文章编号：1003-0077(2010)02-0068-08

基于网页布局相似度的 Web 论坛数据抽取

王 允，李弼程，林 琛

（信息工程大学 信息工程学院，河南 郑州 450002）

摘 要：Web 论坛中蕴含着丰富的信息资源，充分利用这些信息资源依赖于论坛数据抽取技术。该文解决了从 Web 论坛抽取什么数据和如何抽取的问题，提出了一种基于网页布局相似度的 Web 论坛数据抽取方法，有效弥补了目前方法的自动化程度低，或准确率低的不足。该方法充分利用 Web 论坛网页布局结构上的特点，采用分级处理的方式，先识别出主题信息块、再利用待抽取数据的统计规律在主题信息块中完成抽取，整个过程不需要任何人工干预。实验结果表明，新方法对不同的 BBS 站点有很好的通用性，且具有较高的准确率和召回率。

关键词：计算机应用；中文信息处理；Web 论坛；数据抽取；相似度

中图分类号：TP391 **文献标识码**：A

Data Extraction from Web Forums Based on Similarity of Page Layout

WANG Yun, LI Bicheng, LIN Chen

(School of Information Engineering, University of Information Engineering, Zhengzhou, Henan, 450002, China)

Abstract: Web forums contain a wealth of information resources. Making full use of these information resources relies on web forums data extraction technology. This paper solves the problems of what data should be extracted and how to extract from web forums by the proposed method based on the similarity of page layout. The method can effectively avoid the disadvantages of current methods at low degree of automation or low accuracy. The method firstly recognizes the topic block by making full use of the special layout of the web forum pages, then extract data using rules from the topic block. Experimental results show that this method performs well in adjustability, precision and recall.

Key words: computer application; Chinese information processing; Web forum; data extraction; similarity

1 引言

随着互联网技术的不断发展，Web 论坛也由最初的电子布告栏(Bulletin Board System)系统日益成长壮大。目前，我国拥有近 140 万个 Web 论坛，2008 年底中国网络信息中心(CNNIC)的统计报告指出，论坛的使用人数达到了 9 100 万，占网民总数的 30% 以上。每天都有成千上万的人在不同的 Web 论坛探讨问题，交流观点，日积月累使 Web 论坛成为一个巨大的信息资源库。作为互联网的重要组成部分，对 Web 论坛的信息处理也逐渐被人们重视。针对 Web 论坛的应用日趋多样，如 Google、Baidu 等搜索引擎都提供了对论坛信息的检索；Web 论坛在网络舆情传播中的重要作用近年来受到了广泛的关注[1-2]；此外还有网络社区挖掘[3]等。

面对内容、形式多样的 Web 论坛，如何有效地抽取其中的数据是各种 Web 论坛应用的前提。Web 论坛数据抽取必须解决以下两个问题：(1)抽取什么数据；(2)如何抽取。话题是一个 Web 论坛的基本组成部分，话题隶属于不同的版块，由主帖和若干跟帖组成。我们采用元数据来描述话题，话题

收稿日期：2009-06-04 **定稿日期**：2009-11-03

基金项目：国家 863 计划资助项目(2007AA01Z439)；信息工程大学学位论文创新基金资助项目(BSLWCX200802)

作者简介：王允(1983—)，男，硕士生，研究方向为 Web 信息挖掘；李弼程(1970—)，博士，教授，研究方向为智能信息处理；林琛(1981—)，女，博士生，研究方向为网络数据挖掘。

的元数据有话题所属站点、所属版块、作者、标题、发表时间、帖子内容及各回帖的作者、回复时间、内容等。本文只讨论可以从论坛网页中直接抽取出的话题元数据如作者、标题、时间、内容等信息。利用元数据表示话题可以帮助我们更好的利用数据库来存储论坛数据，从而为后续的应用提供便利。

