Research and Implementation of the Active Interactive Web Information Mining Technology based on Virtual Machine

WANG Yuan-sheng[1,2,3], WU Hua-rui[1,2,3*],GU Jing-qiu[1,2,3], HUANG Feng[1,2,3], LUO Lei[1]

[1)] Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China

[2)] National Engineering Research Center of Information Technology in Agriculture, Beijing 100097, China

[3)] Key Laboratory of Agri-informatics, Ministry of Agriculture, Beijing 100097, China

***Abstract*— Intelligent web information mining is an effective method of agricultural information resources integration. Web crawler based data mining technology is a popular way of resource integrate in recent years. But, as the web technique developing and the information resources sites update frequently, the limitations of web crawlers relied mining methods are becoming more obvious. By sufficient investigation and research, in this paper, a method of virtual machine based active interactive network information mining technology was proposed, and solved these problems. This system has been used in the National Modern Agriculture City for Science and Technologies system in agriculture market information network services, and achieves an effective result.**

***Keywords*— virtual machine, interactive, web information, active mining**

基于虚拟机的交互式网络信息主动挖掘技术研究与实现

王元胜 [1,2,3]　吴华瑞 [1,2,3*]　顾静秋 [1,2,3]　黄锋 [1,2,3]　罗磊 [1]

[1)] 北京农业信息技术研究中心，北京 100097，中国

[2)] 国家农业信息化工程技术研究中心，北京 100097，中国

[3)] 农业部农业信息技术重点实验室，北京 100097，中国

摘 要 网络信息智能挖掘是农业信息资源整合中的有效手段，基于网络爬虫的挖掘技术是近年来流行的整合方式之一，但随着信息源站点频繁更新及 Web 技术的不断发展，单纯依赖爬虫的挖掘技术的局限性日益突出，本文通过充分的调研分析，提出了基于虚拟机的交互式网络信息主动挖掘技术，有效地解决了上述问题，并且在国家现代农业科技城农业市场信息网联服务中得到初步应用，取得了较好的应用效果。

关键词 虚拟机，交互式，网络信息，主动挖掘

1. 引言

农业信息量大、面广而分散，农业人员难以快速、准确、有效地获取农业信息，成为制约农业信息资源深入整合应用的难点问题，运用网络信息挖掘技术，从各个农业站点析取数据，汇聚到中央存储数据库服务系统，实现自动、协同的资源整合建库与共享，可在一定程度解决上述问题[1]。近年来，以网络爬虫为核心的网络信息智能挖掘技术是农业信息资源整合中的一种常用手段，可以在分析 DOM tree 结构的基础上，挖掘数据项的起始和结束标记区间信息，建立信息源站点与目标数据库之间的映射关系，通过网络爬虫逐项析取各项数据到中央数据库存储系统中，实现网络信息资源自动汇聚[2-3]，但随着信息源站点频繁更新，Web2.0、Web3.0、富客户端等技术发展，单纯依赖爬虫的挖掘技术的局限性日益突出，维持持续稳定自动化汇聚的维护量增加，迫切需要新型灵活的挖掘技术以适应这一形势的需要[3-10]。本文在充分调研分析农业网站建设现状的基础上，提出了在虚拟机环境下交互式挖掘网

基金项目：国家科技支撑计划“农村农业信息资源整合关键技术集成与应用(2011BAD21B02)”和“国家农村信息专业服务平台开发与应用(2013BAD15B04)” 项目

络信息的技术方法，即，基于数据项在网站中分布位置的唯一性标识，定义数据项与目标数据库字段间的对应关系，在虚拟机中建立数据动态汇聚的 Web 服务，自动将各站点的数据整合到中央数据库管理系统中，有效地实现了上述目标。

