基于树比较的 Web 页面主题信息抽取 *

朱梦麟,李光耀,周毅敏
(同济大学 电子与信息工程学院,上海 201804)

摘 要:为了从具有海量信息的 Internet 上自动抽取 Web 页面的信息,提出了一种基于树比较的 Web 页面主题信息抽取方法。通过目标页面与其相似页面所构建的树之间的比较,简化了目标页面,并在此基础上生成抽取规则,完成了页面主题信息的抽取。对国内主要的一些网站页面进行的抽取检测表明,该方法可以准确、有效地抽取 Web 页面的主题信息。

关键词:信息抽取;相似页面;树比较;抽取规则
中图分类号:TP391.9 文献标识码:A 文章编号:1674-7720(2011)19-0067-03

Topic information extraction from Web pages based on tree comparison

Zhu Menglin, Li Guangyao, Zhou Yimin
(Department of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: In order to automatically extract Web page information from Internet that contains magnanimous information, this paper presented an approach based on tree comparison. This approach compared tree built from the target page with that ones built from its similar pages to simplify the target page. Extraction rules were generated on this basis, and then we used the rules to extract topic information from the target Web page. Experiment result shows this extraction method is precise and efficient.

Key words: information extraction; similar pages; tree comparison; extraction rules

随着 Internet 的飞速发展,Web 已经发展成为一个共享的数据空间,互联网已成为人们获取信息的重要渠道。而在 Web 数据呈几何级数增长的同时,用户查找、定位自己所需的信息变得越来越困难,如何快捷、有效地搜索信息成为亟待解决的问题,Web 信息抽取技术正是在这种背景下应运而生。Web 信息抽取技术的核心是能够从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,使其更为结构化、语义更为清晰的格式。比如从新闻报道中抽取出新闻的时间、地点、主要内容等;从介绍商品的网站上抽取出商品的价格、参数、评价等。通常,被抽取出来的信息以结构化的形式描述,可以直接存入数据库中,供用户查询以及进一步分析利用。当今,Internet 已经成为发布和传播信息的最重要手段,网络上的信息和活动对人们的影响越来越明显。一个良好的 Web 信息抽取系统可以高效地收集所需的网络信息,并加以分析利用,如应用于专业数据获取、股票预测、用户行为爱好分析等。目前,像 Newsbot、Shopbot 等一些针对特定领域的信息抽取/集成软件已经投入了商业应用,帮助人们随时获得最新的新闻消息或收集同一商品的不同价格信息以决定合理的购买方式。

Web 的数据大部分都是以 HTML 形式出现的,这是一种半结构化的数据,缺乏对数据本身的描述,不含清晰的语义信息,模式也不太明确,这使得应用程序无法直接解析并利用页面上的信息;并且由于人们审美和商业的需求,充斥着大量与主题无关的修饰信息,如图片、广告、各种脚本语言等。如何排除干扰,有效地确定 Web 页面中的主要数据区域并从中抽取出大家所关注的主题信息是本文的主要工作。

Web 信息抽取技术发展至今,已经有了很多比较成熟的方法,如基于文本统计的信息抽取技术[1]、基于 HTML 结构的信息抽取技术[2]、基于隐马尔科夫模型的信息抽取技术[3]等。这些方法各有利弊,但有一个需要共同面对的问题是对于目标页面的不定期改版,原有的抽取规则可能会失效。本文提出的基于树比较的 Web 主题信息抽取技术是一种基于 HTML 结构的信息抽取

* 基金项目:上海市科委国际合作项目(10510712500)

方法。通过目标页面与其相似页面的比较训练,简化目标页面并生成抽取规则,以此规则来完成目标页面主题信息的抽取。当页面改版,抽取规则失效时,会自动进行重新学习而生成新的抽取规则。经验证,本抽取系统具有良好的健壮性,能很好地解决这个问题。

1 相关概念

1.1 DOM 树

DOM(Document Object Model)是由 W3C 制定的一种与平台和语言无关的标准接口规范,它允许程序和脚本动态访问、修改文档的内容、结构和类型。它定义了一系列的对象和方法对 DOM 树的节点进行各种随机操作。DOM 树中的节点可分为 4 种不同的对象:(1)Document 对象。作为树的最高节点,Document 对象是对整个文档进行操作的入口;(2)Element 和 Attr 对象。这些节点对象都是文档某一部分的映射,节点的定级层次恰好反映了文档的结构;(3)Text 对象。作为 Element 和 Attr 对象的子节点,Text 对象表达了元素或属性的文本内容。Text 节点不再包含任何子节点;(4)集合索引。DOM 提供了几种集合索引方式,可以对节点按指定方式进行遍历,索引参数都是从 0 开始记数的。DOM 树中的所有节点都是从 Node 对象继承而来,Node 对象定义了一些最基本的属性和方法,利用这些方法可以实现对树的遍历,同时,根据属性还可以得知节点的名称、取值并判断其类型。

