基于页面分类的 Web 信息抽取方法研究

成卫青,于 静,杨 晶,杨 龙
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:通过对现有 Web 信息抽取方法和当前 Web 网页特点的分析,发现现有抽取技术存在抽取页面类型固定和抽取结果不准确的问题,为了弥补以上两个不足,文中提出了一种基于页面分类的 Web 信息抽取方法,此方法能够完成对互联网上主流信息的提取。通过对页面进行分类和对页面主体的提取,分别克服传统方法抽取页面类型固定和抽取结果不够准确的问题。文中设计了一个完整的 Web 信息抽取模型,并给出了各功能模块的实现方法。该模型包含页面主体提取、页面分类和信息抽取等模块,并利用正则表达式自动生成抽取规则,提高了抽取方法的通用性和准确性。最后用实验证实了文中方法的有效性与正确性。

关键词:Web 信息抽取;正则表达式;页面分类;HTMLParser;结点树
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)01-0054-05
doi:10.3969/j.issn.1673-629X.2013.01.014

Web Information Extraction Research Based on Page Classification

CHENG Wei-qing, YU Jing, YANG Jing, YANG Long
(School of Computer Science & Techn., Nanjing University of Posts and Telecomm., Nanjing 210003, China)

Abstract: By means of analysis of existing Web information extraction and the current Web page characteristics, current extraction techniques are found to have problems that the types of extract page fixed and the extract results are not accurate. In order to make up for the deficiency mentioned above, propose a Web information extraction method based on page classification. This method is able to complete the extraction of the mainstream of information on the Internet page. By classifying the Web page and extracting the main body of the page, it overcomes the two problems existing in traditional method respectively. A complete model of the Web information extraction is designed and the details of each functional module are provided. The unique features of the model are containing modules of Web page principle part extraction and Web page classification, as well as using regular expression to generate extraction rules automatically that promote the generality and precision of the extraction method. Experimental results have verified the validity and accuracy of the method.

Key words: Web information extraction; regular expressions; page classification; HTMLParser; node tree

0 引 言

随着互联网的快速发展,Web 已经发展成为一个巨大的信息资源。然而正是由于网络信息量繁多,且网页数据是半结构化的,对人类来说从这样的数据中搜索需要的信息已变得相当困难,更不用说对当下流行的应用程序,应用程序无法直接解析并利用网络上的海量信息。为了增强网络数据的可用性,出现了 Web 信息抽取技术,它将网页上的信息以更为结构化的方式抽取出来,可以供数据分析系统、批量查询系统等应用程序使用。Web 信息抽取技术已成为近年来的研究热点,具有广阔的应用前景。

收稿日期:2012-04-15;**修回日期**:2012-07-20
基金项目:国家自然科学基金资助项目(61170322,71171117);软件开发环境国家重点实验室开放课题(SKLSDE-2011KF-0X);江苏省自然科学基金资助项目(BK2010524)
作者简介:成卫青(1972-),女,副教授,博士,通讯作者,研究方向为网络测量;于 静(1988-),女,山东临沂人,硕士生,CCF 会员,研究方向为 Web 信息抽取。

1 现有 Web 信息抽取方法

1.1 人工获取规则处理方式的信息抽取

人工获取规则处理方式的信息抽取指技术人员依靠自己掌握的知识,审阅某些待处理的文本文档,总结出相关信息出现的规律,再根据信息抽取系统内部的抽取规则格式表达出相关的抽取规则。这类手工方式构造提取模式的方法,由于是人工编写提取模式,所以准确性比较高,但非常耗费人力。目前采用这种处理方式的信息抽取方法有基于本体的信息抽取,由专家对本体(ontology)[1]进行分析、调整并人工制定规则和模板。

1.2 半自动化的信息提取

这种方式的提取模式不是通过人工来编写的,而

是通过半自动化方式产生的。其过程可描述如下:首先获得源网页,在浏览器中显示;接着定义目的模式结构;然后标记源网页中感兴趣的内容,并与目的数据模式之间建立映射;最后通过启发式算法或其他算法,由程序根据映射关系归纳、总结、推导出提取模式[2]。