2. 系统分析与设计

通过网络信息挖掘技术进行数据资源整合的数据主要来源于各省农业网站，这些网站采用的 Web 框架主要为基于 html 标签语言和在 html 语言中嵌入富客户端控件（Flash 和 Slivelight) 两种，从数量上看前者比例占大多数，但后者的比例在逐步上升，爬虫技术对于前者仍然适用，对后者来说则不适用，但随着来源站点的更新换代，带来 html 标签的噪声及无规律性信息在增加，因而爬虫挖掘方式的适应性在显著降低。因此本文结合资源整合研究任务，分析了近年来各农业网站站点 Web 语言规律，重新定义了来源网站与目标数据库的数据映射关系为目标，找到一种通过数据项位置的唯一性标识来建立映射关系的数据挖掘方法，然后将这种网络信息主动挖掘技术按“基于标签(tag)的数据项唯一性标识”和“基于图形识别(OCR)的数据项唯一性标识”两种类型，设计不同的方法，实现数据按专题定向汇聚入库、协同共享的资源“共建共享”目标。(图 1)。

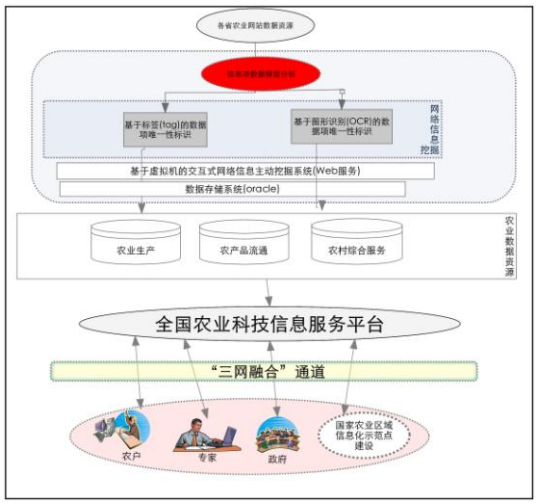

图1 网络信息主动挖掘技术总体框架

2.1 网络信息主动挖掘基础设施

主要包括虚拟机系统、Oracle 数据库系统、数据汇聚平台。部署在全国农业科技信息服务中心数据交换服务器上，以外网方式基于 http 与各省农业站点互联互通，通过 amf 协议与数据汇聚平台进行数据通讯，提供资源自动汇聚 Web Service 服务。

2.2 虚拟机系统

采用 VMware 虚拟机系统，部署在现代农业科技城数据交换服务器上，在虚拟机里部署 Web 交互式数据解析系统，按照数据项唯一性标识解析的数据通过 amf 协议，调用数据汇聚平台的 Adobe Web Service 接口，实时存储到网络服务中心的 Oracle 数据库系统中。

2.3 数据库系统

采用 Oracle 9i 数据库系统，部署在国家现代农业科技城“全国农业科技信息服务中心”虚拟机所在的数据库服务器上，在数据库中建立专门的数据汇聚用户、专题数据库表、数据视图，通过数据汇聚平台上的 Adobe Web Service 提供各专题的汇聚接口，接受虚拟机中的交换系统从各省农业站点中析取的数据资源。

2.4 数据汇聚平台

采用 Tomcat + LCDS (Live Cycle DataServices) 的方式实现资源汇聚，在 Tomcat 6.0.14 中部署数据访问框架 LCDS，基于 AMF 连接协议为每个数据专题定义远程调用的 Java 对象，速度和效率优于 HTTP 的文本模式，可实现各农业站点数据的实时解析与远程快速整合入库。

3. 网络信息主动挖掘相关技术

本文设计的网络信息主动挖掘主要涉及交互式系统、数据项唯一性标识、主动挖掘和数据汇聚等技术。

3.1 交互式系统

在虚拟机上部署交互式程序模块/插件与 IE 浏览器交互，得到来源网站数据项的位置信息，供建立来源网站数据项与数据中心目标数据库映射关系使用。对于本文设计的基于图形的数据解析方式，交互式系统还需要提供来源网站数据项及其所在文档的图形信息，保存在缓存文件夹中，供主动挖掘程序使用。

3.2 数据项唯一性标识

通过交互式系统，得到数据项的位置信息，本文标识数据项有两种方法：一种是根据数据项所在 html 语言中的标签及其在文档中出现顺序号来定位；另一种是根据网页内容在终端计算机屏幕中的图形坐标位置，通过交互式插件，获取数据项的左上角象素坐标及右下角象素坐标，来进行唯一性标识定位。

