基于 Web 数据的农业网络信息自动采集与分类系统

段青玲[1]，魏芳芳[1]，张 磊[1,2]，肖晓琰[1]

（1. 中国农业大学信息与电气工程学院，北京 100083； 2. 北京市农业物联网工程技术研究中心，北京 100083）

摘 要：为了快速、高效地获取农业 Web 信息，解决信息孤岛和信息不对称的问题，重点研究了农业 Web 数据自动采集与抽取、基于 SVM（support vector machine）的文本分类、物联网异构数据采集等技术，并采用统一建模语言（unified modeling language，UML）描述了农业网络信息自动采集与分类系统。该系统实现了农业网站、物联网数据的自动抓取和共享，为用户提供农业资讯、农产品市场行情、供求信息在线查询，环境数据实时监测和个性化信息服务等功能。应用结果表明，该系统对样本集网站的信息抓取准确率为 98.2%，资讯分类准确率为 92.5%，具有数据采集实时性强、用户参与度好、通用性高等特点，该系统为农业信息整合和服务提供参考。

关键词：农业；文本处理；采集系统；信息；物联网

doi：10.11975/j.issn.1002-6819.2016.12.025

中图分类号：TP274+.2 **文献标志码**：A **文章编号**：1002-6819(2016)-12-0172-07

段青玲，魏芳芳，张 磊，肖晓琰. 基于 Web 数据的农业网络信息自动采集与分类系统[J]. 农业工程学报，2016，32(12)：172－178. doi：10.11975/j.issn.1002-6819.2016.12.025 http://www.tcsae.org

Duan Qingling, Wei Fangfang, Zhang Lei, Xiao Xiaoyan. Automatic acquisition and classification system for agricultural network information based on Web data[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(12): 172－178. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2016.12.025 http://www.tcsae.org

0 引 言

中国是传统农业大国，农村信息资源分散，农业产业门类多，个性化差异大，“信息孤岛”和“信息不对称”问题成为中国农业现代化的主要瓶颈[1-2]。因此，构建农业网络信息自动采集与分类系统，将分散于各个网站的农业信息资源进行整合，为用户提供统一的共享平台和个性化信息服务很有必要。

农业网络信息采集技术按照采集方式分为农业网站信息采集和物联网异构数据采集。农业网站信息采集涉及信息抓取[3-4]、抽取[5-9]、分类[10-12]等技术。Yogesh 等[13]研究了针对多种语言的新闻网页的信息抽取方法；刘玉龙等[14]研究了基于文本特征的自动抽取方法，抽取准确率为 91%；SeydaErtekin 等[15]采用 SVM 算法实现文本分类，分类准确率最高为 89.786%。上述研究大多集中于信息采集、抽取、分类的某一种技术，而非专用于农业 Web 数据。

农业物联网就是将物联网技术应用在农业生产、经营、管理和服务中[16-24]，即运用各类传感器，采集大田种植、设施园艺、畜禽、水产养殖和农产品物流等农业相关信息，并将获取的海量农业信息进行融合、处理，实现农业产前、产中、产后的过程监控和科学管理[25]。在物联网信息采集中，由于传感器和无线传输网络等设备厂商众多，存在着感知数据格式多样、量纲不一致、数据组织形式不同[26]等问题，因此，如何把感知数据转换为格式统一、高质量的数据，是实现物联网异构环境数据融合的难点。

针对农业网络资源分散、异构问题，本文研究了农业 Web 数据自动采集与抽取、文本分类、物联网异构数据整合等技术，克服了数据区域识别困难、文本分类准确率不高、感知数据到统一格式的目标数据转换质量低的技术难点，设计了农业网络信息自动采集与分类系统，实现了农业资讯、农产品市场行情、供求信息查询，资讯信息自动分类，环境数据实时监测等功能。该系统已投入运行，系统的特点在于，将获取的互联网数据和物联网数据整合在一个平台上，既为用户提供了互联网资讯、价格、供求等产前信息服务，又提供了环境数据便于产中实时监测和产后生产决策服务。应用结果表明，系统采集数据实时性强，信息分类准确率高，能够将分散、异构的数据实时整合，并为用户提供综合全面、个性化的农业信息服务。

1 系统需求分析

1.1 研究对象

农业网站信息包括科技、市场、资讯和其他信息[27]。科技类信息涉及新品种研发、新技术推广、科研成果等内容，主要由科研单位在网上发布，如中国农业科技信息网；市场信息是指农产品的市场行情、价格和供求信息，由各地的农产品批发市场发布，如北京新发地市场网站；资讯类信息主要是新闻资讯，由新闻网站发布，如中国农

收稿日期：2015-12-11 修订日期：2016-04-24

基金项目：国家高技术研究发展计划（863 计划）资助项目（2013AA102306）；山东省自主创新资助项目（2014XGA13054）；中央高校基本科研业务费专项资金资助项目（2015XD001）。

作者简介：段青玲，女，河南，教授，工学博士，主要从事智能信息处理方面研究。北京 北京中国农业大学信息与电气工程学院，100083。

Email：dqling@cau.edu.cn

业新闻网；其他信息包括政府、企业、行业网站，发布惠民政策、农业商务信息，如农业部主办的中国农业信息网。本文主要研究与农业农村相关的惠农政策、资讯、新品种、推广技术、价格、供求等信息的采集与整合。