1.3　自动的信息提取

在 Web 信息提取过程中,面对的是海量的数据,如果采用人工提取或人工学习的方式进行信息提取并不现实。通过分析网页可以发现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。因此挖掘最大重复模式可以发现非常有用的提取规则[3]。这类利用页面本身特点来自动获得提取模式的方法,称之为基于规则的模式提取,也就是自动信息抽取。

通过以上对各类 Web 信息抽取方法的描述,可以看出自动的信息提取方法具有很大的优势。文中所提出的基于页面分类的 Web 信息抽取方法(Web information extraction based on Web page classification, WIEBC)就是属于这一类的。它也是通过分析网页结构,挖掘出最大的重复模式,继而生成相应的抽取规则进行信息抽取。不同的是,文中采取的对 Web 页面的分析方法更加有效,抽取方法更加简洁准确。

2　基于页面分类的 Web 信息抽取方法

2.1　方法概述

已有研究中比较有效的 Web 信息抽取模型包括两类:一类是针对相似页面结构的 Web 页面的信息抽取,另一类是基于规则生成的 Web 页面信息抽取,文中正是利用了这两类技术的优势,提出了一种应用领域更广、适用性更强的抽取模型——基于页面分类的 Web 信息抽取方法(WIEBC)。

WIEBC 方法包括 5 个步骤:首先将 Web 页面生成页面树,然后通过遍历页面树,过滤并进行页面去噪,之后提取页面主体,再根据页面特征进行页面分类,然后对各类页面应用相应抽取规则,最后将结构化的数据存入数据库,如图 1 所示。该方法与已有方法的不同主要体现在页面主体提取,页面分类和相应抽取规则生成模块。

现在网络上的大多页面都是用 HTML 语言来描述的,为了方便处理,生成页面树是很明智的选择[4]。目前生成页面树的技术也有多种,例如利用 Dom 解析器生成 Dom 树[5,6],利用 HTML 标签生成标签树等,文中利用 HTMLParser[7] 生成结点树。由于网络页面内容的繁杂性,事先对页面进行清理是相当有必要的,模型中表示为页面去噪。然而只进行页面去噪无法将整个 Web 页面的主体净化出来,所以还要对干净的页面进行主体提取,以得到需要的主体块,页面主体提取是文中的一个创新点。主体中包含了各种类型的隐藏数据,根据它们的特征可以进行页面分类。接下来要做的就是对不同类型的数据应用不同的抽取规则,这样可以提高抽取效率和准确性。抽取出来的数据经过处理就可以存到数据库中供应用程序使用。

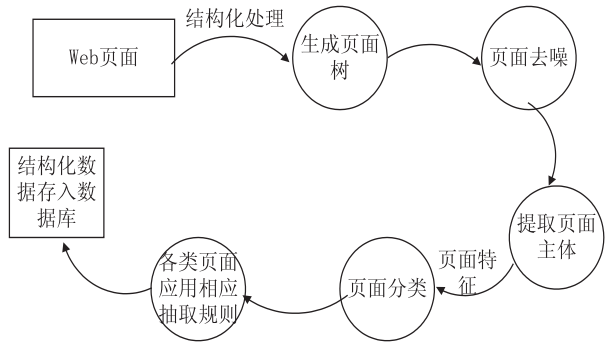

图 1　系统完整模型

2.2　生成页面树、页面去噪及提取页面主体模块技术

HTMLParser[7] 是一个快速实时的 HTML 文档解析器,并且是一个开源项目,可以从 HTMLParser 项目主页[8]下载到 jar 文件,导入 java 库后很容易使用。文中就利用它来生成 Web 页面的结点树,并且利用它对页面进行去噪处理。

Parser 类从高层将网页看成是一棵分层次的结点树,<HTML>是根结点,它的下层结点是<BODY>和<HEAD>,这两个结点又有自己的子结点,这样就形成一棵分层次的结点树。结点之间存在父子、兄弟等关系。

关于页面去噪问题我们有以下规则:

①Html 的 head 标记以及其间的内容是可以删除的;