对于第二个问题，为了适应众多不同风格类型的网络论坛站点，我们旨在找到一种通用性强、自动化程度和抽取精度都比较高的方法。通过对 Web 论坛页面的生成机制、视觉效果、HTML 语法结构的分析，发现了论坛类网页所具有的普遍特征，总结出两点重要结论，并在此基础上提出了一种基于页面布局相似度的 Web 论坛数据抽取方法。本文以下部分将详细介绍该方法的相关内容。

2　相关研究和技术

Web 论坛数据抽取属于 Web 信息抽取中针对网页中某种属性的抽取，比如从新闻报道中抽取标题、正文内容、作者、发表时间等。目前 Web 信息抽取的方法多是基于规则，一般都是针对某一网站制定规则并以此构造分装器（Wrapper）实现自动抽取。典型的系统有 STALKER[4]、WHISK[5] 等。Wrapper 是一种软件构件，它主要通过两种途径来构建，一是知识工程的途径，即通过领域专家来制定抽取规则，这需要耗费大量的人力，成本很高；另外一种是采用机器学习的途径自动构建 Wrapper，根据标注样本，机器学习算法通过自动学习来建立抽取模型，这种方式仍然需要手工标注样本。总之，利用分装器的信息抽取技术都要在一定程度上依靠人工辅助，自动化程度比较低，其系统的适用性较差，目前只在比价购物方面的商业应用中比较成功。由于论坛网页形式多样且不断更新，因此，Wrapper 的维护成本较高，不适合大规模应用。

无监督的 Web 信息抽取主要针对含有多个数据记录的网页，比如产品列表页面，通过发现网页中的重复模式来确定数据区域，比较有代表性的算法有 MDR[6]、DEPTA[7]、NET[8] 等。此外，文献[9]提出了一种自适应的方法以适应更多类型的网页。由于论坛网页具有严格紧凑的结构，因此上述方法也可用于论坛网页。但是，这些方法都是基于编辑距离（Edit Distance）[10] 的字符串比较来发现重复模式。即将信息包含的 HTML 标签顺序连接成字符串，通过比较发现相似的字符串来确定待抽取的数据区域。但是网页结构复杂多变且存在很多局部噪声，仅仅通过字符串的比较将导致比较低的抽取准确率。文献[11]利用视觉特征和语言特征将网页划分成不同的内容块，再进一步确定主要内容块，文献[12]也是对网页进行块状分割，虽然其性能还有待提高，但是这种对网页进行分块处理的思想还是值得我们借鉴的。本文的方法从网页的布局角度出发，利用影响网页的布局结构的 HTML 标签，通过查找网页内部相似的布局结构来确定数据区域，能够有效提高抽取的准确率。

3　基于网页布局相似度的 Web 论坛数据抽取方法

待抽取的话题元数据包含在两种类型的网页中，一种是帖子的主题页面，用来列出帖子的标题并提供指向帖子内容页面的链接，图 1 是一个典型的例子。另外一种是帖子的内容页面，包含了主帖、回帖的内容，如图 2 所示。二者所含内容虽然不同，但通过分析可以发现它们具有相同的结构特点，因此可以采用同样的方式处理。

3.1　Web 论坛网页的布局结构特点

Web 论坛系统一般都在服务端使用 CGI 模块来动态生成同一功能类型的 HTML 页面，这些动态生成的页面往往采用相同的模板，在单个网页内或网页之间，相同类型的信息内容具有相似的视觉效果。因此，整个页面往往都很规整，结构上很紧凑。如图 1、图 2 中矩形框之间结构布局都是相似的，可以看出它们都包含了标题、作者、时间、内容等，并且这些内容在块中的位置都是固定的。从网页的 DOM 树结构同样可以看出这一点，如图 3 所示，图 1 中的矩形区域对应于 Table 节点下的每个 TR 节点，从图中两个展开的 TR 节点可以看到其对应位置的节点是相同的，而且其余的 TR 节点也有同样的结构。这些特征都明显区别于网页中的其他部分，可以帮助我们区分导航条、广告区、相关信息等网页的其他区域，从而有效的滤除噪声。

