

基于智能手机的农业信息服务系统研究

孙小华¹, 王福顺², 杨会英², 赵艳², 王风国³

(1. 河北软件职业技术学院, 河北 保定 071000; 2. 河北农业大学 信息科学与
技术学院, 河北 保定 071001; 3. 河北巨鹿县供电公司, 河北 巨鹿 055250)

摘要:在分析国内外农业信息服务平台的应用现状基础上,研究了我国农民获取农业信息的特点,对基于智能手机平台的农业信息服务系统进行了技术构建,将 Web Services 数据挖掘技术与 Android 平台智能手机技术相结合,利用 Heritrix 框架进行特定农业信息的收集,通过 Web Services 数据挖掘为 Android 平台客户端提供最直接的农业信息服务,使移动通信网络和计算机网络更好地服务于农业科技信息推送与农产品的网络销售过程。

关键词:智能手机;数据挖掘;农业信息推送;农产品的网络销售

中图分类号:TP368 **文献标志码:**A **文章编号:**1671-1807(2013)06-0033-04

信息技术的迅猛发展,给人类生产生活带来了深远的影响,信息技术正在成为中国经济和社会发展的主要推动力量。越来越多的人借助信息网络提高工作效率、领略生活的丰富多彩,网络正在改变中国人的生活和工作方式。在信息时代的路上,广大的农民是相对落后的群体,而农民科技素质决定农业经济的发展,如何做到将最新的科学技术知识、先进的生产技术和致富信息及时的传递到农户手中,是目前研究的热点问题^[1]。

本文以强大的移动网络为基础,在智能手机平台上实现农业信息服务系统的开发,使广大农民轻松获取最新的农业科技信息,以此加强农业信息服务,提高农民科技素质;并着力改变农产品尤其是鲜活农产品供需双方的信息严重的不对称现象,改善单个农户、小规模农业组织在信息的搜集、占有和处理方面的弱势地位,解决农产品流通过程中的“小农户与大市场”的矛盾。

1 系统的整体构架

系统的总体架构如图1所示。它主要由服务、网络和客户三部分构成,其中服务器由数据层和中间服务器构成,网络主要有 Internet, 3G/WLAN 构成。数据层包括农业信息数据库和农业数据仓库,主要信息来自于与农业信息相关的互联网;中间服务器可自动利用 Heritrix 框架进行特定农业信息的收集,并结合基于 web 的数据挖掘技术获取农民直接关心的最新的农业科技信息、市场供求信息、国家政策信息、市

场行情等;客户端是面向 android 智能手机用户而开发应用,主要功能包括两大部分,一是农业信息浏览查询;二是农产品供需信息发布。

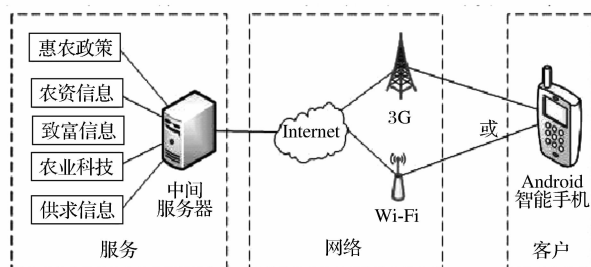


图1 系统总体架构

2 基于 Heritrix 的农业信息聚焦爬虫

针对农业信息服务网站的特点,制定聚焦爬虫(Focused Crawler)爬行策略,需要根据一定的网页分析算法过滤与主题无关的链接,保留有用链接并将其放入等待抓取的 URL 队列。然后它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL,并重复上述过程,直到达到系统的某一条件,从而提高爬行的精度和效率。

2.1 BKDRHash 算法优化多线程爬取

Heritrix 是多线程工作的,每个 URL 被一个线程处理,同时每个线程对应一条处理链,Heritrix 是使用 Berkely DB 来构建链接队列,队列放置在 Bdb-MultipleWorkQueues 中以 Key-Value 形式存在的、URL 被赋予 Key 值,Key 值相同的链接放在同一个

