基于 WEB 信息抽取的主动服务技术研究①

Research on initiative service technology based on web information extraction

于 静 (中国科学院合肥智能机械研究所 合肥 230031)
(中国科学技术大学信息科学技术学院 合肥 230027)
李 淼 (中国科学院合肥智能机械研究所 合肥 230031)

摘 要: 随着互联网的发展和普及,Internet 上的信息急剧增长,能够自动获取适用,简单和精炼的信息,成为人们的迫切希望。同时针对农村互联网条件差而手机越来越普及的情况,我们设计实现了一个基于 WEB 信息抽取和 GSM 的主动服务系统。本文在分析农产品供求信息网页结构的基础上,提出了一种基于内容和 web 文档结构路径(DOM)相结合的信息抽取算法。最后实验结果说明该抽取算法能够很好地制定抽取规则并能够准确的抽取所需要的内容。

关键词: WEB 信息抽取 GSM 主动服务 DOM 包装器

1 引言

随着互联网的发展和普及,Internet 上的信息急剧增长,人们开始迫切的希望找到一种能够在无限的信息海洋中自动获取适用,简单和精炼信息的方法。而仅仅依靠目前的搜索引擎已很难满足人们的需要。信息主动服务技术可以将用户感兴趣的信息及时推送给用户,有效地解决了用户信息服务个性化和信息更新及时性的问题,改变了传统的网上信息服务模式。统计数字显示,截至 2005 年 6 月,我国互联网用户已经居世界第二位,但来自农村的用户只占 0.3% 。而手机在农村的普及率却大大提高,以短信为主的数据业务正在以每年超过百分之百的速度增长。因此我们引进了基于 GSM 的主动服务技术,以解决农村用户在获取农产品供求等信息时对互联网的依赖问题。

信息抽取是指从文本中自动抽取相关的或特定类型的信息,并将其形成结构化的数据以供用户查询。Web 信息抽取是抽取 web 页面信息的过程,旨在从信息量极其丰富的 Web 资源中有效地挖掘出大量的、潜在的、有价值的知识。通常分为抽取自由文本,半结构化网页,结构化网页[1]。我们主要研究的是网络上大量存在的半结构化网页,而相同类型的网页结构具有很大的相似性[2]。

现代信息抽取技术源自于文本理解,然而这类抽取系统只能在很窄的知识领域范围内运行良好,向其它领域移植的性能很差[3]。近年来,针对 web 信息抽取的其他技术也取得了长足的进展:基于文本序列特征模式匹配的抽取系统,如 WIEN;基于 DOM 树结构路径的抽取系统,如 W4F;基于 Ontology 的描述文件,如 BYU。基于 Ontology 描述文件的抽取系统,需要专家的支持,工作量繁重。基于 DOM 树结构路径的抽取系统,对网页的结构依赖性较强,树路径任一层发生变化,就会导致抽取规则失效。本文采用的信息抽取规则依赖于基于内容和 web 文档结构路径(DOM)相结合的信息抽取技术。

2 web 信息抽取过程描述

本文采用包装器技术来对农产品供求信息进行抽取。包装器是将 html 的内容,利用定制好的规则抽取出来,转换成结构化的数据,供信息系统进一步处理。抽取规则是基于某一类特定网页基础上的,抽取规则的描述和处理是包装器中核心的部分。

① 基金项目:中国科学院知识创新工程重要方向项目(编号:KGCX2 - SW - 511)

2.1 待抽取网页描述

供求信息类的网页有其独特的结构特征。如图1所示,该页面显示了一个常见的供求信息网页的格式,把供求信息整体称为一个网页中我们感兴趣的信息块,即供求信息块。在该信息块中包含联系人,联系电话,信息内容等子信息。信息块内的子信息内容多为网页动态生成,有较一致性的结构,但供求信息块的位置却随网页结构的变化而变化。

供 求 信 息 | Supply and demand information

供应信息	
信息名称:	供应各品种 大/中/小杜鹃花
信息分类:	供应信息
联系人:	陈维新
联系电话:	13015600847
发布时间:	2007-1-24 12:56:25
信息内容:	本花场有各品种大、中、小杜鹃花。 有意者请联系！！！ 本场还有其他花卉供应。

