DOI:10.16208/j.issn1000-7024.2008.07.026

基于后缀树的 Web 论坛信息抽取

肖建鹏[1]， 张来顺[1]， 任 星[1]， 宋晓光[2]
(1. 解放军信息工程大学 电子技术学院，河南 郑州 450004；2. 中国人民解放军 65012 部队，辽宁 沈阳 110101)

摘 要：针对现有网上论坛信息抽取的不足，提出一种基于后缀树的论坛信息抽取方法。将标准化后的 HTML 文档转换为后缀树，查找出其中的重复模式并产生分装器，将分装器转换为 NFA(非确定型有穷自动机)达到抽取论坛信息的目的。该方法运用构造后缀树的技术来抽取论坛信息，较好地解决了现有的抽取方法准确性较差、通用性不强的问题。实验结果表明，该方法具有较高的准确性和实用性。

关键词：信息抽取； 分装器； 后缀树； 重复模式； 论坛
中图法分类号：TP311 **文献标识码**：A **文章编号**：1000-7024 (2008) 07-1675-03

Information extraction for web forum based on suffix tree

XIAO Jian-peng[1], ZHANG Lai-shun[1], REN Xing[1], SONG Xiao-guang[2]
(1. Institute of Electronic Technology, PLA Information Engineering University, Zhengzhou 450004, China;
2. China PLA Troop 65012, Shenyang 110101, China)

Abstract: Aimed at the limitation of the current methods to extract the web forum information, an information extraction method for web forum based on suffix tree is proposed. First, the HTML files standardize is converted to the suffix trees, then check to find out the repeat mode and build the wrapper, finally the wrapper is converted to the NFA to attains the aim of extract the web forum information. The method uses the suffix tree technology to extract the web forum information. The method has more accurate and applicability. The experimental result shows this method has high-accuracy veracity and practicability.

Key words: information extraction; wrapper; suffix tree; repeated pattern; forum

0 引 言

Web 论坛的出现为用户提供了信息交互的平台并且已经日益成为一种主要的交流方式。随着 Web 论坛中的用户不断增多，论坛中积存了大量的信息资源，因此急需有效的信息抽取和检索功能来支持对论坛信息的抽取。

对 Web 论坛的信息抽取不完全同于一般的信息抽取，主要目的是抽取论坛中用户所需要的内容而不是抽取细粒度的数据。在这方面，已经有一些方法被提出，其中有一类方法是注重基于领域 Ontology 的网页信息抽取。文献[1]首先通过领域 Ontology 解析器解析领域 Ontology 得到一系列概念和关系集，然后由规则生成器根据 Ontology 的概念关系和语法规则生成标注规则，接着将要处理的文档输入语法分析器进行语法分析处理，随后再将处理过的文档送入信息标注器，根据标注规则和语法分析的结果对文档进行信息标注，最后再使用信息抽取器对标注好的文件进行信息抽取。但是现在的领域 ontology 基本上是展现出来供标注过程使用的，而无法自动接收标注完的反馈信息。因此，该方法使用的效率低，需要增加机器学习的方法加以完善。

还有一类方法依据“对同一个论坛的主题无关的部分常常有着相同的内容和表现风格”这样的事实来注重于探测同一论坛网页中的一般模式。具有代表性的是 Lin 和 Ho[2]提出的系统(Info Discover)首先根据 TABLE 标签把网页分成若干个内容块，然后将词作为特征抽取出来并计算每个词的熵值，进而计算每个内容块的熵值，最后通过设定熵的阈值来划分有关和无关的内容块。尽管提高了效率，但都是其系统只针对单一的站点，有一定的局限性。其它一些现存于 Internet 网上，用于普遍搜索的引擎如 Google、YAHOO 和 Sohu 等以及为论坛而特殊设计的系统如 Lycos Discussions，都只是简单的检索和分类出论坛中的每一个网页，而抽取出用户所需的论坛信息作则显得力不从心。

针对以上不足，本文提出了一种简单有效的方法。该方法是根据以下的设想提出的：网页中的有用信息往往位于具有特定排列方式和次序的结构当中。特别是由搜索引擎产生的搜索结果通常是有规律的重复的模式。因此抽取重复模式可以发现对包装器有用的抽取规则[3-5]。