综合以上分析，我们总结得到以下关于 Web 论坛网页的重要结论：

（1）从视觉效果上看，主题页面和内容页面通常含有大量（通常都有数十个）布局结构相似的内容块。图 1 中这样的块有 71 个，这明显区别于页面的其他部分。我们把这些包含了论坛话题元数据的区

图 1 主题页面

图 2 内容页面

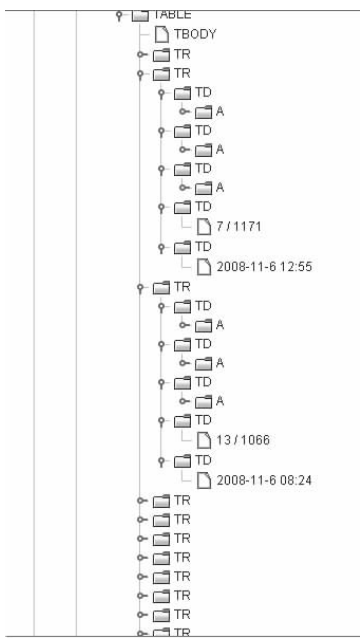


图 3 页面的 DOM 树结构

域称为论坛网页的主题信息块，如图 1 和图 2 中的矩形区域。

（2）从 HTML 语法结构上看，表示这些相似内容块的节点通常都是 Table、Div、UL，Form 等的子孙节点，并且都位于 DOM 树的同一个层次。如图 2 中相似的 TR 节点都是同一个 Table 的子节点。

3.2 数据抽取的具体实现过程

基于上述特点，本文采用分级处理的方式抽取论坛话题的元数据，主要包括两个步骤：

（1）网页级处理，滤除整体噪声，识别出各个主题信息块；

（2）区域级处理，滤除局部噪声，从主题块中提取出元数据。

其优点是将抽取范围缩小到一个小的区域（主题信息块），经过两层过滤提高了精度。下面介绍具体实现过程。

3.2.1 主题信息块识别

人们观察网页时总是能快速准确的定位自己感兴趣的部分，这在很大程度上得益于网页的块状布

局结构。整个网页被划分成各个不同语义内容的区域且一般都有固定的位置，比如导航区通常位于页面顶端，广告一般在两侧，而版权信息和相关链接一般在网页底部。这种布局的实现是由 HTML 语言的块状标签节点来控制的，如 Table、Div、Form 等，这些节点将网页分割成各个相对独立的区域。因此，在将网页解析成 DOM 树时只保留块状节点以提高运算的效率。

由前面分析可知，主题信息块的识别就是要找到网页中那些含有大量相似结构的区域，对应于 DOM 树，就是找到那些含有大量（和预先设定的值相比）相似子节点的节点。对每一个 DOM 树节点，可以通过计算其子节点两两之间的相似度判断是否相似，算法的伪代码如下所示：

```
输入：DOM 树节点 Node；
输出：含有相似子节点的节点集合 NodeSet；
Algorithm FindSimilarChildren (Node)
{
    获得 Node 的 length 个孩子节点 childNode；
    设定常数 K 为满足要求的相似节点个数；
    //两两比较，得到相似节点个数
    for(int i =0;i<length-K;i++)
    {
        int C=0; //相似节点计数
        for(int j =i;j<length;j++)
            if(childNode[i]和 childNode[j]相似)
                C++;
        If(C>K) //K 一般取大于 2 的常数
        {
            将 Node 添加到 NodeSet 中；
            Break; //结束循环
        }
    }
    if(未在 Node 中找到满足要求的相似子节点)
        for(每一个 childNode)
            FindSimilarChildren (childNode)；
}
```

由于主题信息块的个数通常是最多的，因此可以简单的判定 NodeSet 中含有最多孩子节点的节点即为我们要找的节点。实际使用中 K 可取较大的值（主题信息块的数量通常都较大）。

