基于DOM树的可适应性Web信息抽取

李 朝 彭 宏 叶苏南 张 欢 杨亲遥
(华南理工大学计算机科学与工程学院 广州 510641)

摘 要 Web信息抽取通常采用的是一种归纳学习方法,从给定的训练样本网页中学习到抽取规则,这种方法虽然能够准确地抽取出信息,但是当网站的模版发生改变后,必须重新获得抽取规则,因而这种抽取器的维护成本比较高,可适应性差。提出一种新的可适应性Web信息抽取方法,该方法首先通过聚类方法获取商品在网页中频繁出现的关键词组,然后利用网页的DOM树结构来确定包含这些关键词的信息块,从而实现Web信息的自动抽取。对大量商业网站进行信息抽取的实验表明,该算法不仅能有效抽取出商品信息,而且是一种与站点结构无关的可适应性信息抽取方法。

关键词 DOM树,信息抽取,可适应性

Adaptive Web Information Extraction Based on DOM Tree

LI Zhao PENG Hong YE Su-nan ZHANG Huan YANG Qin-yao
(School of Computer Science, South China University of Technology, Guangzhou 510641)

Abstract Many Web information extraction methods are related to wrapper induction. It extracts the items by the rules learnt from the Web pages used for training. Although it can get the information accurately, it is hard to be maintained when the template of the Web site is changed, as it needs to learn the rules again. In our research, we put forward a new adaptive Web information extraction. It determines the block which contains all information about the merchandise by using the keywords of a certain topic, which is based on DOM tree structure. The experiments on a great amount of Web pages show that our method can not only extract the information efficiently, but also is irrelevant to the site structure, which can be widely used for many different Web information extractions.

Keywords DOM tree, Information extraction, Adaptive

1 Web信息抽取的技术

随着Internet的发展,Web已经成为一个全球的、巨大的、分布和共享的信息空间。目前Web上的数据大部分都是以HTML形式存在的。HTML是将内容描述和显示描述信息混和在一起,缺乏对数据本身的描述,语义信息不清晰。而Web信息抽取技术的核心是从Web页面所包含的无结构或半结构的信息中识别用户感兴趣的数据,并将其转化为结构化、语义更为清晰的格式。构建Web信息抽取器的方法有很多,这些方法大多数都是有监督的学习方法,通过学习样本网页,归纳出网页的抽取规则。但是,目前常见的信息抽取器存在的一个主要问题:从样本网页中学习到的抽取规则只适用于样本所在的网站。这些抽取器不仅维护成本高,而且可适应性差。因此建构一种与站点结构无关的可适应性Web信息抽取具有很大的现实意义和商业应用价值。

最近,有研究者提出了几种不同的可适应性信息抽取的方法,旨在站点结构发生变化时,能够扩展旧的规则或者学习到新的规则来自动抽取出Web信息,然而这些方法仍然需要很多人工工作来辅助。例如:文献[1]提出的IEKA方法需要识别出站点不变和可变的性质来获得站点的网页训练集;文献[2]提出的渐进式Web信息抽取,利用关联挖掘算法从训练样本网站中寻找内容关联知识,并利用这种关联知识来识别兴趣信息块,然后通过半指导式学习和无指导式学习,归纳出可以适用于同一领域的不同网站的抽取规则。该方法同样需要花费大量时间来训练样本。文献[3]提出的基于关键词聚类和节点距离的网页信息抽取方法只要获取到该领域的关键词就能有效地抽取信息,但是它必须分析网页的结构,而且仅仅只是针对HTML中的〈table〉标签来建立网页结构树,因此不能抽取出非〈table〉标签中的商品信息。基于文献[3]这种思想,我们提出一种基于DOM树的可适应性信息抽取方法,以对任何一种网页结构进行有效的信息抽取。

2 基于DOM树的可适应性信息抽取

文献[3]中提出的基于关键词聚类和节点距离的网页信息抽取方法是首先针对某特定领域确定关键词组,然后构造一棵以〈table〉标记为节点的二叉树,通过对包含关键词的节点进行聚类来获取待抽取的信息块,也就是中心节点。对于节点外的数据(比如图片信息)则根据节点的距离来抽取。