农业物联网数据是由不同传感器采集，例如，畜禽养殖物联网环境数据包括光照、粉尘、湿度、二氧化硫、硫化氢等；水产养殖物联网水质数据包括温度、pH 值、溶解氧等，这些数据是实现农业生产环境控制和智能化管理的基础。本文主要研究不同传感器网络的数据采集与整合。

1.2　系统需求

1.2.1　功能需求

农业网络信息自动采集与分类系统包括 4 个子系统。1）系统参数配置子系统：一是设置系统用户、数据源网站、数据采集规则等信息；二是设置物联网的数据源、映射规则和采集频率等信息。2）互联网信息采集子系统：根据配置的采集规则抓取资讯、市场行情和供求等信息，并进行分类和存储处理。3）物联网信息采集子系统：根据映射规则采集实时环境感知数据。4）信息服务子系统：实时发布资讯、市场行情和供求等信息，根据用户特征进行个性化信息推荐，并将采集到的物联网数据进行统计汇总。

1.2.2　用户需求

系统描述采用 UML 方式能够详细展示系统需求、结构和业务逻辑[28-29]，农业网络信息自动采集与分类系统的用户分为管理员、企业用户、普通用户 3 类，系统用例图如图 1 所示。管理员对用户进行管理，设置数据源网站信息，制定抓取规则，配置物联网采集点和物联网数据源，制定物联网数据转换规则。企业用户可以通过系统获取农产品资讯等信息，实时监测本企业的生产环境。普通用户主要进行信息浏览和查询统计。

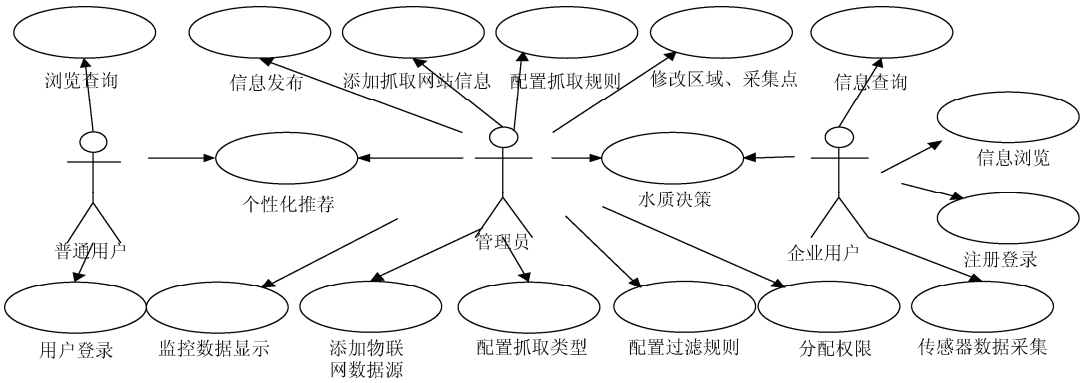

图 1　系统用例图

Fig.1　Use case diagram of system

2　系统设计

2.1　系统总体设计

系统总体由数据源配置、信息采集和信息服务 3 部分组成，如图 2 所示。数据源配置主要完成互联网网站和物联网数据源的管理，互联网信息采集模块抓取网站信息，进行信息抽取，将抽取到的信息进行分类处理，在信息服务模块将采集信息发布，根据农户行业特征进行信息推荐。物联网信息采集模块抓取 XML、文本、Excel 格式的数据，并根据映射规则进行数据转换，数据过滤生成结构化的目标数据，以关系数据库存放，在信息服务模块进行生产监测和生产管理。用户可以对数据进行查询浏览。

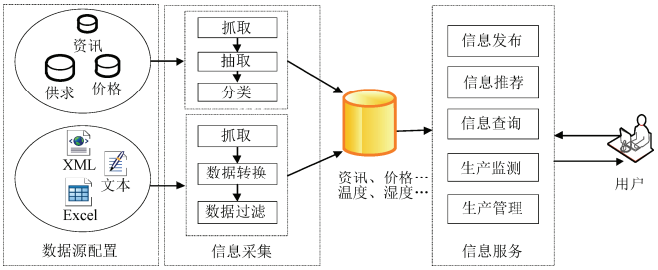

图 2　系统总体设计图

Fig.2　System overall design diagram

2.2　关键技术研究

系统研发的目的是为用户提供信息服务，其关键技术主要包括：信息采集、抽取、分类、物联网数据采集等技术。用户获取信息的途径有 2 种方式，一种是信息检索，用户通过查询获取信息，另一种是信息推荐，用户通过客户端接收信息。信息采集、抽取、物联网数据采集技术使得用户快速查询到有效信息，文本分类技术为个性化信息推荐提供了技术支持。