上述算法中的关键是判断两个节点是否相似。通常我们观察两个物体是否相似时总是习惯先从整体上把握，比如大小、形状是否相同，再分析其内部的特征，比如质地、棱角等。同样观察网页也是如此，是一个从宏观到微观的过程，我们往往先看整体上的结构布局，再逐步深入内部细节。因此，我们定义节点的结构相似度 SoL（Similarity of Layout）来度量两个节点在布局结构上的相似程度。SoL 是 0 到 1 之间的值，越接近 1 表示两个节点的布局结构越相似。设有两个节点 x，y，则它们的 SoL 定义为：

$$SoL(x,y)=\sum_{i=1}^{N}\omega_i\sum_{j=1}^{M_i}\frac{1}{M_i}S_{ij} \tag{1}$$

其中 N 表示比较的深度，即只比较到第 N 层节点；M_i 表示第 i 层子节点的个数；ω_i 为第 i 层子节点对整体结构布局的贡献系数，一般认为越深层次的节点对宏观布局的影响越小，因此它们对应的 ω_i 值就越小，其反映的更多的是节点间细节上的差异，显然 ω_i 应该满足 $\sum_{i=1}^{N}\omega_i=1$。S_{ij} 表示进行比较的两个节点的第 i 层第 j 个节点是否为同种类型的块状节点，其值取 0 或 1。首先判断两个节点是否为使用了同样的 HTML 标签，比如是否同为 TR 或 TD 标签，若不同，则 S_{ij} 为 0；若相同，则继续比较两节点属性是否相同，如 width，style，align 等能反映节点布局结构的属性，若这些属性值也相同，则 S_{ij} 为 1。

实验表明，在 N 取 3，ω_i 分别取 0.6，0.3，0.1 时，若 SoL 值大于 0.9 则可认为两个节点相似，此时可以达到较好的识别效果。如果要求的精度更高，则可以适当的增大 N 的值，同时调整 ω_i 的取值，此时相应的 SoL 值也要更大才能判断节点是否相似。实际上主题信息块之间确实存在一些局部细节上的不同，但这并不影响它们之间的整体相似性。

3.2.2　元数据提取

元数据在主题信息块中的位置是相对固定的而且数据本身各有特点，如查看回复数为阿拉伯数字；时间一般有固定的格式；正文通常含有较多文本且其中很少出现超链接；标题链接的锚文本长度一般要大于其他链接的锚文本长度；作者一般都表现为链接的形式用于指向其主页并且有时会出现“作者”、“Name”等前缀。这些信息都有助于我们正确提取数据，但有些情况不能忽视，比如有些回帖的内容很短，甚至只有一个字；正文中出现的数字，时间；表示作者的链接锚文本可能长于标题的链接锚文本；另外主题信息块中仍有少量噪声，比如一些广告链接和功能性链接等，这些情况都将严重影响抽取的准确性。

由于主题信息块之间是相似的，相同的内容都有相同的表现形式，会表现出一定的统计规律性，比如若某一位置所有的主题信息块中都出现时间则可

认为是时间，从而区别于个别正文中出现的时间。因此，我们考虑所有的主题信息块。

将主题信息块表示成具有明显语义信息的节点的集合，比如文本节点、超链接、图片等，其他节点不予考虑。其中第 i 个主题信息块 B_i 表示为如下形式：

$B_i=\{n_1,n_2,n_3\ldots n_k\}$，　n_i 代表各语义节点。

采用深度优先的方式遍历主题信息块中的所有节点，按照下面的步骤得到 B_i：

(1) 获取下一个要处理的节点。若为空，结束。否则转至(2)。

(2) 若当前节点的子节点只含有文本节点或链接节点，则将其添加到 B_i 中，转至(1)。

由上述方法将所有主题信息块表示成语义节点的集合。先对 B_i 中的节点进一步过滤，若所有的 B_i 中节点 n_i 都相同，则认为 n_i 是噪声节点，再应用以下规则从 B_i 中抽取代表元数据的节点：

