文章编号：1005-8451（2010）09-0037-04

网页信息自动抽取技术的研究

胡少荣，孟嗣仪，刘 云，张彦超，丁 飞
（北京交通大学 网络舆论安全研究中心 100044）

摘 要：在网络舆情分析中，经常要从大量的网页信息中抽取出有用的数据。但一般的网页信息抽取技术都是基于对HTML文档的分析。本文提出网页信息自动抽取的方法，可以滤除网页噪声，快速准确地获取所需要的网页信息。该方法首先将HTML转换为结构化的XML文档，然后结合DOM4J和XPath语言建立网页解析模板库，最后根据模板的抽取规则对网页信息进行抽取。实验证明，该方法具有较高的回召率和查准率。

关键词：自动抽取；网页信息；解析模板；XPath；网络舆情
中图分类号：TP39 **文献标识码**：A

Research on automatic extraction technology of Web information

HU Shao-rong, MENG Si-yi, LIU Yun, ZHANG Yan-chao, DING Fei
(Center for Security Studies Public Opinion, Beijing Jiaotong University, Beijing 100044, China)

Abstract: In online public opinion analysis, it was needed to extract valuable information from large amount of Web source. But the common way of Web information extraction technology was based on the analysis of HTML documents. This paper proposed automatic extraction technology of Web information, it could eliminate noisy content, extract information efficiently. This method transformed HTML into structured XML model, then built Web pages parser template library by DOM4J and XPath, finally extracted the Web information according to rules of the parser template. Result showed that this method was high with recall and precision with retrieving.

Key words: automatic extraction; Web information; parser template; XPath; online public opinion

随着网络技术的飞速发展及其应用的深入，网络成为反映社会舆情的主要载体之一。舆情是指在一定的社会空间内，围绕中介性社会事件的发生、发展和变化，民众对社会管理者产生和持有的社会政治态度。它是公众通过互联网传播的对现实生活中某些热点、焦点问题所持的有较强影响力、倾向性的言论和观点。网络舆情表达快捷、信息多元，方式互动，具备传统媒体无法比拟的优势，因此越来越受到人们的关注。近些年来，随着国内网民数量的日益增长，网络舆论所具有的强大力量在一些重大新闻事件中得到了很大程度的彰显。因此科学分析舆论，对于网络舆情的正确引导和管理，具有重要的现实意义。网络作为巨大的数据源，如何从中提取出人们所关心的信息，滤除无用信息，是当今研究的热点。网络舆情分析中网页信息自动抽取技术的研究应运而生。

收稿日期：2010-01-27
基金项目：基金项目：国家自然科学基金资助项目（60972012）；教育部培育基金项目（707006）；铁道部科技研究开发计划重点课题（2008X019）；北京市教育委员会学科建设与研究生建设项目资助（JXKJD20090001）；通信与信息系统北京市重点实验室资助项目（JSYJD20090001）；教育部哲学人文社会科学研究重大课题（08WL1101）
作者简介：胡少荣，在读硕士研究生；孟嗣仪，副教授。

1 网页信息抽取技术

网络舆情分析中的网页信息抽取技术通过对网页进行处理，用一组信息描述所需要提取的信息，将其结构化后保存到数据库中，方便用户获取和利用这些信息。网页信息抽取的关键是保证信息抽取算法的准确性和健壮性。但是该技术主要的问题是要面对不断变化、更新的海量信息，并且大多数是以用于浏览，而不是用于数据操作和应用的HTML文档的形式出现。这就为网页信息抽取带来了极大的不方便。

目前，比较流行的抽取技术包括：基于隐马尔科夫链理论的HMM（HIDDEN Markov Model）[1]，基于ontology[2]的信息抽取，基于RBF[3]神经网络和关联规则的Web文本分类规则获取方法和基于数据挖掘MDR（Mining Data Records）[4]的算法。

以上算法都基于复杂的数学模型，实施起来比较困难，信息抽取的效率和准确性也不尽如人意。为最大程度地实现信息抽取的自动化，本文提出了网络舆情分析中网页信息自动抽取的方法，主要用于高效、精确地抽取并存储有用信息。目前，网络舆情的主要来源有各大新闻网站、论坛和博客。因此本文所采用的信息自动抽取技术也主要针对这3类网页信息进行处理。