2.2.1　农业网站数据采集

农业网站数据采集主要采集互联网上的资讯、市场行情、供求信息，采集过程如下：首先选择数据源即种子站点，从中获取网站 URL、网站类别信息，保存原始网页，通过网络爬虫，不断扩展到<a>和<frame>标签里的超链接，下载网页。然后采用 HTML parser 工具解析网页，自定义 NodeFilter 对象提取用户感兴趣的内容，包括市场、价格、所在地、标题等，将其保存到数据库服务器中，见图 3。

信息抽取是为了从抓取到的网页中得到结构化的数据，抽取过程如下：用户输入数据源网站信息后，系统对爬取到的目标网页去噪，构建 DOM 树，挖掘出目标数据区域，分割属性，最后对得到的结构化数据进行抽取。如何根据目标数据区域识别方法，从多个具有重复模式

的数据区域里识别目标区域是信息抽取的难点，本文自定义的目标数据区域识别条件如下：1）数据记录条数多。网页包含多条记录，一般大于 3 条，如中国农业信息网。2）每条记录属性个数多。农业网站的价格和供求信息每条记录都会包含多个属性，如品种、批发市场、日期和报价等。3）既包含数字又包含汉字。农业网站价格类信息中都会包含价格，市场等。如果验证通过 3 个条件，为每个特征分配权重，对每个数据区域计算权值并排序，取最大值的即为目标数据区域。如图 4 所示为互联网数据抽取活动图。

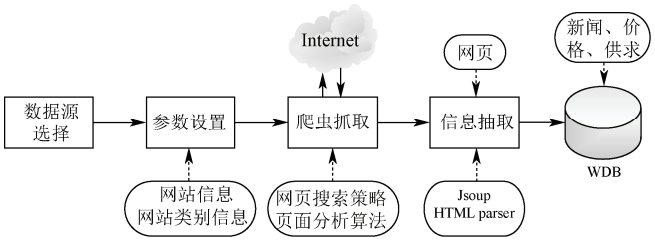

图 3 数据采集过程图

Fig.3 Data acquisition process chart

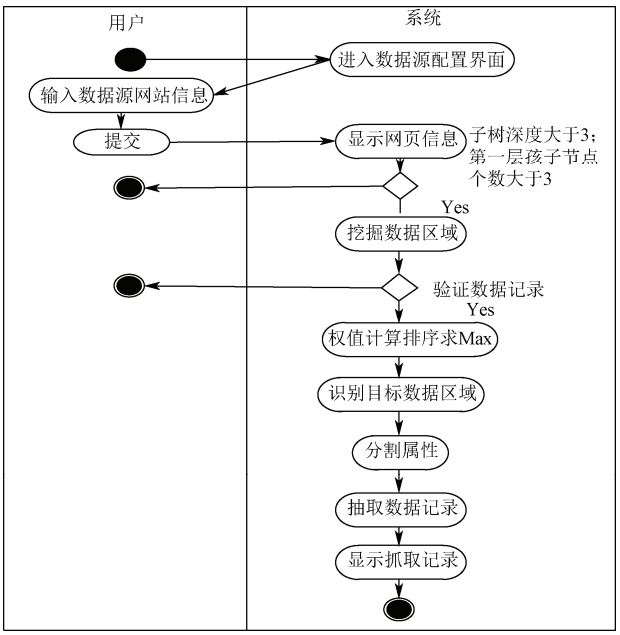

图 4 数据抽取活动图

Fig.4 Data extraction process chart

2.2.2 基于 SVM 的农业文本分类

为了提供个性化的信息服务，农业信息分类需要结合农户的行业特征，系统按照农产品类别实现信息自动分类。基于 SVM 的文本分类流程如图 5 所示，主要分为 2 个阶段：训练阶段和测试阶段。训练阶段主要包括训练样本的分词、特征选择、特征项权值计算等处理，最后采用 SVM 算法获得农业文本分类器。测试阶段测试样本经过同样的处理，再根据训练阶段得到的关键词库进行特征过滤，最后通过训练好的分类器进行分类。

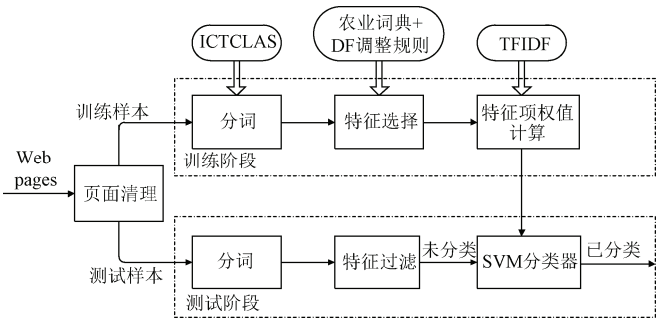

图 5 农业文本分类图

Fig.5 Agricultural text classification diagram

通过 ICTCLAS 分词处理后，本文根据农业生产资料编码构建了农业行业分类关键词库（表 1）。例如，词“牛肉”“鲫鱼”中含有“牛”“鱼”，根据文档频率调整规则调整该词的文档频率。文档频率调整规则如下：假设文档 D 是包含所有特征项的集合，$D(f_1,f_2,f_3,\cdots,f_n)$ 表示 n 维向量空间中的一个向量，其中 f_k（k=1,…,n）表示一个特征项，DF 计算公式如下

