25卷 第6期
2008年6月

微电子学与计算机
MICROELECTRONICS & COMPUTER

Vol.25 No.6
June 2008

基于多Agent技术的分布式信息抽取系统研究

魏保子[1,2],王儒敬[1]

(1 中国科学院 合肥智能机械研究所,安徽 合肥 230031;
2 中国科学技术大学 自动化系,安徽 合肥 230026)

摘 要:讨论了信息抽取的必要性及其现状,并提出一个基于多Agent技术的分布式信息抽取系统模型.系统主要有信息抽取Agent、数据清洗Agent、数据保存Agent等以及相应的知识库组成.并采用分而治之的思想,把信息抽取中遇到的问题分解,分配到各个Agent去完成.提出一种新的规则表示方法,抽取规则可以根据网页结构进行调整,该系统具有一定的自适应性.

关键词:多Agent技术;信息抽取;分布式;抽取规则

中图分类号:TP391 **文献标识码**:A **文章编号**:1000-7180(2008)06-0018-04

Distributed Information Extraction System Research Based on Multi-Agent Technology

WEI Bao-zi[1,2], WANG Ru-jing[1]

(1 Institute of Intelligent Machines, CAS, Hefei 230031, China;
2 Department of Automation, University of Science and Technology of China, Hefei 230026, China)

Abstract: The information extraction necessity and the present situation is discussed, and a distributed information extraction system model based on the multi-agent technology is proposed. This system is composed of the user interface agent, the agent manager, the information extraction agent, the data clean agent; the data preserves agent and corresponding knowledge library. Using thought called divide and conquer, the question which was met in process of information extraction is decomposed and assigned to each agent to solve. At the same time, the extraction rules are able to adjust themselves to the homepage structure; this system has certain adaptability.

Key words: multi-agent technology; information extraction; distributed; extraction rule

1 引言

随着互联网信息资源的不断快速的增长,信息资源的闲置和资源浪费愈加严重,另一方面用户对互联网信息的需求也将会日益增强.而且发现除了大众化的需求,每个互联网用户对个性化服务的要求越来越强烈.根据每个互联网用户的具体信息需求,为其提供相应的信息资料或其他信息服务,正是互联网个性化服务的一个主要内容.

为了实现个性化信息服务,就必须能够从多个信息源中获取用户所需的信息内容,并有效地把它们整合到一起.问题是目前互联网上的信息均是基于html语法而编写的Web网页来进行的,而Web网页的内容描述是针对互联网用户浏览而进行的,格式多种多样.研究如何从Web网页中抽取信息内容,并把此信息内容变成格式良好的数据,就成为信息获取研究中的一个重要内容.

为此文中提出了一个基于多Agent技术的分布式信息抽取系统[1].

收稿日期:2007-08-03
基金项目:国家"八六三"计划项目(2006AA10Z237);国家科技支撑计划项目(2006BAD10A1410)

2 基于多 Agent 技术的分布式信息抽取模型

随着互联网信息的急剧增加,出现了大量的半结构化文本信息资源,典型的就是网页资源.为了合理的利用这些信息资源,使之变成结构化良好的数据,人们提出了各种各样的信息抽取方法或系统,其中比较常用的方法是基于模式匹配的信息抽取,除此之外具有代表性的还有:基于层次结构的信息抽取模型、基于概念模型的多记录信息抽取模型[2].

每种方法都有自己的优缺点,以上两种方法虽在一定程度上可以解决问题,但是当网页结构变化时,所分析的抽取规则就会失效,另外由于不同的网页结构千差万别,而且 html 语法不是很严格,即使存在语法错误也不会影响显示效果.所以基于层次结构的信息抽取实现起来比较繁琐,适用范围较小,对于不同的网页要分别对待,而且并不能很好的适应如此复杂的网络环境.

在这里需要指出的是:一个实用的 Wrapper 方法应该具有处理多变的互联网环境能力,如:格式有错的网页,网络连接失败,网页结构发生变化等等.另外据统计互联网上大部分信息存在于动态网页中,而这些网页是一般的信息采集工具所抓不到的,这样就使得互联网上大量的信息得不到合理得利用,为了解决上述问题和其他方法的不足,本模型采用比较常用的基于模式匹配的方法和文档结构路径相结合,针对动态网页,如 jsp、asp、php 等格式的动态网页,利用多 Agent 技术,分层设计即读取层、抽取层、映射层、同时每层分别有相应的知识分别是读取规则、抽取规则、和映射规则.为了解决此模型的不足,采用分而治之的思想,把模型中的问题分解为多个小问题分别解决,因此模型还包括读取和映射规则的可视化生成工具,使得信息抽取变的更智能化,方便化,可以随意得到网上所需要的任何信息.