本文涉及的网页信息自动抽取技术包括URL模板过滤网页、网页信息结构化、网页解析模板匹配和数据库存储，其操作方便，切实可行。

2 Web信息自动抽取技术的算法实现

2.1 网页信息自动抽取

网页信息自动抽取首先通过URL模板匹配过滤出可以解析的网页，然后将可解析的HTML文档进行网页结构化处理，生成XML文档。最后结合DOM4J和XPath语言建立页面解析模板，从XML文档中抽取指定节点信息，并将其存储进入数据库。抽取流程见图1。

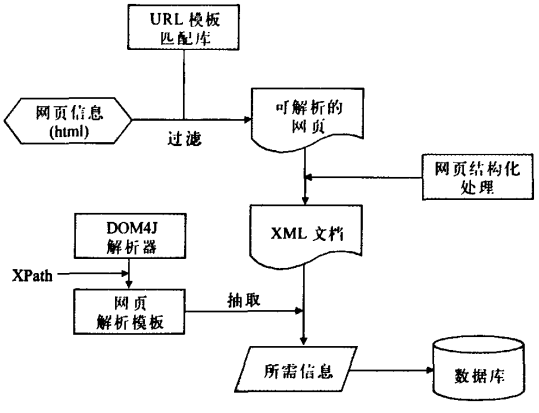

图1 网页信息自动抽取流程图

2.1.1 基于URL的模板匹配

由于在信息抽取中，页面解析模板包含了大量的路径信息，在进行匹配时，会消耗大量的时间。如果能在网页解析前对无关网页（如广告网页、用户没有定制的网页）进行一定的预处理的话，势必会对系统的运行效率有很可观的改善。

本文利用了网页URL模板匹配库来进行URL结构的过滤分析，该模板中主要包含了匹配URL的正则表达式和页面解析模板的选择参数。

正则表达式（regular expression）就是用某种模式去匹配一类字符串的一个公式。正则表达式由一些普通字符和元字符（matacharacters）组成，它被转换成特定的算法，根据这个算法来进行文本匹配。在许多程序设计语言中，正则表达式通常被用来作为检索或替换字符串数据的一种强大的工具。

正则表达式的强大功能不只是表现在特定的字符串匹配，而是字符类型的模式匹配。正则表达式中由很多特殊字符，它们分别用来匹配不同的字符类、制定匹配位置和制定重复字符。因此可以利用它来对需要处理的网址进行过滤。本平台在开发中正是利用了正则表达式的优点来对网页进行筛选的。

URL模板匹配库是一个包含了网站URL特征的XML文件，与待抽取网页的URL进行模板匹配，判断页面是否可以被解析并确定其网页解析模板。图2表示的是匹配网易论坛的URL模板。其中<regex>之间的数据就是网易论坛URL正则表达式的匹配形式，<module>之间的数据bbs_163_topic表示的是网易论坛的主题页面。经过URL模板库的过滤，可以过滤出网易论坛的网页并确定为论坛主题页面。否则，页面则被滤除。

```
<page name="网易论坛" sort="topic">
    <url>http://bbs.163.com</url>
        <regex>
(http://bbs\.\w+\.163\.com/bbs/\w+/\
d+\.html)
        </regex>
    <module>bbs_163_topic</module>
    <idName>163BBS_</idName>
</page>
```

图2 网易论坛URL模板代码

2.1.2 网页信息结构化

由于网络上的多数信息是用HTML语言来表示，其数据的异构性和半结构化使得这种语言不能处理网络上的很多需求。本文将XML应用在网页信息自动抽取中的主要目的就是为了解决这两方面的问题，为舆情分析中提供结构化的数据。

（1）HTML

HTML（超文本标记语言）是用于创建网页和进行信息发布的通用语言。格式和语法比较简单，

规定比较灵活。但是其表现过于简单、扩展性差，缺少语义性，许多功能受到了限制。

（2）XML

XML是一种元标记语言，它将结构、内容和表现分离，提供描述结构化资料的格式，有着良好的数据存储格式、可扩展性、高度结构化、语义性强、便于网络传输等优势，不仅能满足不断增长的网络应用需求，而且还能确保在网络进行交互时，具有良好的可靠性与互操作性。这就为本文的抽取信息方案提供了理论依据，确保其切实可行。