$$DF(f_k)=\frac{a(f_k)}{A}。$$

式中 $a(f_k)$是具有特征项 f_k 的文本数，A 是训练集文本数。

计算出文档中的特征词的 DF 值后，根据特征词是否包含农业行业分类关键词进行调整。假设：农产品分类关键词的一级词汇集合为：$T_1=\{u_i|i=1,2,\ldots,r\}$，$r$ 为一级词汇个数，农产品分类关键词的二级词汇集合为：$T_2=\{v_i|i=1,2,\ldots,p\}$，$p$ 为二级词汇个数，选择后的新特征集合为 $S=\{f_k|k=1,2,\ldots,m\}$，f_k 拆分后的特征子集为：$S_1=\{f_{ki}|k=1,2,\ldots,m;\ i=1,2,\ldots,q\}$，其中 f_{ki} 为 f_k 拆分后的特征词，m 为特征项的个数，q 为拆分后的特征子集包含的特征词个数。

表 1 农业行业分类关键词库

Table 1 Key words of agricultural industry classification

一级词汇 Vocabulary of first level	二级词汇 Vocabulary of second level
粮食 Foodstuff	谷、麦、米、高粱、糜子、豆、薯
蔬菜 Vegetables	菜、菊苣、葱、蒜、茴香、萝卜、山药、芋头、姜莴笋、豆、莲藕、茭白、慈菇、荸荠、韭黄、洋蓟、椒、茄
药材 Medicinal materials	参、香、藤、白、防、三、皮、草、叶、子、仁、蔻、苓、甲、蛇、蝎、胆、黄、川、甘、乌、蚕、母、板蓝根、苍术、柴胡、常山、麻、赤芍、地榆
水产 Aquatic products	海、鱼、虾、蟹、蛙、贝、藻、鲳、尖吻鲈、真鲷、螺、蚌、蛏、蛤蜊、海参
畜牧 Animal husbandry	牲畜、牛、奶、乳、蛋、马、猪、羊、驴、骡、骆驼、家禽、鸡、鸭、鹅、狗、猫、兔、兽、蜜蜂
林产品 Forest	木、育苗、造林、森林、木材、竹材、橡胶、松脂、生漆虫胶、槐、胶
果品 Fruit	果、梨、柿子、山楂、榅桲、葡萄、桃、栗、桔、莓、李、杏、梅、椰、瓜
花卉 Flowers and plants	花、菊、草、葵、掌、兰、冠、蕨、荷、莲、桂、莞、竹、梅、莺萝、香豌豆、牵牛、雪轮、福禄考、美女樱、千日红

假设特征项文档频率阈值为[α, β]，文档频率调整步骤如下：

初始化时，$S=\phi$

For　$i=1,2,...,n$

IF　$DF(f_k)<\alpha$　THEN $f_k \notin S$

IF $f_k \wedge u_i \in S_1$　THEN　$S=\{f_k,S\}, DF(f_k)=\beta$

IF $f_k \wedge v_i \in S_1$　THEN　$S=\{f_k,S\}, DF(f_k)=\beta-1$

利用上述文档频率调整方法，可对特征进行降维，选择出相关度较高的特征词，然后进行 TFIDF 权重计算，构建线性 SVM 分类器模型，实现农业信息自动分类。

2.2.3　基于物联网的异构数据采集

农业物联网已广泛应用于农业生产各个环节中，但是传感器采集的数据存在三大问题[25-26]：1）数据量大：多种传感器不断更新采集的数据，产生海量数据。2）数据类型不一致：提交的数据有些是实际物理值，如温度实际值，而有些是电压电流值，需要经过公式转换。3）数据组织形式不统一：传感器网络的数据有文本、excel、xml 文件等不同组织形式。为了不同物联网采集的数据能够共享，根据农业物联网的特点，本文定义了采集点的编码规则，并规定统一的感知数据量纲。物联网设备编码规则如图 6 所示，采集设备编码采用组合编码方式设计，长度共 18 位，分为四部分，从左到右的含义是区域码、网络类型、网络编号和采集点编码。其中，网络类型长度 2 位，取值范围为 01-99，不同的取值代表不同类型的物联网，例如 01 表示水产物联网、02 表示畜禽物联网。

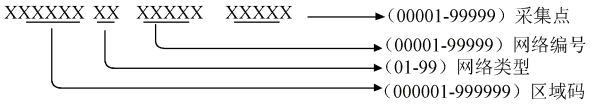

图 6　物联网设备编码结构图

Fig.6　Internet of things equipment code structure diagram

基于物联网的异构数据采集处理过程如图 7 所示：传感器采集的感知数据在采集节点通过无线传感网络传输到指定服务器作为源数据，不同厂商的传感器数据可能表示形式不同，有的是模拟信号，有的是数字信号，这时需要将原始粗糙数据进行统一的数据量纲转换，过滤噪声数据。然后针对具体的服务对象，将远程抓取的数据按照映射规则映射到目标关系数据库中相应字段，实现原始数据到目标数据的逻辑抽象。最后按照服务器指定的抽取频率采集数据。物联网数据整合主要包含以下 4 个步骤。