3.3 主动挖掘

在虚拟机中，研发部署数据主动挖掘的 Java 程序，根据数据项的唯一性标识将数据逐项挖掘出来:（1）标签（tag）方式主动挖掘，从数据项唯一性标识映射模型中得到数据项的标签类型及序号，从来源网站析取网页中将数据挖掘出来；（2）图形方式主动挖掘，从数据项唯一性标识映射模型中得到数据项的象素坐标及相关图形文件，以光学字符识别（OCR）的方式，从交互式系统产生的来源网站镜像图片中将数据挖掘出来。

3.4 数据汇聚

在虚拟机中的数据挖掘程序中，提供通过 Adobe AMF 协议调用数据汇聚平台专题数据存储服务的 Web Service 接口的功能模块，实时将挖掘出数据汇聚到全国农业科技信息服务中心数据库对应专题数据库中。

4. 网络信息主动挖掘技术实现

最后，本文采用 JavaEE 技术，基于 Adobe LCDS(Live Cycle Data Services)技术框架，采用 Tomcat 6.0.14 作为数据汇聚平台的 Web 服务器，对农业领域的信息资源进行编目，建立每个数据专题的元数据描述信息和 Adobe Web Service 数据服务，通过虚拟机中交互式数据挖掘系统，整合各省主要农业网站的信息资源，面向全国共享应用。

为详细阐述网络信息主动挖掘技术实现过程，下面以农产品市场行情为例，详细介绍各环节的技术实现流程：

4.1 数据项位置唯一性标识方法

以农产品市场行情为例，在数据汇聚平台的数据库中，为农产品市场行情专题定义了其数据结构：

表 1 农产品市场行情专题数据库数据结构

字段名称	字段别名	字段类型
KJC_PRICE_ID	序号	NUMBER(10)
NAME	农产品名称	VARCHAR2(255)
LOW_PRICE	最低价格	NUMBER(10,2)
MID_PRICE	中间价格	NUMBER(10,2)
HIGHT_PRICE	最高价格	NUMBER(10,2)
SPEC	规格	VARCHAR2(255)
UNIT	单位	VARCHAR2(255)
TIMES	发布时间	DATE
ORIGIN	URL 来源	VARCHAR2(255)
CREATTIME	汇聚时间	DATE
NETNAME	来源网站	VARCHAR2(255)

按本文设计原理，通过交互式系统，获得在来源网站中该专题来源网站中数据项唯一标识如下：

表 2 数据项位置唯一性标识方法

数据项	定位方法	标识 (tag/ocr)	位置	备注
产品名称	标签	td	9+i*8	i 为整数从 0 开始递增
最低价格	标签	td	10+i*8	i 为整数从 0 开始递增
中间价格	标签	td	11+i*8	i 为整数从 0 开始递增
最高价格	标签	td	12+i*8	i 为整数从 0 开始递增
规格	标签	td	13+i*8	i 为整数从 0 开始递增
单位	标签	td	14+i*8	i 为整数从 0 开始递增
发布日期	标签	td	15+i*8	i 为整数从 0 开始递增

如上表所示，各农产品的名称可以从来源网站 html 中第（9+i*8）个<td>标签中包含的内容中得到，相应的“最低价格”可以从第（10+i*8）个<td>标签中包含的内容中得到，由于加上了序号，标签中内容具有了唯一性，依此，各农产品的市场信息都可以根据上述标识方法得到。以图形方式建立唯一性标识的流程和方法与标签方式类似，只是标识类型为 pixel(象素)，位置是数据项左上角坐标和右下角坐标形成的区域。

