基于重复模式的论坛信息抽取研究

韩 普[1,2],王 泽[2]

(1.南京大学 信息管理系,江苏 南京 210093;2.南京师范大学 教育技术系,江苏 南京 210097)

[摘要] 针对现有网络论坛信息抽取的不足,提出了一种基于重复模式发现算法的论坛信息抽取方法.该方法首先利用 SgmlReader解析器将 HTML文档转换为格式规范的 XHTML文档,然后通过计算 XHTML文档结构中 DOM子树相似度,自动发现论坛页面结构的重复模式.该方法通过自动定位重复模式进行论坛信息抽取,较好地解决了在论坛信息抽取过程中需要人工查找、定位重复模式或者通过人工分析论坛页面代码定制抽取规则的问题.试验结果表明,该方法具有较好的准确性、通用性和实用性.

[关键词] 重复模式,论坛抽取,信息抽取

[中图分类号] TP391 [文献标识码] A [文章编号] 1672-1292(2010)03-0074-04

Information Extraction for Web Forum Based on Repeated Pattern

Han Pu[1,2], Wang Ze[2]

(1. Department of Information Management, Nanjing University, Nanjing 210093, China;
2. Department of Educational Technology, Nanjing Normal University, Nanjing 210097, China)

Abstract: Aiming at the limitation of the current method to extract the web forum information, this paper introduces an information extraction method for web forum based on repeated pattern discovery algorithm. This method used SgmlReader parser to convert the HTML document to XHTML document firstly, and then calculated the similarity between the DOM trees that is in the XHTML document, and automatically found the repeated pattern from the forum pages. The method solved the problem that people have to manually locate the repeated pattern or manually analysis page source code for the extraction rules. The experimental result shows that this method has high accuracy, good universality and practicality.

Key words: repeated pattern; forum extraction; information extraction

随着信息社会的快速发展,网络论坛在人们的生活和学习中扮演着越来越重要的角色.论坛站点已成为信息化社会的重要组成部分.随着论坛的用户数不断增加,论坛中积存了大量的信息资源,论坛的信息抽取成为 web信息抽取的重要组成部分.因此急需有效的信息抽取和分析方法来支持对论坛信息的抽取.

网络论坛正日益受到学术界和社会各界的广泛关注[1].文献[2]通过对论坛建立网站地图的方式实现了论坛的信息抽取,有效地避免了重复链接的问题.文献[3]根据论坛的层次结构特点,设计开发针对论坛结构的爬虫程序.文献[4]根据要抽取的目标论坛设计了一些有效的遍历策略.著名的搜索引擎站点 Google在 2003年就推出了网络论坛的检索功能. Yahoo和 Sohu都为论坛而专门设计了搜索系统.专门的论坛搜索引擎有 Lycos Discussions Search, Qihoo论坛搜索、Teeinn中文论坛搜索引擎等.这些搜索引擎都只是简单的内容检索,存储的是论坛的页面,满足不了对论坛数据进行分析的需要.

网络论坛的信息抽取不同于一般网页的信息抽取.在网络论坛中,对于同一类型的页面,论坛系统采用同一模板生成,相同类型的页面结构非常相似.文献[5]实现了一个面向网上论坛的信息抽取系统,采用了基于 DOM(DocumentObjectModel)树和 HTML页面结构的方法,对网上论坛进行信息抽取,抽取出论坛帖子的“消息、发信人、发布时间、标题、内容”等属性,该系统通过人工查找有关的节点生成规则.文献[6]通过人工分析论坛网页源代码来制定、修改和添加抽取规则,程序根据抽取规则进行论坛信息的抽

收稿日期: 2010-02-20

通讯联系人: 韩 普,博士研究生,研究方向:信息抽取, web挖掘. E-mail: hanpu0725@163.com

取. 这些论坛抽取研究均需使用者有一定的专业知识才能进行使用, 不够简便性和通用性. 针对以上系统的不足, 本文提出了一种更为高效简便的论坛抽取方法. 该方法充分利用论坛页面结构的特点自动定位需要抽取的信息单元, 用户仅需选择需要抽取的帖子属性即可自动生成抽取规则, 不需专业知识也可抽取论坛信息. 该方法首先将文档规范化处理, 然后通过计算两颗标记子树的相似度来自动发现论坛结构的重复模式完成论坛信息的抽取.