经研究，HTML网页均可转换为XML文档，经过转化后，可以清晰地查看到网页节点信息，从而能很方便地定位并抽取这些信息。如图3，这是经过转换后的XML的文档片段。

2.1.3　基于XPath的网页解析模板的设计

XPath（XML Path Language）是一门在XML文档中查找信息的语言，可用来在XML文档中对元素和属性进行遍历。XPath将一个XML文档建模成为一棵节点数，有不同类型的节点，包括元素节点，属性节点和正文节点。根据节点的名字，利用Xpath的导航能力可以直接定位到包含信息的节点，从根节点开始层层深入，逐步遍历，为每个节点构建一个地址，直到返回所需要的结果，从而得到XPath表达式，这可以减小基于文本的信息提取系统的搜索空间。例如，在图3中待抽取信息在模板中定义为：//DIV[@class='outContainer']/DIV/DIV/DIV[2]/DIV/LI/STRONG。

```
<?xml version="1.0" encoding="UTF-8"?>
<HTML>
  <HEAD>1</HEAD>
    <BODY>
      <DIV>2</DIV>
       <DIV class="outContainer ">
         <DIV class="articleleft">
           <DIV class = "articleTop">
            <DIV class="block">
             <DIV class="writerInfo">
              <li>
                <strong>抽取信息</strong>
              </li>
             </DIV>
            </DIV>
           </DIV>
         </DIV>
       </DIV>
    </BODY>
</HTML>
```

图3　经转换后的XML文档

其中，DIV为上层节点名称，class为节点属性，outContainer为节点属性值。表达式从根节点逐步递进到STRING节点，这样结合在一起就构成了待抽取信息相对路径的表达式。

本文针对网页信息划分了3类模板：（1）新闻解析模板；（2）论坛解析模板；（3）博客解析模板。基本上可以囊括大部分网络中的热点话题。在撰写本文前，URL模板库和网页解析模板库中已设计如下模板：论坛解析模板包括新浪、网易、腾讯的论坛主题页面及论坛回复页面模板；博客解析模板包括新浪、网易和聚友网的主题页面及博客回复页面模板；新闻解析模板包括网易、搜狐、凤凰网的新闻模板。这里主要介绍论坛的网页解析模板。

论坛解析模板库同样为XML文件格式，论坛页面的抽取信息一般包括发帖标题、所属板块、发帖作者、发帖时间、发帖内容、回复数量、回帖作者、回帖时间、回帖内容。这些基本上涵盖了我们所关心的重要信息。通过对每条信息指定XPath路径，就可以达到自动抽取网页信息的效果。另外，由于论坛有主帖和回帖之分，因此模板库中记录了区分主帖和回帖的统计信息，图4显示了网易论坛的解析模板部分代码。

```
<bbs name="网易" module="bbs_163_topic">
  <url>http://bbs.163.cn/</url>
    <title>
      <title_xpath comment="标题">
       //DIV[@class='head']/H3/A
      </title_xpath>
     </title>
     <edition>
       <edition_xpath comment="板块">
         //DIV[@class='cru aGray']/A[4]
       </edition_xpath>
      </edition>
-------------------------------------------------
 <retime>
  <retime_xpath comment="回帖时间">
   //DIV[@class='rightCont']DIV[@class='readMode']/H6
  </retime_xpath>
 </retime>
</bbs>
```

图4　网页解析模板部分代码

其中module属性是模板标识，区分主帖回帖，若经过URL匹配后得到的属性后缀是“topic”，那么可判断该页面为主帖，之后利用DOM4J解析器可以从指定的XML文档中自动抽取出主帖信息。若判断为回帖，则用回帖解析模板来抽取回帖信息。