②脚本类标记,以及注释类标记以及其间的内容是可以删除的[9];

③空标签以及 select、input 等标记以及其间的内容是可以删除的[9];

④类型为 hidden 的标记以及其间的内容是可以删除的[9];

⑤img 标签是可以删除的;

⑥用于表示广告,弹出窗口的结点是可以删除的;

⑦处于页面边缘的超链是可以删除的。

HTMLParser 中的 Lexer 包主要负责从 HTML 源读入字符和识别结点位置。Filter 包中定义了多个过滤类[10],可以根据需要从网页中提取特定的类、结点等内容来实现相应功能。Nodes 包定义具体结点的实现,包中所有的类是 Node 接口的具体实现,这个包为文本、标签和注释实现了一些特有的接口。Tags 包包含了详细的标记和标记的实现,且标记实现的功能大大超过了一般标记的实际功能。利用这些 HTMLPars-

er 中包的方法可以灵活地完成所需要的功能,而且网页的主体信息一般包含在<table>、<tr>、<td>、<p>、 等标签中,在进行网页去噪之后再对结点树进行处理就相对简单了一些。

关于页面主体抽取,首先在定义 1 中给出页面主体的定义。

定义 1:理想状态下,页面主体是以页面结点树中满足一定条件的结点 P 为根结点的结点树,结点 P 满足如下条件:P 的每一棵子树的结点个数相等,且在整棵树中不存在满足同样规则的结点 Q,使得 Q 的子树的结点个数大于 P 的子树的结点个数。

可以用广度优先遍历方法对页面结点树进行遍历,将满足规则(每一棵子树的结点个数相等)并且子树结点数最多的结点 P 找出来,以 P 为根结点的结点树就是要提取的页面主体部分。寻找结点 P 的这一算法是在具有规范的 HTML 源码的理想状态下的算法,实际应用中需要进一步的完善。由于这一算法并不是文中的主题,限于篇幅,这里仅作描述,完善的详细算法,将另行文详述。

2.3 页面分类算法

最常被浏览的网页可以分为三类。第一类称为导航页面,页面主体是由一系列排列有序的超链组成。第二类称为文本页面,网页主体是由大段的文本内容组成。第三类网页主体含有相似的块结构,文中将这样的块结构称为重复块,如当当网图书搜索结果页面,每本书的信息块就是一个重复块。利用这三类页面的不同页面特征可以把它们区分开来。

在给定的 HTML 中,首先采用上述方法提取页面主体,并设页面主体叶子结点集合为 M,然后寻找满足下面条件的结点 j:M 中任何一个结点都是 j 的子结点。而对于 j 的任何一个子结点 jc,M 中都存在不是 jc 子结点的结点[11]。根据分析,存在满足条件的结点 j 的页面可能是第一类或第二类页面,再判断其中任意一个结点是文本结点还是超链结点就可以区分出是第一类还是第二类页面。剩下的页面归为第三类页面。

2.4 各类页面的抽取规则

正则表达式(Regular Expressions)是定义一组字符串的字符和符号序列[12]。利用正则表达式可以快速方便地实现字符串的模式匹配,以及对输入域中的值进行数据验证。可以用<a href=". *">这个规则来匹配所有超链。对于所有的标签可以用原样符号进行匹配。而对于汉字的匹配可以用[^\x00-\xff]来匹配两个字节。在 HTML 文本中匹配两个字节就相当于匹配了汉字。这里还要利用正则表达式中“组”的概念,可以用圆括号创建一个或多个组,在信息提取中就是对指定匹配的字符组进行提取,得到想要得到的内容。

第一类和第二类页面主体部分的结构是非常相似的,尽管在显示上看起来好像差距很大,但是对于大段文本的页面,其各行之间基本都存在标签,也就是每行都可以作为一个结点。第二类页面很显然会把一个超链放到一个结点。但是,文中对这两类页面的抽取内容是不一样的。对第一类页面,要抽取的是超链的 URL,采用抽取规则一进行提取。而对于第二类,这里只抽取文章的标题,但不能利用<title>标签来提取,因为<title>标签一般是在<head>标签中的,这早在页面去噪时就已经被净化掉了,但是可以利用文章标题常用<h>标签显示的特点进行提取,相应的提取规则见抽取规则二。