到稿日期:2008-08-25 返修日期:2009-01-23 本课题得到广东省自然科学基金(No. 07006474)资助。

李 朝 博士研究生,主要研究方向为数据挖掘、Web知识发现、机器学习,E-mail:lzjoey@gmail.com;彭 宏 教授,博士生导师,主要研究方向为人工智能。

该方法的一个核心部分就是建立以〈table〉标记为节点的二叉树，然而并不是所有的商业网站的信息都是以〈table〉结构来存储的，因此我们在此基础上提出了一种基于 DOM 树的可适应性信息抽取，从而商品信息无论以何种方式存储，都可以被准确地抽取出来。基于 DOM 树的可适应性信息抽取算法与基于关键词聚类和节点距离的网页信息抽取方法的基本步骤相似。首先采用分层抽取的算法来确定关键词组：针对某个领域的站点，将其分成几个类别，找出同一类的网页集合中频繁出现的关键词，然后确定在各个类别间频繁出现的关键词；其次通过关键词组确定待抽取信息在网页中的位置，即确定抽取的中心节点；最后通过搜索中心节点的最近邻来抽取节点外的信息。所不同的是我们的方法是将网页表示成 DOM 树结构，通过关键词在 DOM 树中的位置来确定待抽取的商品信息，而不是根据〈table〉标签来建构网页二叉树，通过关键词聚类来确定中心节点。我们的方法与站点结果无关，具有很强的适应性。下面我们具体介绍如何利用 DOM 树来获取中心节点以及节点外的信息。

2.1 DOM 树的建立和中心节点的确定

一个 Web 页面可以表示成文档对象模型，即 DOM(Document Object Model)结构[4]。一个 DOM 结构是一个包含两种类型节点的有序树，一种是元素节点(element node)，另外一种是文本节点(text node)。元素节点是用来表示 HTML 的表示信息，如"〈table〉","〈tr〉","〈td〉"等等，而文本节点则是用来表示页面中的文本信息。如图 1 所示的网页 HTML 文本可以表示为图 2 所示的 DOM 树结构。在 DOM 树中，每一个文本节点都对应一条从根节点到叶子节点的路径，路径上所有的元素节点都是文本节点的祖先。

```
<table>
<tr> <td>
<b><font face="Verdana, Arial" size= "2"
color="#666666">Microsoft Excel </font>
</b>
Author: John
<b> Published : </b> 2004
<b>Price : <strike> 49.99 </strike> </b>
</td> </tr>
...
</table>
```

图 1　网页片段文本

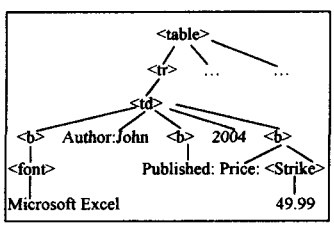

图 2　网页文本的 DOM 树表达

由于每一个关键词都对应到文本节点，因此我们可以利用每一个关键词的路径来得到所有关键词的最近公共祖先，从而有效地定位到中心节点，具体算法如下。

输入：关键词组集 $\{k_1, k_2, \cdots, k_n\}$

输出：中心节点和关键路径

步骤：

1. 对待抽取的网页建构 DOM 树结构，深度优先遍历 DOM 树，找到每个关键词 $k_i(i=1,\cdots,n)$ 对应的文本节点，生成对应的遍历路径集合$\{p_1, p_2 \cdots, p_n\}$。

2. For $i:=1 \to n$;

For $j:=i+1 \to n$;

寻找最长的公共子字符串 p_l'，其中 p_l'字符串必须是从 p_i 首个字符开始严格匹配，并将 p_l'加入到公共子字符串集合$\{p_i'\}$，$i \leqslant n$ 中，然后累计 p_l'出现的次数。

End;

End;