R1：对所有 B_i 中对应文本节点求出其长度的平均值，最大者为正文。

R2：对所有 B_i 中对应链接节点求出其锚文本长度的平均值，最大者为标题。

R3：所有 B_i 中对应某节点其文本中均含有数字则为查看回复数。

R4：所有 B_i 中对应某节点其文本中均含有一定格式的时间字符串则为时间。通过大量调查我们搜集了数十种 Web 论坛中经常使用的时间表达形式。

R5：若所有 B_i 中对应某节点其文本中均出现“作者”、“Name”等字样则为作者，否则 B_i 中链接节点中位置靠前且锚文本长度较短的为作者节点。

4 实验部分

为了验证本方法对于各种 Web 论坛系统的自动抽取性能，我们选取了 100 个论坛站点作为数据来源，这 100 个站点包含了目前比较有代表性的中文论坛，同时还有一部分英文论坛。论坛的类型多种多样，有综合性的门户网站的论坛如新浪、搜狐、网易等，也有专业性较强的如法制论坛、CSDN 社区等，还包括一些地方性论坛如北方论坛、福州论坛等。每个论坛均选取若干版块，内容涉及政治、经济、军事、体育、娱乐等不同领域，利用网络蜘蛛程序共抓取网页 21 088 篇，其中主题页面 8 136 篇，内容页面 12 952 篇。选取部分页面作为实验数据用于方法中参数的确定，其余页面作为测试集来验证方法的性能，数据组成如表 1 所示。实验内容主要包括参数选取、测试集上的抽取性能以及分类实验。

表 1　实验数据和测试数据

	测试网页数	实验网页数
主题页面	7 936	228
内容页面	12 652	272
总计	20 588	500

4.1 参数选取

方法中主要用到的参数有布局相似度阈值，计算布局相似度时需要比较的层数以及每一层对整体布局的贡献系数。其中，比较层数 N 的选取最为重要，因为它影响到其他参数的选择，主要须考虑以下因素：层数偏少，即比较的过于粗略则可能得到很多相似的节点，影响最后抽取的准确度；层数偏多，即比较的过于细致则可能得不到正确的结果并且所消耗的时间会更多。因此，需要综合考虑使得抽取的准确率和召回率都比较高，同时运算处理速度又快。贡献系数的选取要依据以下原则：外层的贡献系数要大于内层，即要满足以下条件：

$$\sum_{i=0}^{N-1}\omega_i=1,\quad \omega_i>\omega_j,\ 0\leqslant i<j\leqslant N \qquad (2)$$

在实验中，我们选择 $N=1,2,3,4,5,6$ 来分别计算处理每个网页的平均时间和抽取的召回率和准确率，而贡献系数的选取则根据式(2)，可根据实际情况作具体调整，这里只给出 $N=3$ 时的一组参考值(0.6,0.3,0.1)，此时布局相似度 SoL 为 0.9 时效果较好。具体实验结果如图 4 和图 5 所示。图 4 显示处理时间随着 N 的增大而不断增大最终趋于平缓，这是因为大部分的网页 DOM 树层数都在一定范围内，因此实际处理的层数不会随 N 的增加而无限增加。由图 5 可以明显看到，N 取太小或太大值时，准确率和召回率均不能达到令人满意的效果，这与我们的分析是一致的。综合考虑，N 一般取 2

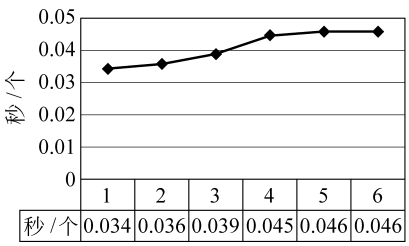
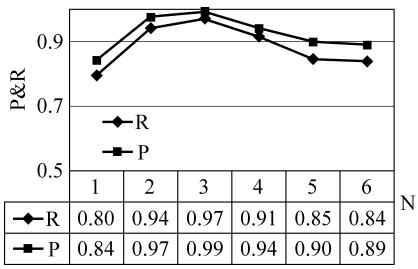