1.2 XPath

XPath 即为 XML 路径语言(XML Path Language),它是一种用来确定 XML 文档中某部分位置的语言。XPath 基于 XML 的树状结构,提供在数据结构树中找寻节点的能力。最常见的 XPath 表达式是路径表达式(XPath 名称的另一来源)。路径表达式是从一个 XML 节点(当前的上下文节点)到另一个节点、或一组节点的节面步骤顺序。这些步骤以"/"字符分开,每一步有三个成分:轴描述(用最直接的方式接近目标节点);节点测试(用于筛选节点位置和名称);节点描述(用于筛选节点的属性和子节点特征)。本文的抽取规则就是以 XPath 的形式给出,使用 XPath 定位所要抽取的信息在 DOM 树中的节点。

用 Xpath 来定义抽取规则,虽然简单明确,但从抽取系统的健壮性来考虑,却存在着一定的隐患。假设要从图 1 这样一棵 DOM 树上抽取商品 iPhone4 的价格,则可以定义 XPath/html/body/div[2]/table/td[2]/text()为抽取规则。但是,当目标页面的布局稍有改变时,该抽取规则可能就不再适用,而需要重新训练学习[4]。比如,第一个 div 被删除,第二个 div 的 table 下新加了一些节点等。本文提出的信息抽取算法在当前的抽取规则失效后,会自动获取改版后的页面重新进行再学习、训练以生成新的抽取规则,确保了信息抽取系统的有效性。

1.3 DSE 算法

对于 Web 主题信息抽取来说,很重要的一步就是简

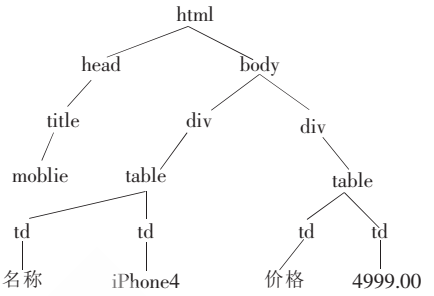

图 1 一个 Web 页面的 DOM 树

化待抽取的 Web 页面,确定主题信息所在的数据区域,删减与主题无关的干扰信息。DSE [5](Data-rich Section Extraction)算法能很有效地完成这个工作。DSE 的提出是基于这样一个事实:在同一个网站下,往往有大量使用同一设计模板的 Web 页面,这些页面具有相同或相似的 HTML 结构。同时,广告、导航信息等与主题无关的内容在这些页面的相同位置不断重复出现。这时,通过对由这些页面构建的 DOM 树进行两两比较,就可以尽可能地排除这些干扰信息,缩小下一步处理的数据集合,提高信息抽取的效率和精度。DSE 算法的基本过程如下:

(1)深度优先遍历两棵待比较的树 A、B。其中树 A、B 是由两个相似的 Web 页面构建所得。

(2)在遍历的同时,不断比较两棵树上相同位置的两个节点,对于相同的两个内部节点,则继续比较它们的子节点。对于叶子节点,如果比较结果相同,则把它们从该树上删除;如果不同,则继续比较下一个叶子节点。只有当一个节点的所有子节点都被删除后,才会删除该节点。

(3)当遍历整棵树后,树 A、B 中重复出现的与主题无关节点均已被删除。

图 2 显示了一个简单的 DSE 算法的 DOM 树比较过程。可以看到,树 A 经一次 DSE 算法比较后,一部分与主题信息无关的重复内容已被删除,页面 A 对应的 DOM 树已得到了很大程度的简化。

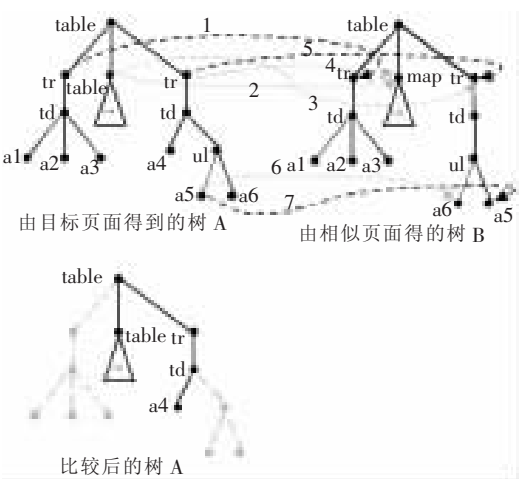

图 2 简单的 DSE 算法例子[5]

2 抽取算法及实现

2.1 抽取算法

本文进行的信息抽取算法具体步骤如下：

(1)构建目标页面的 DOM 树。由网上获得的目标页面的 HTML 源文件并构建其对应的 DOM 树。

(2)获取目标页面的几个相似页面。可利用正则表达式匹配等方法判断是否属于目标页面的相似页面。

(3)用 DSE 算法对目标页面与其相似页面进行比较匹配，简化待抽取的目标页面，具体的比较次数需要看页面的复杂程度，一般为 1~3 次。只有尽可能地简化目标页面的 DOM 树，缩小下一步处理的数据集合，才能有效提高抽取算法的速度和效率。