抽取规则一:([^\x00-\xff] *),其中第一个组匹配 URL,第二个组匹配中文描述。

抽取规则二:<h\d>([^\x00-\xff] *)</h\d>,其中的组用来匹配正文文本。

定义 2:重复体为页面中的一段 HTML 代码,为若干相似块结构的一个块,它包含了一条信息,这一条信息可以包括一个或多个属性。

对于第三类页面,关键信息的提取较复杂,要利用网页主体内块结构的相似性进行操作。首先,将提取到的页面主体部分的源码,以字符序列的形式写入文本文件 f_1,这样做的目的是方便进行匹配以及提取操作。这里需要建立一个知识库,用来存放已生成的抽取规则,每种类型的抽取规则只需要生成一次即可。当有新的页面需要提取时,可以先到知识库中查找有没有匹配的抽取规则,如果有就可以直接应用,如果没有则要生成新的抽取规则。

这类页面的特征决定了文本文件 f_1 中是一个个重复体,可以通过比较各个重复体,提取相同的字符序列来生成抽取规则。这里提取相同的字符序列并不提取超链标签和双字节字符,尽管这两部分也有可能是相同的,而是利用替换规则一,得到标签和正则表达式组合而成的字符序列,这就是一条抽取规则,将它放入知识库文本文件中。用这个抽取规则不但可以对当前分析的页面内所有的重复体进行抽取,还可以对此类所有可以匹配的页面进行抽取,往往对同一个网站中多个页面都是适用的。

替换规则一:用<a href=". *">来代替超链,用([^\x00-\xff] *)来代替任意个汉字部分,用\d * 代替任意个数字。

有了抽取规则,就可以利用它进行页面关键信息的抽取。信息抽取包括两步,首先要获得信息的数据结构,利用数据结构可以生成数据库表格,其次是提取信息。为获得数据结构,先利用抽取规则匹配至少两

个重复体,为了提高准确性可以多匹配几个重复体,对各个重复体提取到的数据进行比较,将相同的字符序列存入数组 t,t 中的元素就是我们想要得到的数据表的列属性,利用数组 t 就可以动态生成数据库表。理想情况下,匹配两个重复体就可以,但是有时候信息里面也会有相同的值或者序列,所以最好多匹配几个重复体。第二步是用抽取规则匹配本页面内的所有重复体,提取信息,并将其存入生成的数据库表格中,得到所要抽取的结构化数据,也即第三类页面的关键信息。

第一类和第二类页面的整体流程相似,但是相对要简单得多,这里不再详述。具体流程图如图 2 所示。

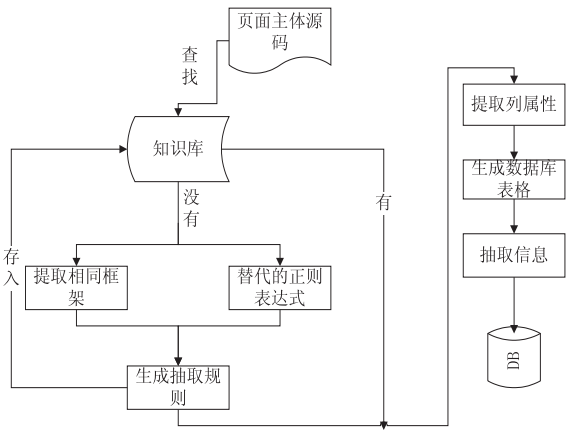

图 2　信息抽取模型

3　实　验

在文中所提的模型以及各部分相应算法的基础上,实现了一个原型系统。由于第一类和第二类页面相对简单,实验中只对第三类页面进行页面抽取。开发环境为 eclipse,并导入了 HTMLParser 2.0 的 jar 文件,数据库为 Mysql。

利用 2.2 中所描述的方法,利用 HTMLParser 生成页面结点树,并进行去噪处理,然后根据提取页面主体的算法,找到满足条件的结点 P,将以 P 为根结点的子树的 HTML 源码存入到一个文本文件中,这一部分对结点处理的方法都可以在 HTMLParser 中找到。