该模型中 Agent 实体主要有:用户接口 Agent、Agent 管理器、规则修复 Agent、信息抽取 Agent、数据过滤 Agent、数据保存 Agent.其总体框图如图 1 所示[3].

系统特点:

(1) 系统采用分布式技术,在进行信息抽取时各个抽取器工作方式采用动态分配方式,即各个抽取器在 Agent 管理器的协调下共同完成目标页面的采集.节省了系统开销,提高了抽取效率.

(2) 管理调度 Agent 根据用户的需求在知识库

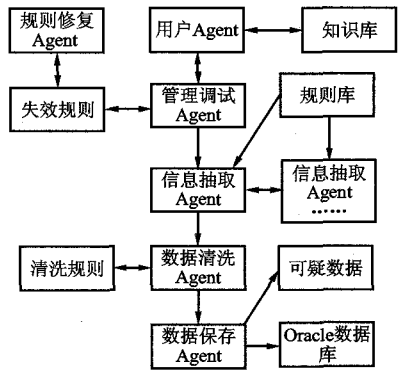


图 1　基于多 Agent 技术的信息抽取模型框图

中查找和用户领域对应的 URL 种子,然后找到和 URL 种子对应的规则库.接着产生相应的信息抽取 Agent,此时信息抽取 Agent 就按照规则进行抽取.

(3) 当抽取到的数据和规则库中所标记的最后一条数据一样时,就停止对本 URL 的信息抽取并把所抽取到的最新的一条数据写入规则库,同时对 Agent 管理器发送信息抽取完毕消息,于是 Agent 管理器就召回此信息抽取 Agent.

(4) 同理,如果用户还需要其他领域的信息,工作流程同上.当信息抽取 Agent 结束对一个 URL 信息抽取的时候,会启动一个数据过滤 Agent 完成对冗余数据的清洗,之后才把数据保存到数据库.至此一个抽取过程结束.

(5) 如果在开始抽取的时候,信息抽取 Agent 发现网页结构变化,无法抽取到数据会把此规则的状态标识"无效"写入规则库,同时 Agent 管理器会调用规则修复 Agent,对规则修复.

2.1 用户接口 Agent 和 Agent 管理器

用户接口 Agent 是与用户交互的接口,它根据用户的输入分析出用户的需求,然后返回给用户等待用户确认,如用户输入"安徽小麦价格",则用户 Agent 就会根据用户输入内容加以分析,处理并从 URL 库中选择含有安徽小麦价格数据的 URL 种子,用户确认之后,用户接口 Agent 就会把用户的确认信息发送到给 Agent 管理器,然后 Agent 管理器就会根据此 URL 种子与相应的抽取规则产生一个价格数据抽取 Agent.Agent 管理器是本模型的核心,所有的 Agent 要由它管理,保证了各个 Agent 之间的协调.

2.2 信息抽取 Agent 与规则库

信息抽取 Agent 是本模型的主要组成部分,它主要分三个步骤进行信息抽取:即读取网页,信息抽

取和信息映射.读取网页是发送一个 http 请求,然后取回相应的网页.信息抽取是负责读取层所读取的网页,抽取出所需的数据.信息映射主要是负责将信息抽取层所抽取的内容映射到有映射规则定义的数据库字段中[4].

同时,和每个步骤相连还有相应的规则库,分别为读取规则,信息抽取规则,信息映射规则,它们都放到一个规则库中.每个规则库和 URL 种子是一一对应的关系,也是和一个具体的 Agent 是一一对应的.规则的具体表示形式采用 xml 格式,并把读取规则,信息抽取规则,信息映射规则放到一个规则库中统一用 xml 格式表示.