3. 在集合$\{p_i'\}$，$i \leqslant n$ 中出现次数最多的 p_l'就是关键路径，关键路径中最后一个标签字符是我们要获取的中心节点。

上述算法中循环遍历路径集合，主要考虑到有的关键字可能在网页的其他位置出现，这样得到的 DOM 树路径就不在待抽取的信息块中，因此通过寻找最长的公共子字符串就可以避免这个问题。

2.2 中心节点外的信息抽取

在电子商务网站中，像商品的图片等信息都不在关键信息块中，我们可以根据基于节点距离的网页信息抽取方法来抽取这些丰富的信息，但是我们的方法是基于 DOM 树中路径来计算节点间的距离，而不是基于二叉树中节点的宽度和深度来计算节点的距离，这种基于 DOM 树中节点的距离的方法既简单又可以减少遍历二叉树的时间。节点的距离是根据节点在 DOM 树中的路径来确定的，具体的算法如下。

输入：任意两个节点在 DOM 树中的路径 p_1，p_2。

输出：节点间的距离 d。

步骤：

1. 采用 2.1 节中的算法求出 p_1，p_2 最近公共祖先节点 n 以及 n 的路径 p；

2. 分别求出 p_1，p_2 中最后一个节点和节点 n 的距离，$d_1=\text{length}(p_1)-\text{length}(p)$以及 $d_2=\text{length}(p_2)-\text{length}(p)$；

3. $d=d_1+d_2$。

考虑到图片的大小、形状等信息，我们根据基于关键词聚类和节点距离的网页信息抽取方法[3]中的公式来计算待抽取的图片信息的权值 $W=\frac{k_1}{1+d}+\frac{k_2 s}{S}+\frac{k_3(h-w)}{w}$，其中 W 是图片的权值，k_1，k_2，k_3 是权值调整系数，d 是节点距离，s 是图片的面积大小，S 是源文件里面积最大的图片的大小，h 是图片的高度，w 是图片的宽度。

3　实验分析

我们采用信息抽取系统中使用最多的两个评测方法作为评价标准，即精度和召回率。定义如下：

$$P=Precision=\frac{\text{正确抽取到的记录数}}{\text{抽取到的记录数}}$$

$$R=Recall=\frac{\text{正确抽取到的记录数}}{\text{记录总数}}$$

我们对多个电子商务网站(包括手机商品网站和五金商品网站)中共有的属性进行测试，包括产品的型号、价格、图片等信息进行抽取，其中手机商品网站中包含商品信息的中心节点的标签是〈table〉，五金商品网站中包含商品信息的中心节点标签是非〈table〉的。实验中 k_1，k_2，k_3 取值分别是 0.6，0.2，0.2。手机商品网站的抽取实验结果如表 1 所列(括号内的数据为基于关键字和节点距离的信息抽取的实验结果)，五金商品网站的抽取实验结果如表 2 所列。

(下转第 210 页)

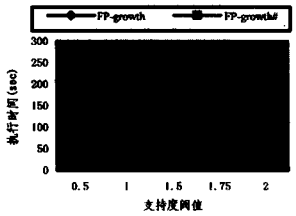

图 2 稀疏数据集上算法比较

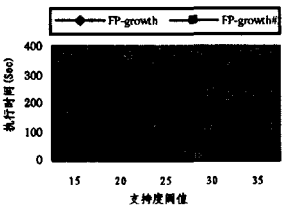

图 3 稠密数据集上算法比较

图 3 显示了两个算法在真实数据集上的运行时间。因为是非常稠密的数据集，FP 树有非常好的压缩性，两个算法的执行时间几乎一样，因此 FP 阵列技术不适合使用在这类挖掘中。

结束语 本文将 FP-tree 和 FP 阵列有效地结合起来，提出了一种改进的 FP-growth 算法。实验结果表明，对于稀疏数据库，本算法具有比 FP-growth 算法更优的性能。如何将 FP 阵列技术应用于最大频繁项集挖掘和闭频繁集挖掘，有待进一步研究。

参 考 文 献

[1] Agrawal R , Srikant R . Fast algorithm for mining association rules[A] // The International Conference on Very Large Data Bases[C]. 1994:487-499