4.2 专题数据模型

在数据汇聚平台中，为“农产品市场行情”专题数据库定义了以下数据模型（图 1-3）：

图 2 农产品市场行情专题数据对象模型

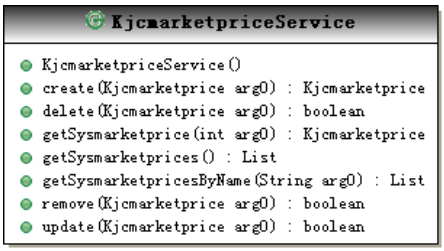

图 3 农产品市场行情专题数据服务模型

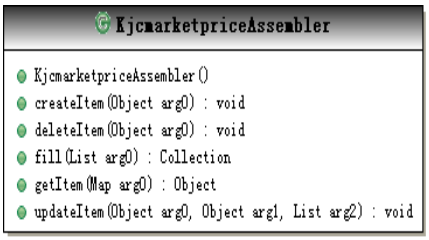

图 4 农产品市场行情专题数据服务适配模型

其中，图 2 为“农产品市场行情”专题数据库的 Java 对象模型，模型中的属性与数据库字段一一对应，模型中方法负责根据输入输出参数维护数据库字段值；图 3 为农产品市场行情专题数据服务模型，在 Adobe LCDS(Live Cycle Data Services)数据访问框架中对外提供 Web Service 服务，本质上会调用图 2 所述的 Java 对象模型；图 4 为农产品市场行情专题数据服务适配模型，将前面两个模型对应的 Java 类安装在 LCDS 中需要用到。

4.3 数据析取

完成上述步骤技术实现后，在虚拟机中研发数据析取程序，对内接受交互式系统挖掘的数据，对外与数据汇聚平台连通，将数据传送到数据中心，核心业务逻辑如下（Java 伪代码方式表示）：

```
public class iMacDataExtract {
  public static void main(String[] args) {
      AMFConnection amfConnection=new AMFConnection();
      System.out.println("开始解析 ");
        try {
              System.runFinalizersOnExit(true);
              ActiveXComponent v2i = new ActiveXComponent("iMacDataExtract");//交互式模块
              System.out.println("转区初始化");
              v2i.invoke("v2iInit");
              System.out.println("执行指令");
              v2i.invoke("v2iPlay", "CMD=URL GOTO=http://web.xmfack.com.cn/price/?md=3&page=738"); //call v2iPlay()
              v2i.invoke("v2iPlay", "CMD=WAIT SECONDS=10"); //call v2iPlay()
  for(int pn1;pn1<1000;pn1++){
        for(int zn=0;zn<40;zn++){
              v2i.invoke("v2iPlay", "CMD=TAG POS="+Integer.toString(9+zn*8)+" TAG=TD CONTENT=* EXTRACT=TXT"); //农产品名称
              String name = v2i.invoke("v2iExtract").toString();
                    (略)
              SimpleDateFormat f=new SimpleDateFormat("yyyy-MM-dd hh:mm:ss");
              java.sql.Date times=java.sql.Date.valueOf(strTimes);
              String origin = "http://web.xmfack.com.cn/price/?md=3";
              java.util.Date creattime = new java.util.Date();
              String netname = "北京市新发地农产品交易网";
              try{
                    amfConnection.connect("http://数据中心IP地址:端口号/fc-ds-samples/messagebroker/amf?id=my-amf&amp;qmp=true");
                    Kjcmarketprice symarketprice = new Kjcmarketprice();
                    symarketprice.setName(name);
                    symarketprice.setLow_price(Double.parseDouble(low_price));
                          (略)
                    Object result=amfConnection.call("kjcmarketpriceService.create",symarketprice);
              } catch(ClientStatusException cse){
                    System.out.println(cse.toString());
              }
              catch(ServerStatusException sse){
                    System.out.println(sse.toString());
              }
        }//for
                    v2i.invoke("v2iPlay", "CMD=TAG POS=1 TYPE=A CONTENT=TXT:下一页"); //翻页
  }// 2nd for
              v2i.invoke("v2iBye");
```

图 5 数据析取业务逻辑

4.4 数据挖掘及资源汇聚效果

在资源汇聚平台中，将 LCDS 包部署为 Tomcat 6.0.14 的一个 Web application，名称为 lcds-sample，在其下面的 WEB-INF 中的“remoting-config.xml”文件中配置以下信息：

```
<destination id="kjcmarketprice">
    <properties>
<source>flex.samples.sysmarketprice.KjcmarketpriceService</source>
    </properties>
</destination>
<destination id="kjcmarketpriceService">
    <properties>
<source>flex.samples.sysmarketprice.KjcmarketpriceService</source>
    </properties>
</destination>
```

将农产品市场行情数据对象模型、数据服务模型和数据服务适配模型对应的 Java 类安装在汇聚平台中，通过 AMF 协议实时响应虚拟机中交互式挖掘系统的数据析取汇聚操作，在平台中对汇聚到的数据进行深入应用分析，研发为农产品市场行情服务系统，对外提供服务：