收稿日期：2007-05-02 **E-mail**：betret@sohu.com
作者简介：肖建鹏 (1979－)，男，辽宁锦州人，硕士研究生，研究方向为 Web 挖掘、信息抽取； 张来顺 (1963－)，男，河北安国人，教授，硕士生导师，研究方向为计算机应用技术； 任星 (1982－)，女，重庆人，硕士研究生，研究方向为数据库安全； 宋晓光 (1978－)，男，辽宁本溪人，助理工程师，研究方向为数据库。

1 体系结构

本文提出的抽取引擎是信息搜索引擎中最主要的部分，允许用户通过搜索引擎抽取出所有感兴趣信息。图 1 是整个搜索引擎体系结构图，主要包括以下几个部分：网络爬虫和网页分类器，抽取模块和数据查询模块。本文涉及到的是其中的抽取模块，由分装器产生器、分装器数据库和抽取器组成。

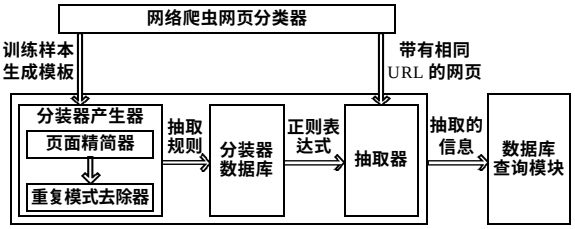

图 1 搜索引擎体系结构

网络爬虫不断的从 Internet 上发现和搜索 HTML 网页，并且利用一个经过学习的网页分类器来判断每一个获得的网页是否为论坛网页，以达到初步的网页过滤。经过滤后的网页被送入由分装器产生器、分装器数据库和抽取器构成的抽取模块，由分装器产生器对 HTML 网页进行学习，同时还要发现网页中重复的模式。分装器产生器生成分装器，并将以规则描述的形式存储在分装器数据库中。当一个带有相同 URL 前缀的网页再次到达抽取模块时，只需直接查找分装器数据库中相应的规则而不需要再运行分装器产生器。信息的抽取方法是通过对 HTML 网页进行模式匹配来完成的。抽取后的数据被定向到文本数据库中构造索引以便有效的支持检索。

2 数据抽取功能的实现

抽取模块的抽取过程分为两个步骤：①学习网页结构并生成相应的网页分装器；②使用生成的分装器抽取网页内的信息。

2.1 分装器的产生

现今，Internet 上的信息发布主要还是以 HTML 的形式为主，而 HTML 页面格式编排不合理的结果是使现有的 Web 浏览器在进行 HTML 语法分析时非常不严谨，因此本文的分装器产生器采用 XML 技术规范进行处理。分装器产生器包括两个部分：页面精简器和重复模式发现器。通过分析网页结构来获取网页的原始内容从而产生分装器。具体的工作流程为：经过滤的网页被送入后，首先由页面精简器对页面进行精简处理，然后由重复模式发现器在此基础上构造一个符号化的后缀树，再使用本文提出的方法进行重复模式的查找。

2.1.1 页面精简器

论坛网页中常常包含很多的图片、字体设置和动态脚本语言等修饰信息，这些与论坛帖子中浏览和回复的信息关系不大，但会极大的增加抽取计算量，因此抽取前要对原始页面进行精简。精简的方法是使用 HTML Tidy 提供的标准类库，将 HTML 文档转换为 XHTML 文档(XHTML 文档为 XML 的子集，符合 XML 规范，格式良好)初步实现轻量级的 Web 数据抽取。精简后的页面实质上就是把 HTML 标签和标签间的文本作为标记串的符号化的 XHTML 页面。接下来要实施的过程就是在该 XHTML 文档上也就是对标记串进行数据抽取的过程。