收稿日期:2013-04-01

基金项目:2013 年保定市科学计划研究与发展计划项目

作者简介:孙小华(1979—),女,河北涿鹿人,河北软件职业技术学院数字传媒系,讲师,硕士,研究方向:计算机软件。

队列中、默认情况下 Heritrix 使用 Host-name-QueueAssignmentPolicy 根据域名来生成 Key 值,即相同 Host 名称的所有 URL 都会被置放于同一个队列中,这就会使某一个队列太长,存在抓取效率低的问题。因为在 Heritrix 中,当一个线程从该队列中获取一个 URL 链接后,总是会先从队列的头部取出第一个链接,然后该队列阻塞,直到该链接处理完才从队列的头部取出下一个链接,而线程池中的其他线程因为没有可取的 URL 而处于等待状态,故造成 Heritrix 抓取效率低,会在某个时间段过后处于一直没有进度的状态^[2]。

因此,需要改变 Key 值的生成方式,使得所有的 URL 比较平均地散列到不同的队列中,以提高抓取效率。文中采用的方法是在 Heritrix 中扩展一个继承 Queue-Assignment-Policy 的子类,覆写其中的 getClassKey() 方法,该方法将一个链接对象处理后,再调用散列算法生成一个 Key 值,相同 Key 的链接存于同一个队列中。这里采用 BKDRHash 算法生成 Key 值,可以将所有的 URL 散列到不同的队列中去,提高抓取效率。

2.2 Heritrix 扩展

文中以河北省农业产业信息网 (<http://www.hbnongyechanyexinxi.org.cn>) 为例进行分析,改进 Heritrix 爬行策略。在 HeritrixProject 项目下建立 Nyxx.extractor 包,在此包内新建类 Hbnyxx Extractor,该类继承父类 Extractor,覆写 extract 方法。在 extract(CrawlURI curi) 方法中判断传入其中的参数,如果是该网的汇集页,则解析出页面下的链接部分,生成对应不同农业信息种类的信息汇集页面地址,然后将其加到 FrontierScheduler 处理的列表中,等待处理。

Frontier Scheduler 其功能是将 Extractor 中所分析出的链接加入到 Frontier 中以便后续处理,它存在于包 org.archive.crawler.postprocessor 中。扩展 FrontierScheduler 实现特定网页的抓取,在 HeritrixProject 项目下建 Nyxx.processor 包,在包下新建类 FrontierSchedulerForHbnyxxe,该类继承父类 FrontierScheduler,重写 schedule() 方法,只有满足条件的 URL 才允许加入到等待队列中^[3]。

利用 BKDRHash 算法进行 URL 散列,实现了面向农业信息的网页信息搜索,即提高了搜索效率又实现了网页的多线程抓取,为农业信息的数据挖掘提供数据基础。

3 Web Services 农业信息数据挖掘

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在的有用的信息和知识的过程,它是知识发现(Knowledge Discovery in Database, KDD)的关键步骤。将数据挖掘技术引入到农业信息服务领域,对数据源进行处理,深入寻找各种因素的相互联系,根据农民获取信息的特点,挖掘出农民最需要的农资供求、农业科技、市场供求、市场行情、致富信息、国家政策等信息。

3.1 Web Services 技术介绍

Web Services 也称为 Web Services 的服务描述,是使用标准的、规范的 XML 概念描述的,囊括了与服务交互需要的全部细节,包括消息格式、传输协议和位置。Web Services 体系结构是面向对象分析与设计的一种合理发展,是面向体系结构、设计、实现与部署而采用的组件化的合理发展,包括 SOA 中的 3 种角色和操作、UDDI 和 WSDL^[4]。SOA(service-oriented architecture),即面向服务的体系结构是一个组件模型,它将应用程序的不同功能单元(称为服务)通过这些服务之间定义良好的接口和契约联系起来。接口是采用中立的方式进行定义的,它应该独立于实现服务的硬件平台、操作系统和编程语言。这使得构建在各种这样的系统中的服务可以以一种统一和通用的方式进行交互。