图 4　N 取不同值时的平均处理时间



	1	2	3	4	5	6
R	0.80	0.94	0.97	0.91	0.85	0.84
P	0.84	0.97	0.99	0.94	0.90	0.89

图 5　N 取不同值时的召回率和准确率

或 3，此时可在处理速度和抽取精度上达到一个较为平衡的状态。

4.2　抽取性能

(一) 总体实验

在参数确定后，我们在更大的测试集上来验证本方法的性能。按照本方法的流程，本部分实验按照以下步骤进行：

(1) 主题信息块的识别准确率和召回率。主题信息块的识别是本方法的关键，其性能直接影响到最终的抽取结果。我们将与目前比较流行的自动化抽取工具 MDR[4] 进行对比，采用准确率(REC)和召回率(PRE)[13] 来衡量抽取性能。结果如表 1 所示。

$$\Pr ecision = \frac{ACE}{AER}, \quad \mathrm{Re}\, call = \frac{ACE}{ACR} \qquad (3)$$

其中 ACE 表示所有正确抽取的结果，AER 表示所有抽取的结果，ACR 表示所有正确的结果。

(2) 元数据抽取准确率。在正确识别主题信息块的基础上，我们只关心元数据的抽取准确率。表 2 给出了实验结果。

最后，将上述两步的结果相乘可以得到最终的抽取准确率，它反映了本方法对 Web 论坛数据抽取的总体性能。结果如表 3 所示。

表 2　主题信息块识别结果对比

	总数	本文方法		MDR	
		P/%	R/%	P/%	R/%
主题页面	7 936	99.23	98.50	90.47	82.88
内容页面	12 652	98.02	97.26	85.36	79.15

表 3　元数据抽取准确率

	抽取元数据的平均准确率/%				
	作者	标题	时间	内容	查看回复数
主题页面	98.49	99.26	99.25	—	99.37
内容页面	98.38	—	99.21	98.12	—

表 4　总体性能(准确率/%)

	作者	标题	时间	内容	查看回复数
主题页面	97.73	98.50	98.49	—	98.60
内容页面	96.43	—	97.25	96.18	—

说明：表 2，3 中缺失项表示该项数据不在此类页面中或未从此类页面中抽取。

(二) 分类实验

为了进一步验证本方法的性能，我们对采集到的论坛网页进行更进一步的分类、细化，分别按照语种(中文和英文)、所涉及领域、不同时间段来组织网页进行实验。

图 6 是对中英文论坛的对比结果，从中可以看出，对英文论坛数据的抽取在召回率和准确率上都要高于中文论坛。通过对比网页我们发现，这主要是由于相对于一些中文论坛，很多英文站点的网页设计都很简洁，语法也相对更规范，同时广告等噪声信息较少，这在很大程度上有助于提高抽取的精度。

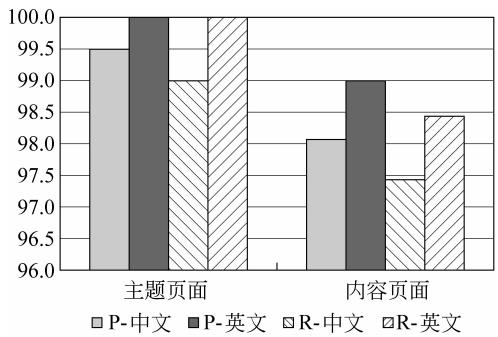


图 6　中英文论坛对比

考虑到有些网站可能会改版，因此我们对采集到的网页按照不同的时间段来分类以检验本文方法的适应性。将网页按照月份来分类，共分为 4 月～9 月六类，结果如图 7 所示。可以看到，在不同的时间段仍然能保持较高的召回率和准确率，但在 6 月份以后的结果却有一定程度的下降。对比这两部分的数据我们发现两点不同：1)版面的调整，使用了不同的模板，显示方式有所不同；2)部分网站改用脚本程序来动态显示数据，我们下载到的网页与通过浏览器看到的效果不一样，真实数据被隐藏。对于第一种情况，本方法表现出良好的适应性，前后变化不大，真正影响抽取结果的是第二种情况，本方法几乎无法给出满意的结果，这也是我们改进的方向，毕竟动态网站是一种趋势，脚本技术更是被广泛应用。