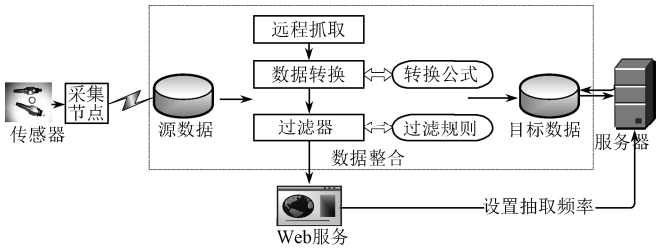

图 7　数据整合过程

Fig.7　Data integration process chart

1）数据源配置：配置源数据库，各个通道的传感器类型，如温度、湿度、pH 值和氨气等。

2）远程抓取：抓取传感器数据到采集节点，支持单通道多次采集和多通道单次采集 2 种方式。单通道多次采集是采集频率快的传感器在指定时间内采集多次，多通道单次采集指的是采集频率慢的多个传感器在指定时间内采集一次。

3）数据转换：建立源数据与目标数据之间的映射如图 8 所示，映射 1 源数据与目标数据字段直接对应，映射 2 根据临时表中相同的字段名，多条记录映射在目标表的同一字段下。每条映射对应一条 textarea 类型的数据格式转换公式，用户以 *y=f(x)*的格式输入，把测量值非浮点型数据转换为浮点型，并将数据量纲转换为统一量纲。映射 3 是把 XML 类型的感知数据转换为关系数据库。

4）数据过滤：根据过滤规则对测量值超出传感器量程范围的异常值进行处理。

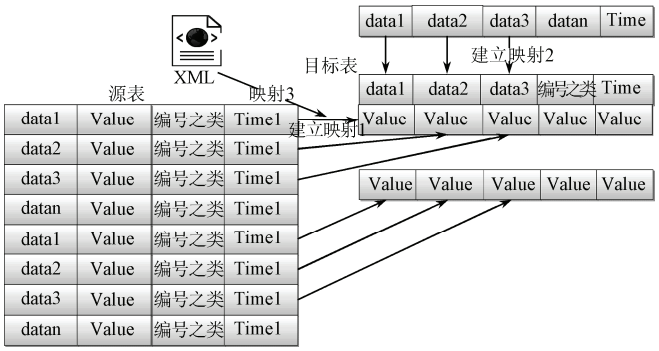

图 8　数据映射过程

Fig.8　Data mapping process chart

3　核心功能实现与测试

农业网络信息自动采集与分类系统开发环境为 JDK1.6+MyEclipse2014，服务器采用免费开源的 Tomocat6.0，数据库为 MySQL5.5。测试环境为 windows server2008+JDK1.6+ Tomcat 6.0。核心功能实现主要从农业网站信息采集、基于 SVM 的农业文本分类、物联网异构环境数据采集等 3 个模块阐述。

3.1　农业网站信息采集

本文采集了中国水产网、价格行情网、农产品信息网、中国农业网等 10 个网站的数据如表 2 所示，其中，A 表示单个网页数据记录数，C 表示实际抽取到的数据记录数，系统测试了 10 个网站中 219 条数据记录，抓取准确率为 98.2%。

3.2　基于 SVM 的农业信息分类

本文试验过程中通过爬虫程序获取到中国农业新闻网、蔬菜商情网、中药材信息网、河南水产网、农林网、水果信息网、园林资料库等网站的 3 046 个网页后，通过 HTML Parser 进行内容解析提取标题和正文，并使用 Java 的正则表达式过滤 script、样式等无用信息。最后将纯文本作为训练语料保存到本地。训练语料分词后得到 2 955 篇资讯，如表 3 所示为文本分类的测试统计数据，分为

粮食 c1、蔬菜 c2、药材 c3、水产品 c4、畜禽 c5、林产品 c6、果品 c7、花卉 c8 共 8 个类别。调用 Weka 进行分类结果如表 4 所示，*P* 表示分类准确率，*R* 表示分类召回率，*F* 表示分类准确率和召回率的调和平均值。结果表明：相对于其他算法，SVM 更适合用于农业文本分类；线性 SVM 分类效果最理想，分类准确率为 91.8%、召回率为 90.3%，*F* 值为 90.7%；加入规则后 SVM 分类效果比未加规则的 SVM 分类效果好，线性 SVM 分类准确率为 92.5%，召回率为 91.3%，*F* 值为 91.6%。

表 2 数据采集结果

Table 2 Data acquisition result

网站地址 (http://) Website address	网页记录数 Webpage records number *A*	实际记录数 Actual records number *C*
www.agrosc.com/Price/IFrame	25	24
nc.mofcom.gov.cn/channel/gxdj/jghq/jg_list.shtml	14	14
www.nongnet.com/list_40_0_0_0_0_0_0_2_1.aspx	30	29
cy.agronet.com.cn/Trade/List?area=430000	10	10
www.bbwfish.com/news_hot.asp?RecSortid=107986	20	20
www.zgchawang.com//news/list/34/	25	24
www.farmers.org.cn/zhuanti/ShowClass.asp?ClassID=403	20	20
www.zysnw.cn/a/yangzhijishu/	30	29
yxsc.6636.net//more.asp?typeid=1	20	20
www.tjagri.ac.cn/index.php?m=content&c=index&a=lists&catid=8	25	25
总计 Total	219	215
准确度 Precision/%		98.2