(4)在简化后的 DOM 树上进行遍历，寻找信息量最大的节点，并生成从根到该节点的 XPath。

(5)由 XPath 生成抽取规则和模板，并储存相关模板信息，用于今后该类页面的信息抽取。

(6)用生成的规则完成信息抽取，并把数据保存到数据库中。

2.2 系统的实现

如图 3 所示，根据设计目标，将系统分为以下模块：

(1)页面浏览模块：实现用户对 Web 页面的浏览和标记功能。用户可以在内置的浏览器中访问该页面，也可以在页面中进行标记。同时，在界面上方构建生成的 DOM 树中，也可以对各节点进行选择查看和标记。

(2)相似页面获得模块：获得与目标页面模板相同、结构一致的页面，用于后续的抽取规则训练算法。

(3)抽取规则生成模块：用 DSE 算法进行相似页面的比较训练，寻找待抽取信息所在的节点，生成 XPath，形成抽取规则。

(4)信息抽取模块：由抽取规则进行抽取，显示结果，并存入数据库。

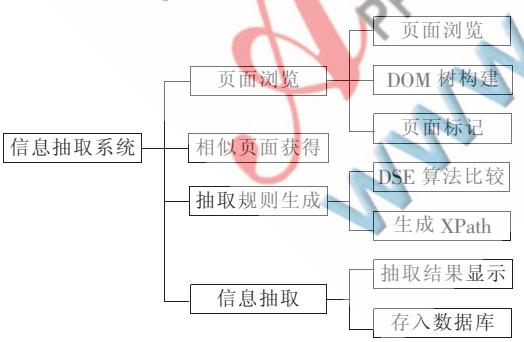

图 3 系统功能模块图

本信息抽取系统具体实现使用 Java 编程，以 Java Swing 制作界面。运行程序后，可以输入任意网址打开页面，并生成该页面的 DOM 树于界面左上方。比如，输入 http://www.sina.com.cn 后，信息系统抽取主界面如图 4 所示。

2.3 实验结果及分析

为了验证本算法的有效性，运用本系统对新浪、搜狐等网站的近千个新闻页面进行了试抽取，并人工检验了抽取的有效性。实验结果表明，大约 98.2%的页面都能正确抽取页面的主题信息，只有极少数的页面抽取失败或无法抽取。可见，本抽取算法具有一定的推广应用价值。

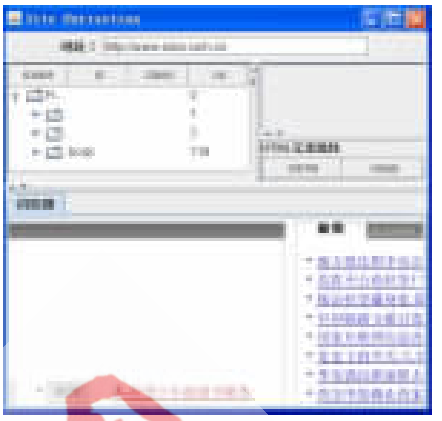

图 4 Web 信息抽取系统主界面图

本文提出了一种基于树比较的 Web 页面主题信息抽取算法，该算法能快速、准确、有效地抽取目标页面的主题信息。如何将该算法更好地应用于信息检索、数据挖掘的各方面是今后的主要工作。如应用于搜索引擎的搜索算法中，提高搜索引擎的检索速度和精度；或对已获得的页面信息进行进一步的数据挖掘，以发现其中有用的信息和知识。

参考文献

[1] 孙承杰，关毅.基于统计的网页正文信息抽取方法的研究[J].中文信息学报，2004，18(5)：17-22.

[2] 张彦超，刘云，李勇，等.基于自动生成模板的 Web 信息抽取技术[J].北京交通大学学报，2009，33(5)：40-45.

[3] 祝伟华，卢熠，刘斌斌.基于 HMM 的 Web 信息抽取算法的研究与应用[J].计算机科学，2010，37(2)：203-206.

[4] DALVI N, BOHANNON P, SHA F. An approach based on a probabilistic tree-Edit model [A]. Proceedings of the 35th SIGMOD International Conference on Management of Data (SIGMOD'09)[C]. New York：ACM Press，2009：335-348.

[5] Wang Jiying, FRED H. LOCHOVSKY.Data-rich section extraction from HTML pages [A]. Proc 3rd International Conference on Web Information System Engineering (WISE'02)[C].Singapore：IEEE Computer Society Press，2002：1-10.

(收稿日期：2011-05-06)

作者简介：

朱梦麟，男，1981 年生，硕士研究生，主要研究方向：计算机仿真、数据挖掘。

李光耀，男，1965 年生，研究员，博士，主要研究方向：计算机辅助设计分析与仿真、城市仿真与城市规划设计、图形图像技术。

周毅敏，男，1981 年生，博士，主要研究方向：计算机辅助设计与仿真。