网络论坛的表现形式多样，所涉及的领域也不

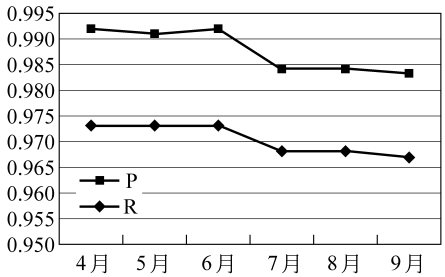


图7 不同时间段召回率和准确率的变化图

尽相同。我们对采集到的网页按不同领域来分类，主要包括综合类、军事类、技术类、娱乐类，其余网页数量较少的归为其他类，实验结果如图8所示，不同领域的抽取结果在召回率和准确率上均相差不大，这也说明了本方法可适用于不同类型的论坛。

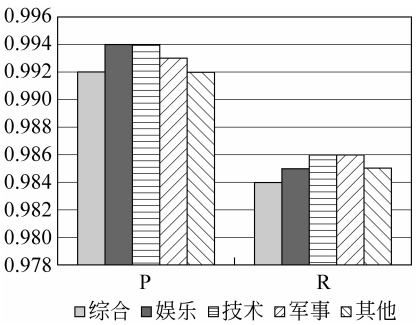


图8 按不同领域分类的抽取性能

4.3 结果分析

总体实验表明，本方法在数据样本量较大的时候依然能表现出良好的性能。从表2可以看出本方法对不同论坛站点的主题页面和内容页面中主题信息块的识别准确率和召回率都很高，这充分说明了论坛网页所具有的布局结构特点能很好的帮助我们正确提取信息，也验证了我们所得结论的正确性。其中对内容页面识别的准确率稍低，原因是一些内容页面中由于回帖数少导致相似的主题块个数少(有些甚至没有回帖)从而导致错误。对比结果显示，本文方法在准确率和召回率上都要明显优于MDR，其原因就在于本文方法只考虑了对布局结构有影响的HTML标签，从而能够有效滤除其他标签的影响，相对于MDR基于字符串比较的方式，受到网页内部噪声的影响小，因而准确率明显高于MDR，更适合用于对论坛网页的数据抽取。

表3显示在主题信息块中对元数据的抽取能达到较高准确率。这是由于时间、数量特征明显，很容易区分，而通过统计平均使得内容、标题、作者的准确率也能令人满意。表4说明了方法的总体性能良好，在没有人工干预的情况下可以达到实用要求。但是由于各站点对数据的组织形式不尽相同，而且一些站点使用脚本程序来实时动态的显示数据，这使得我们虽然能正确找到主题信息块，但对某些元数据的抽取效果不尽如人意。

分类实验进一步证明了本方法并不局限于对某一领域、某一时间段、某种特定语言的论坛数据抽取。由于其利用的是论坛网页普遍具有的布局结构特点，因而对于各种论坛站点均能表现出良好的适应性。

5 结束语

本文明确了论坛数据抽取的任务(抽取什么数据)并利用论坛网页结构布局结构上的特点，提出了一种全自动的数据抽取方法，该方法通过两级处理有效的滤除了网页噪声的影响，达到了令人满意的抽取结果。实验表明该方法具有较强的实用性，能适用于不同的论坛站点。但是，在一些方面仍然需要改进：由于很多网页大量使用客户端脚本程序来显示数据，使得仅仅对网页进行分析是不够的，还需要增加脚本执行功能以获取最终的数据；在确定主题信息块时仅仅根据数量的多少可能会导致错误，可考虑一些其他规则如主题信息块中的文本长度等；此外，为了更进一步地提高抽取准确率，还需要对更大规模的论坛数据进行分析，总结出适用于论坛数据抽取的规律。

参考文献

[1] 薛玮. 网络舆情信息挖掘系统的研究[D]. 北京：北京交通大学，2008.