图 1　供求信息网页结构

2.2 扩展的 DOM 树结构描述

文档对象模型 Document Object Model(DOM)树是一个对象化的 HTML 数据接口,一个与语言无关,与平台无关的标准接口规范,它定义了 HTML 文档的逻辑结构,给出了一种访问和处理 HTML 文档的方法[4]。DOM 树结构简单清晰,意义表述明确,成为描述和操作 HTML 文档的最流行的方式之一。它展现 HTML 层次化的文档结构,将 HTML 语言里面的标记(TAG)作为 DOM 树的节点,形成一种层次化的 DOM 树。

图2表示的是一棵简单、标准的 DOM 树。树中的每一个节点对应于 HTML 语法里的 TAG 元素。图中的 DOM 树是图1网页对应的 HTML－DOM 树的一个部分,该 DOM 树对网页的 DOM 树进行了剪枝和清洗,删除原有的部分与抽取无关的节点和对显示效果起修饰性作用的节点。由图可以看出,对于各个子信息,有共同的父节点,该节点以下包含所有的子节点共同构成了所要抽取的信息块。该信息块是一棵以其父节点为根节点的 DOM 子树。本文的抽取规则是基于 DOM 树结构,使用了扩展的路径表达式。

2.3 抽取规则的描述

采用包装器的抽取规则来实现对网页内容的抽取。基于内容的信息抽取是针对网页中那些不易变化

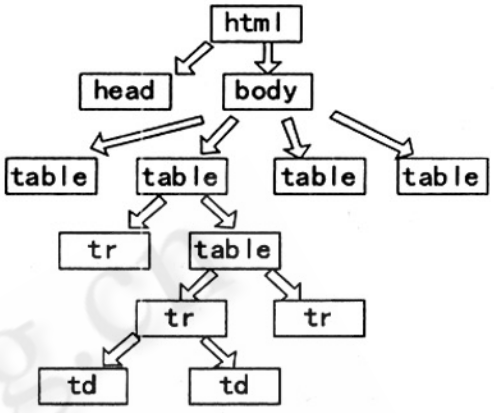

图 2　一个简单的网页 DOM 结构

的信息来对待抽取信息进行定位的方法,不需要文档的结构层次信息,对于网页结构变化多变的一类的网页能够很好地处理。但是,它缺少对文本的精确定位信息。基于 DOM 树提取路径表达式的抽取规则具有良好的结构性,可以根据树中的节点准确定位 HTML 页面中的 TAG 标记,能够准确定位到所要抽取信息的位置。该抽取方法抽取数据明确,不会产生歧义。但该方法不能适应 HTML 文档结构的甚至是细微的变化,树路径越长抽取规则的健壮性越差。综合基于内容抽取的灵活性和 DOM 树结构定位的精确性的优点,本文采用了基于内容和 web 文档结构路径(DOM)相结合的抽取规则来实现信息抽取。

2.3.1 信息块的规则表示

信息块规则是对网页页面内容中的信息块的定位。我们通过网页信息块中那些不易变化的文本信息,可以简单而快速地定位到信息块的位置,即所有子信息的共同的最近的祖先节点。如图1网页中,我们通过"信息名称:"可以很快的定位到供求信息块的位置。对于信息块不能唯一定位的情况,可以适当增加匹配的文本内容信息。

2.3.2 子信息的规则表示

子信息的抽取规则表示是对供求信息块内各子信息的块内路径表达式的描述信息。下面是一个子信息的规则表示:Title = .TR[0].TD[1]。子信息规则表达式是一个二元组,由子信息名、子信息定位表达式组成,分别对应于规则中的"Title"".TR[0].TD[1]"。子

信息定位表达式是对子信息在信息块内的位置的描述，我们采用 DOM 树的路径表达式来表示。这里不采用从页面根节点到子信息节点的绝对路径，而是采用信息块树内的相对路径，相对路径较短，对网页结构不是很敏感。按先根次序遍历根节点下的所有节点，抽取该子树所包含的所有文本信息。