1　基于重复模式发现的信息抽取

1.1　相关的概念说明

版面列表页面: 在论坛站点中, 为方便用户直接进入自己感兴趣的版面参与讨论, 通常会将论坛进行版面划分, 有些较大的论坛就有数十个版面.

主题列表页面: 在论坛站点中, 主题列表页面是同一版面中主题帖子标题的列表, 列表中的文章标题同时作为热字提供到相应的帖子内容页面的链接. 除了帖子的标题作为链接之外, 主题列表页面一般还会有帖子的浏览次数、回复次数、发帖人、发帖时间等跟主题贴相关的属性.

帖子内容页面: 在论坛站点中, 帖子内容页面显示某一主题贴及其回复的帖子, 帖子的标题、帖子内容、回帖人、回帖时间等都要在该页面抽取.

1.2　论坛页面结构的特征分析

从技术角度, 论坛页面有动态和静态两种. 不管是动态还是静态, 论坛的页面一般都由模板生成. 论坛页面结构的相似性不仅包括页面与页面的相似性, 还包括同一页面内信息块结构的相似性. 对于前一种相似性, 在基于模板的信息抽取中已有不少研究. 对于后一种相似性, 是本方法着重分析利用的对象. 如在论坛主题列表页面中, 每增加一个主题帖, 页面结构会增加一个重复结构模块, 我们把这种重复结构模块称为重复模式. 只要能发现论坛页面的重复模式, 通过用户在重复模式中选择需要抽取的属性, 抽取规则便可以自动生成, 如图 1所示. 图 1(a)是一个简化了的帖子列表页面的 DOM树, 除第一个 table节点外, 其它每个 table节点均为一个主题帖, 图 1(b)是将其中的一个 table节点展开后的 DOM结构. 通过对大量论坛结构分析表明, 每个主题贴 (或回帖、跟贴)有以下的特征: (1)每个主题贴 (或回帖、跟贴)都是一个相对独立的 DOM子树. (2)这些 DOM子树在同一父节点下, 并互为兄弟节点. (3)这些 DOM子树具有相同的内部结构, 变化的只是信息的内容.

图 1　主题列表 DOM 树结构图

Fig.1　DOM tree structure diagram of topic list

1.3　重复模式

当前, 网络论坛的信息承载主要还是以 HTMI的形式为主, 它是一种半结构化的描述语言, 语法结构较为宽松. HTMI最初的设计目的是为了显示, 所以在对 HTMI结构处理方面就不尽人意. 本文利用了一款开源的基于. NET平台的 Sgml Reader解析器, 该工具可以将 HTMI转换成格式规范的 XHTMI结构. XHTMI是 XMI语言的特例, 这样, 通过转换后, 我们可以像操作 XMI文件一样来操作 HTMI结构. 现在问题就是求解 XHTMI文档中节点子树的相似度问题. 对如何发现重复模式, 本文参照了文献 [7, 8] 中计算相似度的方法, 提出了计算论坛重复模式公式 (1), 该算法是同一页面内重复模式的发现算法.

$$s(t_1, t_2) = \frac{X_1^T \cdot X_2}{\| X_1 \| \| X_2 \|}, \tag{1}$$

t_1、t_2均为标记子树, X_1、X_2为标记子树 t_1、t_2的特征向量. 并且 $\| X_1 \| \| X_2 \| = (X_1^T X_1 X_2^T X_2)^{1/2}$.