根据 2.3 节的页面分类算法,判别此页面是第三类页面;然后查找知识库,没有匹配此页面的抽取规则,则根据替换规则一动态生成此页面的抽取规则。先提取重复体的相同框架,用<a href=". * ">来替换超链,用([^\x00-\xff] *)来替换任意个汉字部分,用\d * 替换任意个数字。并将生成的抽取规则存入知识库。接着利用此抽取规则对主体的源码进行匹配,首先抽取出数据表所需要的列属性,然后利用得到的列属性在数据库中动态生成数据表。最后,利用抽取规则匹配此网页中所有的重复体,抽取所需要的信息,并存入数据库表中。对应每一个重复体,所抽取出来的信息存入数据表的一行中,作为一条记录。

对 22 个重复体进行了抽取,而表 1 中的抽取结果显示抽取到了 20 条记录,其中第 9 条显然是错误的,可能是因为第一项没有匹配,而导致后边的项发生了错位。

采用信息抽取系统中使用最多的两个评价方法作为评价标准,即召回率(recall)和查准率(precision)[13]。本实验的召回率和查准率为:

R = Recall = 正确抽取到的记录数/记录总数 = 19/22 = 86.36%

P = Precision = 正确抽取到的记录数/抽取到的记

表 1　网页信息抽取结果表

序号	书名	作者	出版社	出版日期	原价	蔚蓝价	折扣
1	计算机网络基础	张晓婷高棪	人民邮电出版社	2003/11/1	¥19.0	¥16.7	¥88 折
2	计算机网络与因特网	科默	机械工业出版社	2000/8/1	¥40.0	¥35.2	¥86 折
3	计算机网络与通信	张元编	电子工业出版社	2004/1/1	¥21.0	¥18.5	¥88 折
4	计算机网络	美	人民邮电出版社	2004/1/1	¥52.0	¥45.8	¥88 折
5	计算机网络技术基础教程	龚桂平	清华大学出版社	2004/3/1	¥23.0	¥20.2	¥88 折
6	计算机网络实验操作教程	蒋理编	西安电子科技大学出版社	2006/12/1	¥31.0	¥27.3	¥88 折
7	计算机网络专业英语	张筱华编	北京邮电大学出版社	2002/6/1	¥18.0	¥19.4	¥88 折
8	计算机网络安全教程	石志国	北京交通大学出版社	2007/1/1	¥21.0	¥24.2	¥78 折
9	杨尚森	高等教育出版社	2005/4/1	¥24.0	¥17.6	¥75 折	¥6.0
10	计算机网络技术	刘敏涵	西安电子科技大学出版社	2004/7/1	¥23.0	¥18.5	¥88 折

录数 = 19/20 = 95.0%

本实验选取的 Web 页面相对规整,可能会使抽取精度相对偏高,但是实验结果已经足以表明文中所设计模型是合理有效的。

4 结束语

文中提出了一种新的 Web 信息抽取的方法,即 WIEBC 方法,并设计了一个完整的系统模型。与已有模型的不同之处在于模型中增加了页面主体提取模块,引入了页面分类方法,并且利用正则表达式自动生成各类抽取规则,能够动态生成数据库中的表格,提高了抽取方法的通用性和准确性。最后用实验证实了 WIEBC 方法的可行性。文中下一步工作是将提出的模型以及其中的算法加以改进和完善,并推广到更复杂网页的信息抽取中。

参考文献:

[1] 陈 静,朱巧明,贡正仙. 基于 Ontology 的信息抽取研究综述[J]. 计算机技术与发展,2007,17(10):84-86.

[2] 周合明,奚建清. 基于模板的 Web 信息提取系统的设计与实现[J]. 计算机技术与发展,2011,21(11):105-108.

[3] Xie Tao, Shi Shengsheng, Quan Fuliang, et al. Research on Complex Structure-oriented Accurate Web Information Extraction Rules[C]//International Conference on Progress in Informatics and Computing. [s. l.]:[s. n.],2010.