3 短信平台描述

根据欧洲电信标准协会(ETSI)制定的 GSM 协议标准，采用 PDU 编码，在计算机串口通信的基础上，结合多线程技术和缓冲池技术，使用事件驱动方式开发出了计算机与短信设备的通信组件，该组件可以实现通过计算机 RS232 串口与工业级短信设备(GSM MODEM)或与支持 AT 指令的手机之间的通信和数据传输。

4 短信推送平台描述

4.1 平台系统实现

短信推送平台主要是针对杜鹃花，梨和猪这三种农产品的供求信息进行 web 抽取，然后将抽取结果以短信的方式发送给特定用户。发送内容包括联系人、联系方式、信息类别等。该平台主要包括 2 个部分，供求信息 web 抽取部分和短信发送部分。其中供求信息 web 抽取部分由 3 个模块完成：网络爬虫模块，规则制定执行模块和抽取结果保存模块。

(1) 网络爬虫模块。该模块是从一个原始或若干初始网页的 url 种子开始，采用广度优先策略不断取出 url 对应 html 页面中的子链接。同时避免网络环路和重复 url 连接等问题。

(2) 规则制定执行模块。基于内容和 web 文档结构路径(DOM)的方法制定抽取规则。读取 url 对应的 html 文本信息，根据抽取规则定位信息块，并将块内 HTML 解析为 DOM 树，抽取出块内子信息。最后基于关键字匹配算法过滤出三类农产品的供求信息。

(3) 抽取结果保存模块。该模块是将抽取的信息保存起来，方便以后对抽取信息的进一步处理，如检索和分类等。在此我们使用 mysql 数据库进行存储。

短信平台主要有以下几部分组成：

(1) 设备管理。显示使用的设备名称及其参数，修改、删除设备参数，启动、停止设备。

(2) 系统设置。设置是否开机自动运行，首次发送失败是否自动重发，系统日志显示管理，登录密码修改等。

(3) 短信发送。定时扫描数据库中的相应数据表，发现未发送的短信后就取出该表中的内容发送给特定用户。不能一次发送成功时，可在限定次数内重新发送。

4.2 实验结果

选取中国农业网等几大农业门户网站的农产品供求信息，作为我们实验的对象。信息检索中的度量标准召回率和准确率作为对实验效果的评测。

表 1 供求信息抽取实验结果

Web resource	total	real	true	recall	precision
中国农业网	216	216	204	1.000	0.944
金农网	209	209	197	1.000	0.943
中国杜鹃花交易网	96	96	95	1.000	0.989
中国猪网	124	124	118	1.000	0.952

召回率是指实际抽取出来的供求信息的数目占所有需要抽取的供求信息数目的比率。准确率是指抽取出来的正确的信息占抽取出来的所有供求信息的比率。在表 1 中，total 表示待抽取网页的总的数目；real 表示抽取出来的供求信息的数目；true 表示正确的抽取结果的信息数目；recall 表示召回率；precision 表示准确率。

由表 1 可知，本文所提出的抽取算法能够很好地制定抽取规则并能够准确的抽取所需要的内容。但农产品种类在抽取的结果信息进行过滤时是采用的基于关键字匹配的算法，不能对文本信息进行理解，所以准确率略有所下降。

5 结束语

本文介绍在基于内容和 web 文档结构路径(DOM)的基础上，设计了一种新的抽取规则，结合两者抽取技术的优点，来实现对农产品供求信息的抽取。并通过 GSM 技术实现了基于短信平台的主动服务系统。

(下转第 60 页)

(上接第 56 页)

参考文献

1 Muslea I. Extraction Patterns for Information Extraction Tasks: A Survey[C]. AAAI - 99 Workshop on Machine Learning for Information Extraction, 1999.

2 李效东、顾毓清，基于 DOM 的 Web 信息提取[J]，软件学报，2002,25(5).

3 Eikvil L. Information Extraction from World Wide Web——A Survey[R]. Norwegian Computer Center, Tech. Rep: 945, 1999 - 07.

4 World Wide Web Consortium: The Document Object Model[EB/OL]. http://www.w3.org/DOM, 2004.

5 Chang Chiahui, Lui Shaochen. IEPAD: Information Extraction Basedon Pattern Discovery[C]. Proceedings of the Tenth International Conference on World Wide Web, Hong Kong, 2001 - 05.