2.1.2 重复模式发现器

重复模式发现器是在精简后的页面上构造一个符号化的后缀树查找其中重复的字符串(重复模式)。后缀树是一种数据结构。一个具有 m 个单词的字符串 S 的后缀树 T，就是一个包含一个根节点的有向树，该树恰好带有 m 个叶子[6-7]。构建长度为 m 的字符串 S 的后缀树，首先将后缀 S[1..m]作为一条单边加入到树中。然后将后缀 S[i...m]加入到成长的树中，其中 i 从 2 增长到 m。考虑到后缀树中的循环总是以一个头标签为开始，所以在构造后缀树的过程中仅仅将带有头标签的子串插入到后缀树即可。这样构造的后缀树减小了规模，也相应的缩减了遍历后缀树的时间，提高了抽取的效率。为查找后缀树中重复的模式，需要遍历后缀树和每一个非叶子节点以便检查其所有的孩子节点是否有连续的子串能被发现。例如标记串为<body><table><tr><td>text</td><td>text</td><td>text </td><td></td></tr></table></body>，图 2 为标记串对应的后缀树。在图 2 的后缀树节点 3 下发现的 3 个子串，它们都以<td>开始并且是连续的。这样，它们就构成了一个连续重复的模式。也就是说被发现的重复串为<td>text</td>，输出到分装器数据库的标记串变为<body><table><tr>(<td>text</td>)*<td></td></tr></table></body>。

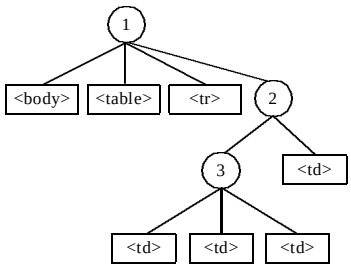

图 2 标记串对应的后缀树

具体算法描述如下：

输入：长度为 m 的字符串 S[1…m]

输出：后缀树字符串

Root=S_1 //定义根节点

Insert_Node(S_i) //插入字符 S_i

Check_Node(N(S_j)) //检查节点 N(S_j)，N(S_j)为字符 S_i 所对应的节点

if N(S_i)=True //如果节点 N(S_i)已存在，其父节点有后缀连接

Root=Parent_Node(S_i) //将节点 N(S_i)的父节点作为根节点

Insert_Node(S_{i+1}) //插入新节点

else N(S_i)=New //节点 N(S_i)在插入 S_i 过程中产生，其祖父节点有后缀连接

Root=Grandparent_Node(S_i) //将 N(S_i)祖父节点作为根节点

Path=P(Grandparent_Node(S_i),Parent_Node(S_i)) // N(S_i)父节点和祖父节点之间的串

UL=Check(UnLeafage(Path)) //查找 Path 之间的非叶

子节点

Insert_Node(S_{i+1}) //从节点UL以后查找并插入新节点

repeat //重复上述的过程

Emark(T(S)) //对后缀树的节点进行标记

if (Emark(T(S))=True)

Check_Small(S_i)=LL'(S_i) // LL'(S_i)表示节点 S_i 下含有最少叶子节点 S_k'到 S_i 的串长度 k

if ((S_t+D(S_i))在 LL(S_i)中) // LL(S_i)为节点 S_i 的叶子列表，S_t 为 LL'(S_i)中的节点，D(S_i)为 S_i 的深度

if (S_t!=S_t+2D(S_i))

if ((S_p-D(S_i))在 LL(S_i)中) //S_p 为 LL'(S_i)中的节点

Find_repeated_pattern(S_t, 2D(S_i)) //发现重复模式串

以上的算法所找出的重复模式仅仅是在标签层次上的发现，并不能确定头标签和尾标签之间的文本串是否重复。为发现这样的重复文本串，需要扫描所有作为重复结构中的文本串I，如果发现一些常见的单词G不断的出现在文本串I中，则G为一个重复文本串。例如存在两个重复串User1,Name: Tom和User2,Name:John ，其中Name重复的出现在串中，所以我们认为 Name 是一个重复文本串。考虑到写入分装器模板的可选元素总是出现在重复模式中，因此为了发现这种可选元素，只需要将重复模式同它邻近的标记进行比较，如果标记相似，则可以得出两个模式之间的不同应归属于可选元素。