表 3 农业文本分类数据集

Table 3 Agricultural text classification data set

类别 Category	训练集 Training set	测试集 Test set	总数 Total
粮食 Foodstuff c1	86	50	136
蔬菜 Vegetables c2	298	198	496
药材 Medicinal materials c3	198	186	384
水产 Aquatic products c4	294	232	526
畜禽 Animal husbandry c5	195	148	343
林产品 Forest c6	49	32	81
果品 Fruit c7	264	206	470
花卉 Flowers and plants c8	297	222	519

表 4 农业文本分类试验结果

Table 4 Agricultural text classification test results

算法 Algorithm		准确率 Precision *P*/%	召回率 Recall *R*/%	*F* 值 *F*-Measure/%
未加入规则 No rules	神经网络	68.2	68.1	67.6
	SMO	85.7	82.3	83.1
	贝叶斯	85.9	85.5	85.1
	C4.5 决策树	88.3	87.8	87.9
	RBF-SVM	88.2	86.1	84.7
	线性 SVM	91.8	90.3	90.7
加入规则后 With rules	RBF-SVM	88.6	86.4	85.5
	线性 SVM	92.5	91.3	91.6

3.3 基于物联网的信息采集

本文以畜禽养殖物联网的环境参数采集为例进行测试。畜禽养殖物联网使用 SmeshLink 传感器网络采集温度、湿度、粉尘和光照数据，数据以 MySQL 数据库存放；使用科尔诺传感器网络采集氨气、二氧化硫和硫化氢等气体数据，数据以 Access 数据库存放。通过建立源数据与目标数据的映射，并规定异常值处理规范和量程转换规则，将 Access 数据和 MySQL 数据感知数据整合，畜禽养殖物联网感知数据整合结果如图 9 所示，包括光照、温度、湿度、氨气、粉尘等，以 MySQL 数据库存放。该方法通过映射可以将不同的传感器网络数据进行自动整合，不需要开发不同的接口程序。

图 9 畜禽养殖物联网数据整合结果

Fig.9 Data integration result of livestock and poultry breeding IOT

4 结 论

本文构建了农业网络信息采集与分类系统，实现了互联网信息和异构环境数据采集，并建立了专用于农业信息的文本分类模型，主要结论如下：

1）在农业网站数据获取方面，系统实现了农业新闻、惠农政策、新技术、农产品价格、供求信息的抽取，改善了抽取准确率，对样本集网站测试抓取准确率为 98.2%。

2）基于线性 SVM 模型实现了农业资讯的文本分类，利用本文构建的农业特征词库和文档频率调整规则进行特征选择提高了 SVM 的分类准确率和召回率，分类准确率为 92.5%，召回率为 91.3%。

3）系统在物联网异构数据采集方面，通过建立源数据、目标数据之间的映射解决了感知数据异构问题，实现了农业物联网环境数据的采集。

4）系统主要用于采集农业网络信息，但系统开发实现采用的关键技术信息采集、抽取、分类以及物联网感知数据整合技术亦可应用于其他行业领域。

本文针对农业网站的研究旨在整合半结构化的市场行情数据、供求数据和非结构化的农业资讯信息，这些数据是涉农用户最关心的信息。通过上述信息采集、抽取、分类技术可以获取到结构化的、分类效率高的农业网站数据。系统针对物联网数据整合的目的是屏蔽感知数据异构性，通过本文源数据与目标数据之间的映射转换方法可以获取统一格式的目标数据，并已经实现畜禽、水产养殖物联网数据的整合。此外，系统实行注册机制，注册过的企业认为已经通过数据授权，可以通过该系统获取实时环境数据。

[参 考 文 献]

[1] 曹丽英，张晓贤，赵月玲，等. 云计算在农业信息资源整合模式中的应用[J]. 中国农机化，2012，241(3)：141－144.
Cao Liying, Zhang Xiaoxian, Zhao Yueling, et al. Application of cloud computing in agricultural information resources

integration mode[J]. Chinese Agricultural Mechanization, 2012, 241(3): 141－144. (in Chinese with English abstract)

[2] 刘慧悦，李后卿，肖雨滋. 新农村农业产业链信息不对称问题研究[J]. 农业经济与科技，2013，24(6)：39－41.

[3] 闫兴龙，刘奕群，方奇，等. 基于网络资源与用户行为信息的领域术语提取[J]. 软件学报，2013，24(9)：2089－2100.
Yan Xinglong, Liu Yiqun, Fang Qi, et al. Domain-specific terms extraction based on web resource and user behavior[J]. Journal of Software, 2013, 24(9): 2089－2100. (in Chinese with English abstract)

[4] Duan Qingling, Yang Rengang, Chen Ying. Automatic identifying query interfaces of deep web based on PreClassification-SVM[J]. Sensor Letters, 2013, 11: 1－7.

[5] Wang Tiantian, Duan Qingling, Lian Jinghua. Extraction data records in agricultural web pages[J]. Sensor Letters, 2014, 12: 795－801.