其中信息抽取的关键是有效信息块的路径的确定.在生成规则时首先把不规范的 html 转化为格式良好的 xml,取得有效信息块的路径.为了取得有效信息块内的用户所需信息,在有效信息块内应用模式匹配的方法,抽取出有效信息.同时和模式匹配方法对应的还有一个文档路径(相对于有效信息块路径的相对路径),这样当模式匹配方法失效时,此规则可以自学习具有自适应性.规则具体表示如下所示:

<TableName>AgriPrice</TableName></Init>
<InfoPath>**html.body.table**[2].**div**[1] <InfoPath>
<Field-one><Before>***<td***</Before>
<After>***</td>***</After>
<InfoPath1>**tr**[1:**n**].**td**[1].**font**[1]</InfoPath1>
<DataFeild>Name</DataFeild><FieldName>名称</FieldName></Field-one>
<Field-two>
<Before>***<td class="z">***</Before>
<After>***</td>***</After>
<InfoPath2>**tr**[1:**n**].**td**[2].**font**[1]</InfoPath2>
<PageField>Price</PageField>
<DataFeild>Price</DataFeild>
<FieldName>价格</FieldName></Field-two>
<LastItem><Data>上次抽取的最后一条记录</Data>
</LastItem>

注:粗体且加有下划线的内容为有效信息块路径,粗体无下划线为具体信息的路径.粗体且斜体的为所要抽取信息的前后标识.

信息抽取 Agent 进行抽取时,首先读取在规则库中和此 Agent 对应的规则,并把规则的各个参数读入内存,得到 URL 种子所对应的 html,然后进入抽取阶段,首先把 html 转化为格式良好的 xml 的形式,根据 InfoPath 元素内容取得有效信息块,分析有效信息块 html 按照模型进行匹配,即和(B1,A1),(B2,A2),(B3,A3)…进行匹配.如取出有效信息块中和 Field-one 的子元素 Before ,After 及<Field-two>下的子元素 Before,After 匹配的有效信息块中的内容记作一条记录.如果有效信息块中还有这样的模式,要一直循环下去直到再也找不到匹配的模式为止.此时进入了信息映射阶段,按照信息映射规则,把采集到的数据和数据库里的字段进行映射.并把从网页中得到的记录以 xml 的格式存储起来,和数据库里的字段对应.

根据上面的规则数据存储格式如下:

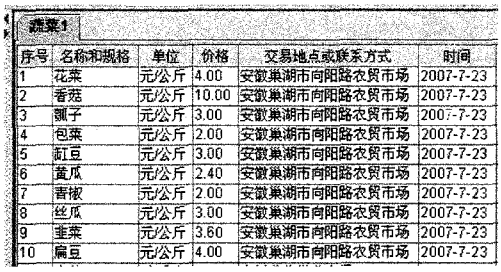
<? xml version="1.0" encoding="UTF-8"? >
<Init><TableName>AgriPrice</TableName><Init>
<Data><Name>黄瓜</Name>
<Price>2.00元/公斤</Price></Data>

因为是动态网页,所以信息抽取 Agent 改变 URL 中的参数读取下一页,然后同上一样取出所有记录,如此下去直到所抽取的记录和上次抽取是的最后一条匹配为止.

在抽取过程中如果发现抽取规则失效,如:在读取规则有效的前提下,无法抽取到数据或找不到匹配的模式就认为抽取规则失效.此时信息抽取 Agent 就会对 Agent 管理器发送关于规则失效的消息,于是 Agent 管理器召回此信息抽取 Agent,并把此规则标识为"失效状态",放入失效规则库,然后产生规则修复 Agent,规则修复 Agent 对此规则测试,在抽取过程中如果发现可以找到有效信息块,但是在抽取具体的有效信息时在有效信息块内进行模式匹配无法抽取到信息.此时读取此有效信息的相对路径进行抽取,如果能成功抽取可根据此路径重新修改匹配规则,并写入规则库.否则在有效信息快路径下的所有子结点查找和对应模式匹配的结点,找到之后修改具体信息相对路径,写入规则库.通知 Agent 管理器此规则修复成功,于是产生信息抽取 Agent 重新开始抽取.如无法找到有效信息块,则根据 before 和 after 中的值对规则修复,如无法修复则把此规则放在失效规则库,等待人工修复.同时,规则修复 Agent 会定时对规则库中的规则进行测试并试图修复.图 2 为信息抽取 Agent 执行抽取任务后的可视化结果.

2.3 数据处理 Agent

互联网上的数据来源多种多样,并不能保证所有的数据都是正确的,如果从网上抽取到的数据是错误的,就失去了抽取的意义.所以对抽取到的数据进行必要的清洗是非常必要的.