2.2 数据抽取

分装器产生后，清洗过的 XHTML 页面和分装器(正则表达式的形式)被送入抽取器。在这里，我们利用一个 NFA 来进行信息抽取。抽取过程分为两步：①将分装器转换为一个 NFA；②构建一个 NFA 抽取器来抽取信息。本文使用 Thompson's 算法把分装器(正则表达式的形式)转换成 NFA。Thompson's 算法是将正则表达式转换为 NFA 的一种算法。输入为字母表 $\sum$ 上的正规式 r，输出为接受 L(r)的非有穷状态自动机 N。方法是先分解 r，然后构造非确定型有穷自动机 N。通过 NFA 转换器转换的 NFA 连同 XHTML 页面一起被送入 NFA 抽取器。在遍历 NFA 的同时通过分装器使用匹配的方法绘制出原始页面的内容。一旦发现匹配，NFA 抽取器则由原路返回。当一个文本串在原路返回时被发现，则完成数据的抽取。

3 实验结果与分析

信息抽取的主要评价指标是召回率 (REC) 和准确率 (PRE)，召回率等于系统正确抽取的结果占所有可能正确结果的比例；准确率等于系统正确抽取的结果占所有抽取结果的比例。为了综合评价抽取引擎的性能，通常还计算召回率和准确率的加权几何平均值，即 F 指数，它的计算公式如下[8]

$$F=(\beta^2+1)PR/\beta^2P+R$$

式中：β ——召回率和准确率的相对权重。β等于 1 时，二者同样重要；β大于 1 时，准确率更重要一些；β小于 1 时，召回率更重要一些。本文选取国内 3 个著名的论坛进行测试，它们分别是新浪论坛、搜狐社区和网易论坛。当取β = 1 时我们对每个论坛抽取 50，100 和 200 这 3 个数量的话题进行测试，如表 1 所示。

表 1　3 个论坛的抽取测试结果

论坛名称	抽取话题数量	可能正确结果	所有抽取结果	正确抽取结果	P(%)	R(%)	F
新浪论坛	50	33	66	33	100.00	100.00	1.0000
	10	66	66	66	100.00	100.00	1.0000
	200	128	127	127	99.22	100.00	0.9961
搜狐社区	50	35	35	35	100.00	100.00	1.0000
	100	65	65	65	100.00	100.00	1.0000
	200	121	121	121	100.00	100.00	1.0000
网易论坛	50	41	38	38	92.68	100.00	0.9620
	100	84	76	76	90.48	100.00	0.9500
	200	156	142	142	91.63	100.00	0.9530

从表 1 不难发现，本方法无论是召回率还是准确率都能够达到一个较高的比例。由此可以看出，对于多种风格类型的论坛站点，本文提出的抽取方法都能正确地学习抽取规则并完成抽取任务。

4 结束语

本文提出一种面向网上 Web 论坛的信息抽取方法。首先把网页编码成标记字符串，然后使用后缀树从标记字符串中发现重复模式并产生规则，最后将规则转变为 NFA，从而达到准确抽取出 Web 论坛信息的目的。该方法无需人工干预，具有自动、可靠和高扩展性，可以用于不同风格的论坛。

参考文献:

[1] 陈兰,左志宏,熊毅,等.一种新的基于 Ontology 的信息抽取方法[J].计算机应用研究,2004,21(8):155-157.

[2] Lin Shian-Hua, Ho Jan-Ming. Discovering informative content blocks from web documents[C]. Alberta, Canada: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002:588-593.

[3] Arasu Arvind, Garcia-Molina Hector. Extracting structured data from web pages[C]. San Diego, California: Proceedings of the ACM SIGMOD International Conference on Management of Data table of Contents, 2003:337-348.

[4] Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo. RoadRunner: Towards automatic data extraction from large web sites [C]. Roma: Proceedings of the 27th International Conference on Very Large Data Bases, 2001:109-118.

[5] Wang Jiying, Lochovsky F. Data extraction and label assignment for web databases[C]. Proceedings of the 12th International Conference on World Wide Web. New York: ACM Press, 2003: 187-196.

[6] 张吉.基于后缀树模型的流文本表示研究及其应用[D]. 北京:中科院计算所,2005.

[7] Stoye Jens, Gusfield Dan. Simple and flexible detection of contiguous repeats using a suffix tree [J]. Theoretical Computer Science, 2002,270: 843-850.

[8] 李保利,陈玉忠,俞士汶.信息抽取研究综述[D].北京:北京大学计算机科学与技术系计算语言研究所,2003.