[6] Wang Tiantian, Li Guo, Duan Qingling, et al. Deep web integrated query interface construction method based on apriori algorithm[J]. Journal of Information & Computation Science, 2013, 10(15): 5063－5075.

[7] Emilio Ferrara, Pasquale De Meo. Web Data Extraction, Applications and Techniques: A Survey[D]. Knowledge-Based Systems, 2014: 301－323.

[8] SwaranLata, Bhaskar Sinha, Ela Kumar, et al. Semantic web query on e-governance data and designing ontology for agriculture domain[J]. International Journal of Web & Semantic Technology, 2013, 4(3): 201－209.

[9] Wu Gongqing, Wu Xindong. Extracting Web News Using Tag Path Patterns[C]// Proceedings of the first International Conferences on Web Intelligence and Intelligent Agent Technology, 2012, 1: 588－595.

[10] Keyur J Patel, Ketan J Sarvakar, Gujarat, et al. Web page classification using data mining[C]//International Journal of Advanced Research in Computer and Communication Engineering, 2013, 2(7): 2513－2519.

[11] Patrick Kenekayoro, Kevan Buckley, Mike Thelwall, et al. Automatic classification of academic web page types[J]. Scientometrics, 2014, 101(2): 1－12.

[12] 魏芳芳，段青玲，肖晓琰，等. 基于支持向量机的中文农业文本分类技术研究[J]. 农业机械学报，2015(增刊 1)：174－179.
Wei Fangfang, Duan Qingling, Xiao Xiaoyan, et al. Classification technique of Chinese agricultural text information based on SVM[J].Transactions of the Chinese Society of Agricultural Machinery, 2015(Supp.1): 174－179. (in Chinese with English abstract)

[13] Yogesh W Wanjari, Vivek D Mohod, Dipali B, et al. Automatic news extraction system for Indian online newspapers[C]// Proceedings of the 3rdIEEE International Conference on Reliability, INFOCOM Technologies and Optimization, 2014.

[14] 刘玉龙. Web 信息抽取规则的设计和实现[D]. 南京：南京大学，2013.
Liu Yulong. Design and Implementation of Web Information Extraction Rules[D]. Nanjing: Nanjing University, 2013. (in Chinese with English abstract)

[15] Seyda Ertekin, C Lee Giles. A Comparative Study on Representation of WebPages in Automatic Text Categorization[D]. http://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.154.9946&rep=rep1&type=pdf.

[16] Liu Ping . Agricultural drought data acquisition and transmission system based on internet of things[C]// Proceedings of the Fifth IEEE International Conference on Intelligent Systems Design and Engineering Applications, 2014: 4004－4008.

[17] 孟志军，王秀，赵春江，等. 基于嵌入式组件技术的精准农业农田信息采集系统的设计与实现[J]. 农业工程学报，2005，21(4)：91－96.
Meng Zhijun, Wang Xiu, Zhao Chunjiang, et al. Development of field information collection system based on embedded COM-GIS and pocket PC for precision agriculture[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2005, 21(4): 91－96. (in Chinese with English abstract)

[18] 赵庆展，靳光才，周文杰，等. 基于移动 GIS 的棉田病虫害信息采集系统[J]. 农业工程学报，2015，31(4)：183－190.
Zhao Qingzhan, Jin Guangcai, Zhou Wenjie, et al. Information collection system for diseases and pests in cotton field based on mobile GIS[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2015, 31(4): 183－190. (in Chinese with English abstract)

[19] 尚明华，秦磊磊，王风云，等. 基于 Android 智能手机的小麦生产风险信息采集系统[J]. 农业工程学报，2011，27(5)：178－182.
Shang Minghua, Qin Leilei, Wang Fengyun, et al. Information collection system of wheat production risk based on android smartphone[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2011, 27(5): 178－182. (in Chinese with English abstract)

[20] 孙忠富，曹洪太，李洪亮，等. 基于 GPRS 和 WEB 的温室环境信息采集系统的实现[J]. 农业工程学报，2006，22(6)：131－134.
Sun Zhongfu, Cao Hongtai, Li Hongliang, et al. GPRS and web based data acquisition for greenhouse environment[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2006, 22(6): 131－134. (in Chinese with English abstract)

[21] 熊本海，罗青尧，杨亮. 家畜精细饲养物联网关键技术研究[J]. 中国农业科技导报，2011，13(5)：19－25.
Xiong Benhai, Luo Qingyao, Yang Liang. Studies on key thing internet technology for precise livestock feeding[J]. Journal of Agricultural Science and Technology, 2011, 13(5): 19－25. (in Chinese with English abstract)

[22] 张伟，何勇，刘飞，等. 基于物联网的规模化畜禽养殖环境监控系统[J]. 农机化研究，2015，2：245－248.
Zhang Wei, He Yong, Liu Fei, et al. The environmental control system based on IOT for scale livestock and poultry breeding[J]. Journal of Agricultural Mechanization Research, 2015, 2: 245－248. (in Chinese with English abstract)

[23] 刘双印，徐龙琴，李道亮，等. 基于物联网的南美白对虾疾病远程智能诊断系统[J]. 中国农业大学学报，2014，19(2)：189－195.
Liu Shuangyin, Xu Longqin, Li Daoliang, et al. Research on remote system for disease diagnosis of penaeus vannamei based on internet of things[J]. Journal of China Agricultural University, 2014, 19(2): 189－195. (in Chinese with English abstract)

[24] 王贵荣，李道亮，吕钊钦，等. 鱼病诊断短信平台设计与实现[J]. 农业工程学报，2009，25(3)：130－134.
Wang Guirong, Li Daoliang, Lü Zhaoqin, et al. Design and implementation of SMS-platform system for diagnosis of fish diseases[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2009, 25(3): 130－134. (in Chinese with English abstract).