该算法利用了余弦相似度的算法思想, 从几何的角度, 要求解的是 X_1和 X_2这两个特征向量的夹角. 在

这里使用二进制表示向量维度的特征值,即:当标记树具有某一特征,其值就为 1,为 0表示标记树没有某一特征.公式(1)中 $x_1^T \cdot x_2$表示 x_1和 x_2向量所共有的特征. $\| x_1 \| \| x_2 \|$表示 x_1和 x_2向量所拥有特征的总和.关键问题是如何构建子树的多维特征向量.如何表示标记树的特征向量,我们约定如下:

(1)节点特征维:树内所有节点按照从父节点到叶节点逐层进行特征标记,如图 1(a),对拥有同一父节点 td的 table兄弟节点,按照如下标记: /html/body/table/tr/td/table;对于两棵标记子树的共同路径部分,可以不写,只从子树的根结节点开始.

(2)属性特征维:如 class="skinRed"特征维表示为"/table@class="skinRed"".(这里省略子树前面路径部分).

(3)文本特征维:即对于标记树中文本内容的处理.如图 1(b)中文本维表示如"/table/tr/td/text()".

现以图 2两颗子树为例,分析的过程如下.

```
<table>
  <tr>
    <td class="1">发贴人</td>
    <td class="2">内容</td>
  </tr>
</table>
```

图 2　(a) t_1标记子树

Fig.2　(a) t_1 tagged subtree

```
<table>
  <tr>
    <td>
      <td class="q">时间</div>
      <td class="2">地点</div>
    <td>
  </tr>
</table>
```

图 2　(b) t_2标记子树

Fig.2　(b) t_2 tagged subtree

按照前面的规则约定,图 2两棵标记树特征如表 1所示.

表 1　特征维及值

Table 1　Feature dimensions and values

特征维	t_1	t_2	特征维	t_1	t_2
/table	1	1	/table/tr/td/div	0	1
/table/tr	1	1	/table/tr/td/div@class=q	0	1
/table/tr/td	1	1	/table/tr/td/div@class=2	0	1
/table/tr/td@class=1	1	0	/table/tr/td/div/text()	0	1
/table/tr/td@class=2	1	0	/table/tr/td/text()	1	0

根据公式(1),计算两颗子树的相似度如下: $s(t_1, t_2) = 3/\sqrt{6*7} = 0.46$.即上面两棵子树的相似度为 0.46.

在求解重复模式中,需要设定一个阈值,当两颗子树的相似度大于设定阈值时,加上其它规则,可以判定该子树通过模板生成,为需要的重复模式.在计算子树相似度时,需要一些规则判断:在提供学习训练论坛页面时,要选择主题列表页面的主题帖子数目超过一定数值(如大于等于 10个),这样可以有效过滤一些其它的干扰项;子树字节大小须超过一定的字符数(如可以限制子树的字符须大于 20个字符);此外,这些子树的根结点互为兄弟节点.有了这些限制,通过公式(1),便可以定位页面内的重复模式.

1.4　抽取过程实现

根据该算法思想,我们设计开发了论坛抽取系统.系统的流程如图 3所示:

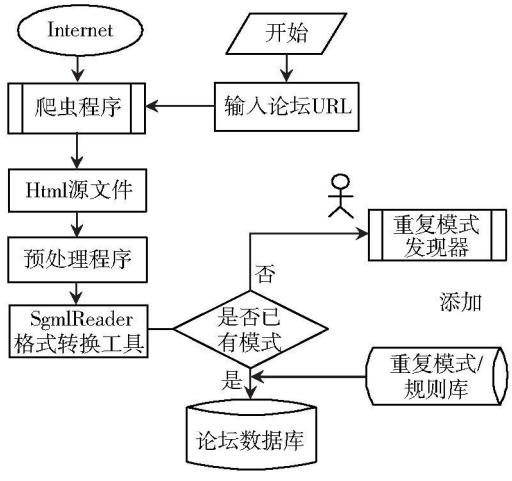

图 3　论坛抽取系统流程图

Fig.3　BBS extractor flow chart

2　实验与分析

为了验证实验模型在不同论坛结构的正确性和通用性,本系统测试了下面几个有代表性的论坛.为减小其它因素的干扰,本系统的抽取均在选定的版面下进行,并且没有考虑分页抽取的情况.实验重点抽取的页面包括主帖列表页和帖子内容页(阈值均设为 0.8).帖子主题列表抽取结果如表 2所示.