2.2　数据处理及存储

由于网络舆情分析中需要处理的数据达到数亿级，因此对数据存储算法及数据库的优化设计就显得极为重要。在面对海量数据存储的过程中，最主要的是对重复的网页信息不再进行保存，这样可简化数据存储时的负担，并且为之后分析数据提供方便。文中网页信息自动抽取技术在存储数据时对数据库进行了优化，在解决避免重复数据的存入时，采用hashcaode（哈希值）作为表的索引，以论坛为例，通过对作者、时间、标题这3个字段组成的字符串进行哈希运算，由于不同的对象有不同的哈希值，因此在数据存储时能使信息数据的重复率大大降低，并且可以提高数据库查询效率。

3 实验结果分析

3.1 数据抽取评价指标

消息理解会议（MUC）为信息检索和信息提取领域内的算法性能测试提供了一系列的评估参数，主要参数是回召率（Recall）R_e和查准率（Precision）P_r，公式如下：

$$R_e=\frac{\text{被正确抽取出来的信息数}}{\text{正确的信息总数}} \quad (1)$$

$$P_r=\frac{\text{被正确抽取出来的信息数}}{\text{正确的信息总数}} \quad (2)$$

通常，查准率和回召率需要一起考虑，因此为了使得评估结果更全面、更具说服力，将二者结合成一个综合性的数据F，能计算R_e和P_r的加权几何平均值，其计算公式：

$$F=\frac{(1+\beta^2)*P_r*R_e}{\beta^2*P_r+R_e} \quad (3)$$

其中β为R和Pr的相对权重，决定了R_e和P_r的比值。通常β是一个预设值，决定对Pr侧重还是对R_e侧重。通常设定为1，这样用F一个数值就可看出系统的好坏[5]。

3.2 实验结果及分析

在如下平台中测试本系统的性能：实验机器的CPU为4.2 GHz，内存2.0 G，操作系统是Windows XP。运行环境为MyEclipse 6.5，数据库是MySql 5.0，程序使用JAVA语言。实验数据来源于北京交通大学红果园论坛（http://bbs.njtu.edu.cn/）2009年10月31日至2009年11月1日的网页信息。实验结果见表1。

表1 性能测试结果

N(页)	EN(页)	T(S)	Re(%)	Pr(%)	F(%)
166 856	163 873	118 466	98.21%	93.74%	95.92%

表1中：N表示待处理的网页数量；EN表示经过URL匹配可以解析的页面的数量；T表示抽取时间；R_e表示回召率；P_r表示查准率；F表示R_e和P_r的加权几何平均值。

实验结果表明，使用网页信息自动抽取方案可以有效地完成信息抽取任务，处理速度较快，准确率较高，基本上达到了实验预期的目的。

4 结束语

网络舆情分析越来越受到大众的关注，如何能高效抽取网页有效信息成为研究的热点之一。本文提出了网页信息自动抽取方案，通过网页结构化处理将HTML文件转换为易于数据交换的XML文档，结合DOM4J和XPath语言建立网页解析模板，根据模板的抽取规则对网页信息进行自动抽取。实践证明，该方法能精确高效地自动抽取网页信息，并且实现方便，具有较高的工程应用价值。当然该方案还处于初级使用阶段，算法功能还不够完善。因此，如何提高对多种网页结构的适应性，完善算法自动化和智能性，同时减少算法复杂性，是今后的主要研究方向。

参考文献：

[1] 王 雷，陈治平，李志成. 基于文本分块的多模板隐马尔可夫模型的文本信息抽取[J]. 山东大学学报（理学版），2006，41（3）：25.

[2] 王 昕，熊光楞. 基于本体的设计原理信息提取[J]. 计算机辅助设计与图形学学报，2002，14（5）：429.

[3] 王 煜，徐建明. 基于RBF神经网络和决策树的文本分类方法[J]. 计算机工程与应用，2005，42（14）：175.

[4] Liu B., Grossman R., Zhai YH, Mining Data Records inWeb Pages[C].Proceedings of the Knowledge Discoveray and Data Mining（KDD）2003：601.

[5] Laender A H F,Ribeiro- Neto B A,Da Silva A S,et al.A Brief Survey of Web Data Extraction Tools[J].SIGMOD Record，2002，31（2）：84.