[2] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[A] // The 2000 ACM SIGMOD International Conference on Management of Data[C]. 2000:1-12

[3] Han Jiawei. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2002:158-161

[4] Wang K, Tang L, Han J. Top down FP-growth for association rule mining[C] // Proc. of the 6th Pacific Area Conference on Knowleage Discovery and Data Mining(PAKDD). 2002

[5] Pei J, Han J, Lu H. H-mine: Hyper-structure mining of frequent patterns in large databases[A] // Proc. of IEEE Intl. Conference on Data Mining[C]. 2001:441-448

[6] Pietracaprina A, Zandolin D. Mining frequent itemsets using patricia tries[C] // Proceedings of the 1st Workshop on Frequent Itemset Mining Implementations. Melbourne, Fl, Nov. 2003

[7] http: // www. almaden. ibm. com/software/quest/resources/index. shtml

[8] http: // fimi. cs. helsinki. fi

(上接第 203 页)

表 1 手机网站的抽取结果

网站	图片正确率	价格正确率	型号正确率	召回率	时间(ms)
21cn	100% (100%)	100%(99%)	100% (100%)	99%(98%)	220 (317)
3553	99% (99%)	99% (98%)	99% (99%)	95% (93%)	332 (464)
pudou	100% (100%)	98% (97%)	100% (100%)	96% (94%)	136 (252)
shouji	99% (99%)	100% (100%)	100% (100%)	95% (95%)	253 (374)
younet	99% (99%)	100% (100%)	98% (98%)	95% (95%)	152 (231)

表 2 五金网站的抽取结果

网站	图片正确率	价格正确率	型号正确率	召回率	时间(ms)
chaolong	100%	100%	100%	99%	71
chinawj	100%	100%	100%	97%	164
Wjb2b	98%	99%	100%	94%	126
wujin. xd55	100%	99%	100%	96%	207
cn. easthardware	99%	98%	100%	95%	54

实验表明，无论商品信息在网页中以何种标签方式存储，基于 DOM 树路径的网页信息抽取方法都具有较高的精度和召回率，而且在效率上也有一定的提高。因为它不需要对 HTML 网页源文件的结构树进行简化和修剪处理，也不需要通过聚类算法来找到包含商品信息的关键块，只需要通过每个关键词组在 DOM 树中的路径来寻找最近祖先节点即可。因此基于 DOM 树路径的网页信息抽取方法是一种与站点结构无关的抽取方法，具有很强的适应性。

结束语 基于 DOM 树路径的网页信息抽取方法是对基于关键词组聚类和节点距离的网页信息抽取方法的一种简化和扩充。它利用对待抽取的网站的部分网页集来获取到关键词组，然后构建每个网页的 DOM 树结构，通过遍历 DOM 树结构来获取每个关键词在 DOM 树中的路径，这些路径的最近祖先节点包含了我们待抽取的信息块。同时对块外的信息，比如图片等等，通过比较图片节点与最近祖先节点之间的路径来获取它们之间的距离，然后利用基于节点距离的网页抽取方法来获取正确的信息。因此基于 DOM 树路径的网页信息抽取方法能够获取比较高的正确率和效率。同时该方法不局限于以〈table〉标签存放商品信息的网站，对其他各种标签存放的商品信息也能准确地抽取，因此具有很强的适应性和扩展性，在各种电子商务网站中都具有比较广阔的应用前景。

参 考 文 献

[1] Wong T-L , Lam W . Adapting Web Information Extraction Knowledge via Mining Site-Invariant and Site-Dependent Features [J]. ACM Transactions on Internet Technology, 2007, 7(1)

[2] Yang Pei, Zheng Qilun, Peng Hong, et al. A Stepwise Learning Approach to Automatic Discover Interest Data Block[C] // The third International Conference on Machine Learning and Cybernetics(ICMLC). 2004

[3] 邓健爽，郑启伦，彭宏，等. 基于关键词聚类和节点距离的网页信息抽取[J]. 计算机科学，2007，34:213-216

[4] [OL]. http: // www. w3. org/DOM/