图 6 基于网络数据主动挖掘方式的资源整合应用效果

本文通过上述一整套完整的技术流程，即可完成每个农业专题的数据析取与建库，实现资源自动汇聚的目标。

5. 结论与讨论

本文以农业信息资源整合为研究背景，研究了在虚拟机中建立 Web 数据析取交互机制的网络信息主动挖掘技术，提出了按数据项位置唯一性进行交互式挖掘的方法，将来自全国各农业网站的数据资源，向全国科技网络服务中心汇聚，建立农业生产、农业产业和农村综合信息服务等类别的专题数据库，对外共享服务，为农业信息资源整合提出了新的解决方案。

本文取得了以下几方面研究成果：

(1) 分析农业信息资源整合中网络信息挖掘技术难题，提出了以虚拟交互式析取为载体，以数据项位置唯一性标识为基础，构建农业网络信息挖掘资源汇聚平台的技术框架，初步研究了交互式系统、数据项唯一性标识、主动挖掘和分布式异构数据自动汇聚的技术方法，克服了传统网络信息挖掘的技术缺陷，解决了农业信息资源的高效汇聚的技术难题。

(2) 构建了 Tomcat + LCDS(Live Cycle DataServices) 的资源汇聚平台，提出了在 VMware 虚拟机环境下构建

Web 数据交互式析取系统的技术方法，基于 Adobe 二进制 Web Service 技术，研发集成了实用的资源汇聚服务总线及资源应用服务系统，为资源整合的深入应用提供了高效的技术支撑。

(3) 通过本文设计的资源汇聚平台，在国农现代农业科技城从 8 个主要省市整合了农业生产、农业产业和农村综合信息服务 3 大类别 30 个专题的农业信息资源，通过软件服务平台和信息门户向全国应用辐射，取得了较广泛的社会效益。

参考文献(References)

[1] Sun Su-fen, Luo Chang-shou, Zhang Jun-feng, Yu Feng, Zhang Shu-liang. “Research and application of agricultural information resource integration system”. *Anhui Agricultural Sciences*, 2007 ,35 (22) :6993 - 6994 , 6997

[2] Wen Xiao-yan, Ma Guan-gsi. “Flex and J2EE vertical search engine design and implementation”. *Computer knowledge and technology*,2011,7(10) :2293-2294, 2317

[3] Ceng Wei-hui, Li Miao. “Deep web crawler research review”. *Application of computer system*, 2008, (5): 122-126

[4] Peng Cheng, Wu Hua-rui, Zhu Hua-ji. “Rural industry information automatically acquired and visualization display methods”. *Computer Engineering*, 2011, 37(1): 270-272

[5] Li Si-ming. “Research on mining of network agricultural information based on Intelligent Agent”. *China Agricultural University*, 2003.

[6] Feng Yong. “Research on a parallel architecture for data mining”. *Chongqing University*, 2003.

[7] Kuang Xiang-ling. “Application of OLAM model active incremental data mining”. *Huazhong University of Science and Technology*, 2004.

[8] Liang Chuan, Wang Wen-sheng, Xie Neng-fu. “Application of agricultural information resources data mining”. *China Agricultural Science Bulletin*, 2009, 25, (11).

[9] Li Wen-pu Liao Gui-ping. “Application of data mining techniques in agricultural information website”. *China Agricultural Science Bulletin*, 2012, 28, (06).

[10] Li Chuan-xi. “Research on adaptive Web information extraction method based on Ontology”. *University of Science and Technology of China*, 2012.

作者简介：

王元胜（1973-），男，湖北武汉人，博士，副研究员，研究方向为农业信息技术研究与应用、3S 集成技术与应用；

通讯作者：吴华瑞（1975-），男，山东聊城冠县人，博士，研究员，研究方向为农业信息技术研究与应用；

顾静秋（1977-），男，河北秦皇岛人，硕士，副研究员，研究方向为农业信息技术研究与应用；

黄锋（1981-），男，河南南召人，博士生，助理研究员，研究方向为农业智能系统；

罗磊（1987-），男，河北保定人，硕士。

E-mail：wangys@nercita.org.cn