[2] 姚晓娜. BBS热点话题挖掘与观点分析[D]. 大连：大连海事大学，2008.

[3] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms[C]//Proc. 16th WWW, Banff, Canada, May 2007. 2007:221-230.

[4] I. Muslea, S. Minton, C. Knoblock. A Hierarchical Approach to Wrapper Induction[C]//Third International Conference on Autonomous Agents, (Agents' 99), Seattle, May 1999.

[5] S. Soderland. Learning Information Extraction Rules for Semistructured and Free Text[J]. Machine Learning, 1999.

[6] Liu B., Grossman R., Zhai Y. Mining Data Records in Web Pages [C]//KDD 2003: 601-606.
[7] Z. Yanhong and L. Bing, Web Data Extraction Based on Partial Tree Alignment [C]//Proceedings of the ACM, 2005: 76-85.
[8] Liu, B. and Zhai, Y., NET - A System for Extracting Web Data from Flat and Nested Data Records [C]// WISE 2005, 2005: 487-495.
[9] Justin Park and Denilson Barbosa. Adaptive Record Extraction From Web Pages[C]//WWW 2007.
[10] Gusfield, D. Algorithms on strings, tree, and sequence[M]. Cambridge. 1997.
[11] 韩先培，刘康，赵军. 基于布局特征与语言特征的网页主要内容块发现[J]. 中文信息学报，2008，22(1)：15-21.
[12] 瞿有利，于浩，徐国伟，等. Web 页面信息块的自动分割[J]. 中文信息学报，2003，18(1)：6-13.
[13] 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.

“综合型语言知识库”再次获奖

北京大学计算语言学教育部重点实验室俞士汶教授等人研制的“综合型语言知识库”继获得 2007 年度教育部科技进步奖一等奖之后，再次荣获 2008 年度第 12 届北京技术市场金桥奖项目二等奖。

金桥奖是北京市科学技术委员会于 2003 年设立的，用于奖励北京市对科技成果转化做出突出贡献的集体和个人。与其他科技成果奖项不同，金桥奖特别关注科技成果转化所产生的效益。“综合型语言知识库”是中文信息处理领域的一项基础性研究成果，北京大学科技开发部之所以在北京大学诸多成果中选中它申报金桥奖，是因为它符合申报金桥奖的 12 个条件中的至少以下两个条件：

2. 通过开发、转让重大技术成果项目，特别是拥有自主知识产权的项目，对地区经济或行业技术进步产生重大影响的；

7. 大力开拓国际市场，积极开展国际技术交流和技术贸易，特别是在技术出口创汇方面取得显著成绩的；

“综合型语言知识库”所包含的各项语言数据资源、语言知识库和语言信息处理工具软件向境内外学术界和企业界广泛转让了许可使用权，包括国际著名的大学、研究机构以及 IT 界的大公司，取得了可观的经济效益，促进了计算语言学和自然语言处理技术的进步。“综合型语言知识库”已完成许可使用权的协议有偿转让 200 次左右，其中以其第一块基石《现代汉语语法信息词典》的转让次数最多，它的第一份协议签于 1996 年 2 月 2 日，最后一份于 2010 年 2 月 8 日生效，前后历时 15 年，还有新的协议正在探讨洽谈中。在 IT 领域，一项研究成果存活如此长的时间，确实难能可贵。

“综合型语言知识库”还在继续发展。国家重点基础研究项目(973)“数字内容理解的理论与方法”(2004－2009)于 2009 年 11 月结题时，将“综合型语言知识库系统”推荐为 3 项代表性成果之一。

期望学术界和企业界继续关注和扶植“综合型语言知识库”，期望它在以汉语为中心的语言信息处理技术的发展历程中发挥更多、更有效的作用。

北京大学计算语言学教育部重点实验室 供稿
2010 年 2 月 10 日