表 2　实验结果

Table 2　Experimental results

论坛	帖子主题列表数量	所抽取主题数量	正确抽取数量	召回率/%	准确率/%
清华大学网络学堂课程论坛	65	65	65	100.00	100.00
	231	230	230	99.57	100.00
CSDN论坛	50	50	48	100.00	96.00
	98.99	100	99	98	99.00
天涯社区	100	97	96	97.00	98.97
	98.49	200	199	196	99.50

从实验结果可以看出,对于清华大学网络学堂课程论坛的抽取准确率都达到 100%,而其它两个社会性质的论坛召回率和准确率稍低,其原因是因为课程论坛的页面结构较为简单一些,有关的干扰因素(如广告)较少,所以抽取的正确率和召回率都较高.要提高其它论坛召回率和准确率,需要根据具体情况添加有效的规则.

3　结语

本文提出了一种基于重复模式的论坛信息抽取方法.方法首先采用 SgmlReader解析器把网页转换为格式规范的 XHTML文档,通过计算 XHTML文档结构中 DOM子树相似度自动发现论坛页面结构的重复结构,即论坛的重复模式.并将重复模式保存进模式库.该方法解决了人工查找重复模式生成规则的问题,充分利用论坛页面内的结构特点自动发现论坛的重复模式,简便了论坛的信息抽取,提高了论坛信息抽取的通用性和准确性.为论坛的数据分析(如舆情分析、论坛发帖规律分析)提供了可靠的数据保障.

[参考文献](References)

[1] 王海明,韩瑞霞.目前国内 BBS研究现状评述[J].兰州石化职业技术学院学报,2004(4):25-29.
Wang Haiming, Han Ruixia. Review of present condition of domestic researches on BBS[J]. Journal of Lanzhou Petrochemical College of Technology, 2004(4): 25-29. (in Chinese)

[2] Cai R, Yang JM, Lai W, et al. iRobot: An Intelligent Crawler for Web Forums[C]// In Proc 17th WWW. Beijing: ACM, 2008: 447-456.

[3] Guo Yan, Li Kui, Zhang Kai, et al. Board forum crawling: A web crawling method for web forum[C]// In Proc 2006 IEEE/WIC/ACM Int Conf Web Intelligence. Hong Kong: IEEE, 2006: 745-748.

[4] Wang Y, Yang JM, Lai W, et al. Exploring traversal strategy for web forum crawling[C]// In Proc of SIGIR. Singapore: ACM, 2008: 459-466.

[5] 奚伟鹏,李昕,蒋凯.面向网上论坛的信息抽取技术[J].计算机工程,2005,31(4):66-68.
Xi Weipeng, Li Xin, Jiang Kai. Information extraction technology for web forums[J]. Computer Engineering, 2005, 31(4): 66-68. (in Chinese)

[6] 陈挺,刘嘉勇,夏天,等.基于平板型 Web论坛的信息抽取研究[J].成都信息工程学院学报,2009,24(2):1-4.
Chen Ting, Liu Jiayong, Xia Tian, et al. Information extraction research based on panel structured Web BBS[J]. Journal of Chengdu University of Information Technology, 2009, 24(2): 1-4. (in Chinese)

[7] Duda R O, Hart P E, Stork D G. Pattern Classification[M]. 2nd ed. Hoboken: John Wiley and Sons, 2000: 27-29.

[8] 杨少华,林海略,韩燕波.针对模板生成网页的一种数据自动抽取方法[J].软件学报,2008,19(2):209-223.
Yang Shaohua, Lin Hailue, Han Yanbo. Automatic data extraction from template-generated web pages[J]. Journal of Software, 2008, 19(2): 209-223. (in Chinese)

[责任编辑:刘　健]