Web Services 是采用 SOA 的体系架构来构建模型的,面向服务的体系结构中的协作如图 2 所示。

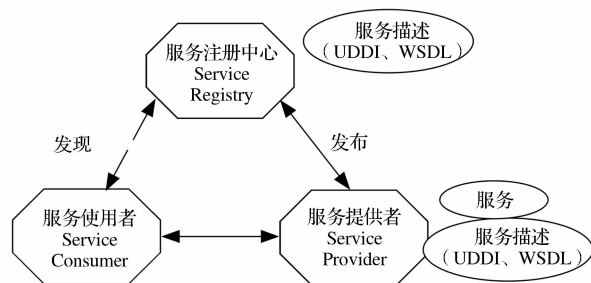


图 2 SOA 体系结构中的协作

SOA 体系结构中的角色包括:

1) 服务提供者: 是一个可通过网络寻址的实体, 它接受和执行来自使用者的请求。并将自己的服务和接口契约发布到服务注册中心, 以便服务使用者可以发现和访问该服务。

2) 服务注册中心: 是服务发现的支持者。包含一个可用服务的存储库, 并允许感兴趣的服务使用者查找服务提供者的接口。

3) 服务使用者: 可以是一个应用程序、一个软件

模块或需要一个服务的另一个服务。它发起查询注册中心里的服务,传输绑定服务,执行服务功能。

SOA 体系结构中的每个实体都扮演着服务提供者、使用者和注册中心这 3 种角色中的某一种(或多种)。

SOA 体系结构中的操作包括:

1)发布:发布服务描述以使服务消费者可以访问。

2)发现:查询服务注册中心来找到满足使用者标准的服务。

3)绑定和调用:服务使用者检索完服务描述后,还可继续根据服务描述中的信息来调用服务。

SOA 体系结构中的构件包括:

1)服务:可以通过已发布接口使用服务,并且允许服务使用者调用服务。

2)服务描述:指定服务使用者与服务提供者交互的方式,指定来自服务的请求和响应的格式。服务描述可以指定一组前提条件、后置条件或服务质量级别。

SOA 面向服务的体系结构,除了动态服务发现和服务接口契约的定义之外,还具有服务自包含和模块化、服务支持互操作、服务松散耦合、服务位置透明等特征,这些特征也是满足电子商务按需操作环境的要求的主要特征。

SOA 体系结构所涉及的技术包括 CORBA、DCOM 和 J2EE,其活动舞台已经扩展到 WWW 和 Web 服务。

3.2 Web Services 和数据挖掘相结合

系统基于 Web Services 的 SOA 进行设计,包括“左中右”三层,即用户应用层、业务逻辑层和数据层,具体的原理框架如下所示。系统图如图 3 所示。

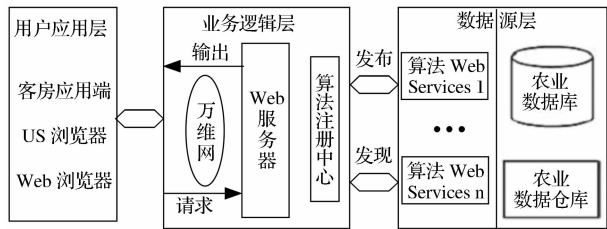


图 3 Web Services 农业信息数据挖掘原理图

1)数据源层:是业务逻辑层的数据源,是用户要进行挖掘处理的数据。系统中的农业数据库与农业数据仓库是运用 Heritrix 的农业信息聚焦爬虫技术进行收集和整理的。