[4] 周 登,戴玉刚,付 涛. 基于树结构的 Web 信息抽取[J]. 计算机技术与发展,2009,19(9):38-41.

[5] 李效东,顾毓清. 基于 DOM 的 Web 信息抽取[J]. 计算机学报,2002,25(5):526-533.

[6] Liu Yaqing, Chen Rong, Yang Hong. Web Information Extraction Based on Hierarchical Model[C]//Second International Conference on Computational Intelligence and Software Engineering. [s. l.]:[s. n.],2009.

[7] 曾维佳. 基于 HTML Parser 的 Web 信息提取系统的设计和实现[J]. 电脑知识与技术,2007,7(4):970-971.

[8] HTMLParser[EB/OL]. 2006-09-17[2012-03-12]. http://htmlparser. sourceforge. net.

[9] 任仲晟,薛永生. 基于页面标签的 Web 结构化数据抽取[J]. 计算机科学,2007,34(10):133-136.

[10] Lin Shan, Hu Yanzhong. An Approach of Extracting Web Information Based on HTMLParser[C]//Second International Conference on Information Technology and Computer Science. [s. l.]:[s. n.],2010.

[11] 蔡捷飞,陈啟泓,梁志宏,等. 主题型网页发现以及网页内信息块发现[EB/OL]. 2010. http://www. doc88. com/p-64281396657. html.

[12] Watt A. 正则表达式入门经典[M]. 北京:清华大学出版社,2008.

[13] 李 丹. 基于序列比对的动态 Web 信息抽取算法研究[D]. 长春:吉林大学,2009.

(上接第 53 页)

4 结束语

文中讨论了一种利用 Markov 随机场下构建的多变量高斯模型来对过分割后的数字图像的特征向量进行建模,通过对模型的训练学习得到模型的参数向量,实现了单幅数字图像在多尺度空间下的场景深度的估计,在此基础上统计了该方法在一些场景下的误差。

实验证明,该方法可以有效地估计场景的深度值,且证明了随着尺度空间的变大,会减小其所获得的深度值误差。理论上,改进图像的概率模型可以缩小误差,因此,之后在对模型的改进上仍有许多不足需要改进。

参考文献:

[1] 朱伟利,朱 枫,郝英明. 基于单幅建筑物图像的三维信息提取[J]. 仪器仪表学报,2008,33(29):33-40.

[2] 吴凤和,张晓峰,施法中. 单幅图像三维表面重建算法的研究与实现[J]. 计算机应用,2009,15(12):56-62.

[3] 赵冬斌,陈 强,陈善本. 由单幅图像恢复物体三维形状的应用研究[J]. 光学技术,2001,11(9):78-82.

[4] 高月芳,罗 飞,曹建忠. 由单幅二维灰度图像重构物体表面形状[J]. 计算机应用,2007,36(7):15-21.

[5] 李 波,王祥凤,李本山. 基于单幅图像的三维重建技术[J]. 信息与电子工程,2006,4(4):71-77.

[6] 孙宇阳. 基于单幅图像的三维重建技术综述[J]. 北方工业大学学报,2001,23(1):9-13.

[7] Sudderth E B, Torralba A, Freeman W T, et al. Depth from familiar objects: A hierarchical model for 3D scenes[J]. Computer Vision and Pattern Recognition (CVPR),2006,24(3):229-237.

[8] Saxena A, Schulte J, Ng A Y. Depth estimation using monocular and stereo cues[C]//International Joint Conference on Artificial Intelligence (IJCAI). [s. l.]:[s. n.],2007:112-118.

[9] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. International Journal of Computer Vision,2004,59(2):167-181.

[10] Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images[J]. Neural Information Processing System,2005,18(8):66-78.

[11] Davies E R. Laws´ texture energy in TEXTURE[M]//Machine vision: theory, algorithms, practicalities. San Diego: Academic Press,1997.

基于页面分类的 Web 信息抽取方法研究

作者： 成卫青， 于静， 杨晶， 杨龙
作者单位： 南京邮电大学 计算机学院,江苏 南京 210003
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(1)