蔬菜1

序号	名称和规格	单位	价格	交易地点或联系方式	时间
1	花菜	元/公斤	4.00	安徽巢湖市向阳路农贸市场	2007-7-23
2	香菇	元/公斤	10.00	安徽巢湖市向阳路农贸市场	2007-7-23
3	瓢子	元/公斤	3.00	安徽巢湖市向阳路农贸市场	2007-7-23
4	包菜	元/公斤	2.00	安徽巢湖市向阳路农贸市场	2007-7-23
5	缸豆	元/公斤	3.00	安徽巢湖市向阳路农贸市场	2007-7-23
6	黄瓜	元/公斤	2.40	安徽巢湖市向阳路农贸市场	2007-7-23
7	青椒	元/公斤	2.00	安徽巢湖市向阳路农贸市场	2007-7-23
8	丝瓜	元/公斤	3.00	安徽巢湖市向阳路农贸市场	2007-7-23
9	韭菜	元/公斤	3.60	安徽巢湖市向阳路农贸市场	2007-7-23
10	扁豆	元/公斤	4.00	安徽巢湖市向阳路农贸市场	2007-7-23

图2　可视化结果示例

数据处理Agent主要包括数据清洗和数据保存Agent,数据清洗和数据保存Agent协调工作,它们共用一块内存区域share,数据清洗Agent清洗完一条数据就会把他放在share中,当数据保存Agent发现share中有数据存在时就把它写入oracle数据库中,然后从share中清除此条数据.关于清洗规则放在知识库中,符合规则的便保留.如发现可疑数据(如:发现白菜价格是100元/公斤便认为是错误数据,发现小麦是3元/公斤便认为是可疑数据)便写入可疑数据表中,等待用户判断,经过对数据进行分析挖掘,去发现其中的知识.并把知识转化为具体的规则,然后把用新的规则对可疑数据进行处理,并把规则写入清洗知识库[5].

3　实验测试

选取5个农业网站的价格数据(主要是表格数据)进行测试,其中有2个网站结构发生了变化.实验对两种系统进行了测试,一种是基于模式匹配的一般的信息抽取系统,另一种是本系统.实验结果统计如表1、表2所示.

表1　本系统实验统计结果

	正常	正常	正常	变化	变化
召回率/%	100	100	100	92.9	71.4
准确率/%	100	100	100	91.6	80.9

表2　一般系统实验统计结果

	正常	正常	正常	变化	变化
召回率/%	100	100	100	5	0
准确率/%	100	100	100	20	0

实验过程中发现,对于一般系统,只要网页结构稍微变化抽取规则就会失效,几乎抽取不到任何数据.

对于本系统,在两个已经变化的网页中,其中一个已经完全变化,抽取规则标注的有效信息块路径已经失效,但是经过规则修复Agent修复之后,规则可以使用,只是召回率比较低.另一个已经变化的网页则只是局部部分变化,模型则可以根据网页结构变化调整抽取规则,且抽取规则工作良好,具有较强的适应性.对于本系统,当网页结构变化时,准确率虽然没有达到100%,但是经过数据处理Agent处理之后,数据进入数据库准确率基本上达到100%.对于结构没有变化的网页,只要模式匹配规则完全正确,本系统和一般系统召回率都可以达到100%.同时发现,本系统抽取信息效率远远高于一般系统.

4　结束语

本模型利用多Agent技术及模式匹配和文档结构路径相结合的方法进行信息抽取.同时采用了分而治之的思想,把在常规的进行信息抽取的过程中遇到的问题分解,各个击破.同时本模型具有一定的自适应性,可以根据网页结构的变化调整抽取规则.实验证明,本系统能够很好的进行海量信息的抽取,满足用户需求.

参考文献:

[1] Jennings N, Faratin P, Norman T. Autonomous agents for business process management[J]. International Journal of Applied Artificial Intelligence,2000,14(2):145-189.

[2] Michael Wooldridge. 多Agent系统引论[M].石纯一,张伟,徐晋晖,等译.北京:电子工业出社,2003:15.

[3] 张云勇,刘锦德.移动Agent技术[M].北京.清华大学出版社,2003:44.

[4] Chen L, Sycara K. Web mate: a personal agent for browsing and searching[C]// Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems. USA: Pennsylvania State University, ACM, 1998(5):132-139.

[5] 李向阳,张亚非.一种基于自举原理的语义模式自动获取方法[J].微电子学与计算机,2005,22(2):188-192.

作者简介:

魏保子　男,(1983-),硕士研究生.研究方向为智能Agent、信息获取、复杂系统.

王儒敬　男,(1964-),研究员,博士生导师.研究方向为数据挖掘、人工智能、信息获取等.