2)业务逻辑层:是系统的核心,主要任务就是把

数据库的数据按照用户需要的方式读取出来。包括算法的发布、发现和运行,负责处理客户端的请求并返回结果。算法注册中心的主要功能是实现算法的发布和发现,主要包括对算法的基本描述。用户通过算法注册中心对算法进行搜索、查询,在算法描述的基础上,根据数据的实际情况,选择合适的算法,进行数据挖掘^[5]。另外,业务逻辑层通过接口编程,将业务逻辑处理和数据处理分离。关系映射框架封装了底层数据库访问细节,应用程序直接调用函数完成一般的数据库操作,而且屏蔽了对具体数据库的依赖。

3)用户应用层:直接面对用户,负责把数据以特定的形式显示给用户,是用户与系统之间进行交互的桥梁。它通过向用户提供访问接口,使用户可以方便地在算法注册中心搜索、查询满足数据挖掘的算法,并对数据进行上传、处理等管理,数据挖掘的结果也可以在该层进行表现。

基于 Web Services 的农业数据挖掘平台具有良好的可扩展性,能够提供统一访问接口以实现后台多种数据库,具有很强的灵活性。

4 Android 终端设计

Android 是 Google 公司开发的基于 Linux 内核平台的开源手机操作系统,也是一款基于智能手机平台的操作系统,是完全开放并且支持所有硬件的。Linux 移动开发平台包含了五大核心部分,分别是应用程序、应用框架、开发库、Android 运行环境与 Linux 内核。因此,要充分研究并掌握这些核心部分,并在此基础上基于 Java 语言开发并搭建符合构建理念和框架的移动农业信息服务平台^[6]。

4.1 搭建 Android 应用开发环境

首先下载和安装 JDK 以方便使用 Java 语言开发;其次,为使用 Android 的开发插件 Android Development Tools (ADT),需下载并安装 MyEclipse 和 Android SDK,通过 ADT 可以进行集成开发,包括代码的自动生成、调试、编译、打包、拖曳式界面生成等功能;另外,系统还需安装 Tomcat 模拟服务器。安装 Android SDK, MyEclipse, JDK, ADT 以及 Tomcat,并进行配置以完成 Android 的开发环境搭建^[7]。