[25] 李道亮. 物联网与智慧农业[J]. 农业工程，2012，2(1)：1－7.
Li Daoliang. Internet of things and wisdom agriculture[J]. 2012, 2(1): 1－7. (in Chinese with English abstract).

[26] Jeonghwan H, Changsun S, Hyun Y. Study on an agricultural environment monitoring server system using wireless sensor network[J]. Sensors, 2010, 10(12): 11189－11211.

[27] 雍春玲，杨晓容，文竹，等. 网络农业新闻信息的采集与发布方式初探[J]. 三农论坛，2014，31(3)：14－16.

[28] 王馨樱，蒋夏军. 基于 UML 活动图及混合遗传算法的测试场景生成[J]. 计算机工程与应用，2015，9(7)：1－7.
Wang Xinying, Jiang Xiajun. Generating test scenarios from UML activity diagram using hybrid genetic algorithm[J]. Computer Engineering and Applications, 2015, 9(7): 1－7. (in Chinese with English abstract)

[29] 柳毅，麻志毅，何啸，等. 一种从 UML 模型到可靠性分析模型的转换方法[J]. 软件学报，2010，21(2)：287－304.
Liu Yi, Ma Zhiyi, He Xiao, et al. Approach to transforming UML model to reliability analysis model[J]. Journal of Software, 2010, 21(2): 287－304. (in Chinese with English abstract)

Automatic acquisition and classification system for agricultural network information based on Web data

Duan Qingling[1], Wei Fangfang[1], Zhang Lei[1,2], Xiao Xiaoyan[1]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*;
2. *Beijing Agricultural Networking Engineering Technology Research Center*, *Beijing* 100083, *China*)

Abstract: The purpose of this study is to obtain agricultural web information efficiently, and to provide users with personalized service through the integration of agricultural resources scattered in different sites and the fusion of heterogeneous environmental data. The research in this paper has improved some key information technologies, which are agricultural web data acquisition and extraction technologies, text classification based on support vector machine (SVM) and heterogeneous data collection based on the Internet of things (IOT). We first add quality target seed site into the system, and get website URL (uniform resource locator) and category information. The web crawler program can save original pages. The de-noised web page can be obtained through HTML parser and regular expressions, which create custom Node Filter objects. Therefore, the system builds a document object model (DOM) tree before digging out data area. According to filtering rules, the target data area can be identified from a plurality of data regions with repeated patterns. Next, the structured data can be extracted after property segmentation. Secondly, we construct linear SVM classification model, and realize agricultural text classification automatically. The procedures of our model include 4 steps. First of all, we use segment tool ICTCLAS to carry out the word segment and part-of-speech (POS) tagging, followed by combining agricultural key dictionary and document frequency adjustment rule to choose feature words, and building a feature vector and calculating inverse document frequency (IDF) weight value for feature words; lastly we design adaptive classifier of SVM algorithm. Finally, the perception data of different format collected by the sensor are transmitted to the designated server as the source data through the wireless sensor network. Relational database in accordance with specified acquisition frequency can be achieved through data conversion and data filtering. The key step of data conversion can be implemented on the basis of mapping rules between source data and target data. The mapping rules include 3 kinds of rules. The first is the source data directly corresponding to the target data; the second is that we create a temporary table, which corresponds to target table if they have same field name; and the third is converting perception data of XML (extensible markup language) type to relational database. Besides, data filtering is required to process abnormal values of the measured value beyond the sensor range. In this paper, unified modeling language (UML) is used to describe the agricultural network information automatic acquisition and classification system. User requirement analysis is described by the system's use case diagram. Web data extraction process is described by the system activity diagram. These help the system's key function implement of automatic information acquisition from Internet. The IOT data sharing module is implemented based on the proposed data conversion and filtering rules. The system can supply the services of on-time agricultural news, agricultural product prices, supply and demand information browsing query, real-time agricultural environment monitoring and personalized information statistics. The preliminary application shows that the agricultural network information automatic acquisition and classification system improves the accuracy of information extraction and text classification. The information acquisition accuracy rate for sample web sets is 98.2%, and the accuracy rate of text classification with rules is 92.5%. Compared with sequential minimal optimization (SMO), Bayesian, C4.5 decision tree and radial basis function (RBF) based SVM algorithm, linear SVM is more suitable for agricultural news classification. The system has high real-time performance and good user participation for IOT applications, which will expect to be applied to agricultural information integration and intelligent processing.

Keywords: agriculture; text processing; information systems; information; the Internet of things