMParser parser = new DOMParser();
//创建 DOM 解析器
parser.parser( "URL" );
//解析网页
print( parser.getDocument(), "" );
//从根节点遍历
.....
}

```

采用遍历整个 DOM 标签树,判断当前的结点是否属于 HTML 的标签结点,然后对 DOM 树节点显示属性分析,再根据显示属性对 DOM 节点进行操作。具体过程为:(1)先获取一个要去噪的页面 DOM 树结构;(2)然后,对此树状分级结构中叶节点进行显示属性分析,如果该节点的子节点中,所有子节点都具有相同的显示属性,则合并该节点所有子节点,如果不相同,则可以根据实际情况进行合并或者不合并。以上是基于 DOM 树结构挖掘信息的过程,即先是解决对网页内容的结构分析,然后利用网页的显示属性和 DOM 技术相结合简化生成 DOM 树;(3)最后利用此 DOM 树结构对网页信息进行处理,最终实现预期效果。

图 3 为原网页的部分截图,图 4 为原网页的 HTML 源代码部分截图。



图 3 部分网页截图

```
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=gb2312" />
<title>金牛座 每日运势_星座频道_新浪网</title>
<link href=" ../css/mamacity.css" rel="stylesheet" type="text/css" />
```

图 4 部分网页代码截图

根据上述内容提取得到网页的标题和正文如图 5 所示。

```
-----分析网页结果如下-----
得到网站标题为: 金牛座_每日运势_星座频道_新浪网
有效日期:2012-09-24
与小动物的相处让人愉快忘掉生活压力。
家中有养宠物的人今天就好好跟它们玩一下,能够帮忙清理宠物的居家环境更好。
没有养宠物的人看看这方面的照片、网站也有帮助,与朋友们聊天时动物也是很好的话题呢。
```

图 5 实验结果截图

3.4 其他信息处理

以上采用对 DOM 树的遍历实现了对网页信息的分类显示处理,实际上其他信息处理(统计、分析、归纳等)同样可以采用将 DOM 树遍历和程序处理相结合的方式实现,如与人工智能信息技术结合,还能实现对数据的智能化应用,这是传统搜索引擎挖掘 WEB 信息难以做到的。

4 结束语

本文基于 XML 的 DOM 树结构来提取网页的基本信息,主要利用了 HTML 标签语言的结构化的特点,针对规范解析后的网页进行正则过滤以及对主题标签提取,就可以得到想要提取的内容了。由于 XML 语言具有良好的结构性,我们也可以利用其对网页数据进行提取和处理。在新一代的语义网中,本文方法将发挥更大的作用。

参考文献:

- [1] 陈金森,原福永,张园园.XML 搜索引擎研究[J].图书情报工作,2007,51(1):114-117.
- [2] 钱程,阳小兰.HTML 到 XML 转换研究[J].计算机与现代化,2011(8):39-41.
- [3] 秦鸿.基于 Web 的数据挖掘[J].电子科技大学学报,2002,31(7):56-59.
- [4] 黄磊,黄汉永.XML 技术在 Web 挖掘中的应用[J].信息技术,2003,27(5):6-7,13.
- [5] 李霞,蒋盛益.基于 DOM 树及行文本统计去噪的网页文本抽取技术[J].山东大学学报:理学版,2012,47(3):38-42.
- [6] 熊一利,徐鹏.基于 XML 的网页数据挖掘[J].科技广场,2010(1):73-75.
- [7] 余静,刘万军.基于网页分块的主题爬虫研究[J].计算机与信息技术,2008(10):83-84.
- [8] 李龙,李丽丽,高玲.一种网络课程答疑系统分词器的设计[J].河北工程大学学报:自然科学版,2012,29(2):68-70.

Application of XML's DOM Tree in WEB Data Mining

LU Yuan-zheng^a, YE Xiao-tong^b

(a. School of Automation and Electronic Information; b. Network Administration Center, Sichuan University of Science & Engineering, Zigong 643000, China)

Abstract: Facing with the rapidly development of the information age, WEB data mining become increasingly important, and traditional search engines can not do the mining processing of data. So the method that takes advantage of the DOM tree for WEB mining is put forward based on good structure and level of XML. First WEB data is transformed into XML file for good structure by tool library, DOM tree is simply produced, then the heedell WEB information can be extracted through the traversal and parsing of DOM tree structure of XML to realize the WEB data mining. Experiments show that the method is easy for structured data storage and information processing.

Key words: WEB Mining; XML; Tidy; DOM Tree