4.2 终端模块划分

Android 终端可以通过网络(3G, WiFi)与农业信息服务中心服务器进行通信。图 4 描述了系统终端平台的四大模块。

1)农业资讯频道:包括资讯的发布、编辑、查看和删除。

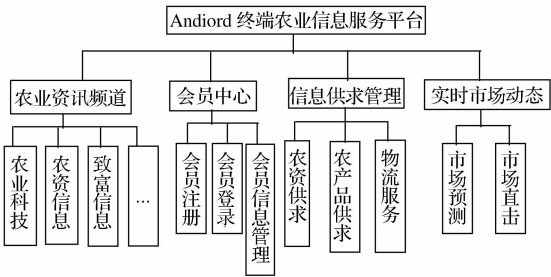


图 4 Andiord 终端功能框图

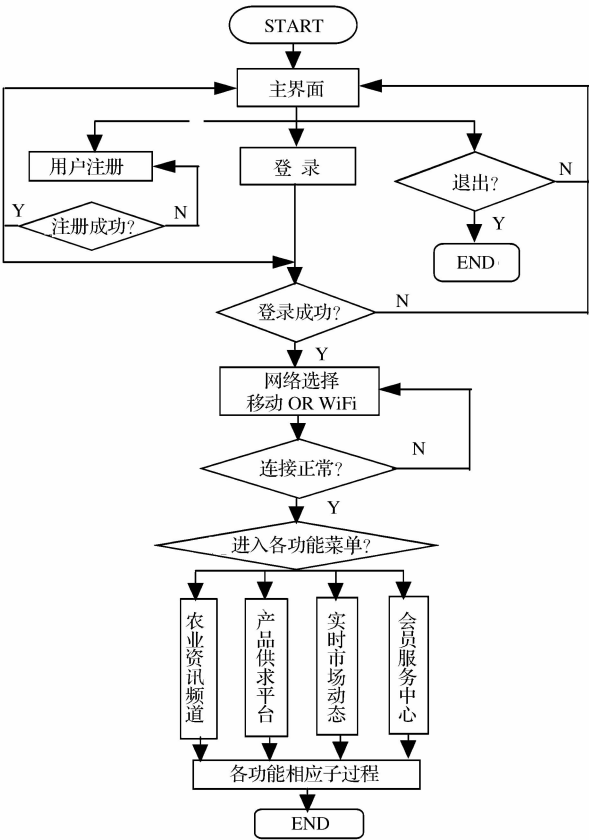


图 5 Andiord 终端软件流程图

3)供求信息管理:包括查看、搜索、发布、编辑、删除和激活供求信息的操作。

4)实时市场动态:包括市场预测和市场直击。

4.3 终端软件流程

用户首次进入该应用程序需要进行注册,然后登陆进入到控制界面。系统总体流程如图 5 所示。

5 结论

基于智能手机的农业信息服务系统结合 Web Services 数据挖掘与 Android 智能手机技术,利用 Heritrix 框架进行信息的收集,使移动通信网络和计算机网络更好的服务于农业科技服务信息的推送与农产品的网络营销过程。利用本系统,用户可以随时随地通过手机学习和了解最新的农业科技信息,也可对农产品的需求、价格等信息进行发布与收集,还可实现物流配送的往来信息交流。利用网络为媒介,给不同需求的用户提供一个在线交流平台,最终解决农业信息推送和农产品流通难题,提高农民在市场竞争中的竞争力,增加农民收入。

参考文献

[1] 许莉丽. 基于安卓的三农移动学习平台研究[J]. 软件导刊, 2012,11(11):87-88.

[2] 袁小洁. 基于 Heritrix 的商品信息搜索的网络爬虫系统的设计[J]. 电脑编程技巧与维护, 2012(22):74-75.

[3] 朱敏,罗省贤. 基于 Heritrix 的面向特定主题的聚焦爬虫研究[J]. 计算机技术与发展, 2012,22(2):64-68.

[4] 刘丽娟,刘仲鹏. 数据挖掘技术在农业数据挖掘平台中的应用研究[J]. 福建电脑, 2012(2):170-171.

[5] 刘天全,刘世洪,崔运鹏. 基于 B/S 结构的农业数据挖掘系统的研究[J]. 农机化研究, 2011(11):178-182.

[6] 徐姗,谢丹丹,姜代红. 基于 Android 系统智能手机的网页地图服务设计[J]. 软件导刊, 2012,11(11):108-109.

[7] 靳岩,姚尚明. Google Android 开发入门与实战[M]. 北京:人民邮电出版社,2009.

2)会员中心:会员注册、会员登录和会员管理。

Reasearch of Agricultural Information Service System Based on Smartphone

SUN Xiao-hua¹, WANG Fu-shun², YANG Hui-ying², ZHAO Yan², WANG Feng-guo³

(1. Digital Transmit Department of Hebei Software Institute, Baoding Hebei 071000, China;
2. College of Information Science & Technology Hebei Agricultural University, Baoding Hebei 071001, China;
3. Power Supply Company of Julu in Hebei Province, Julu Hebei 055250, China)

Abstract: By studying the characteristics of farmers for agricultural information in China, build technology structure for the agricultural information service system on the smartphone platform; made combination of data mining technology based on Web Services and Android smartphone technology. Heritrix technology was used to collect the specific agricultural information, through the Web Services data mining, system provide the most direct agricultural information service to the Android platform client, and made mobile communication network and computer network to better serve information pushing of agricultural science & technology and agricultural products online sales process.

Key words: smartphone; data mining; information pushing of agricultural science and technology; agricultural products online sales