doi:10.11920/xnmdzk.2018.01.011

基于文本分类的农业种植信息集成推荐方法研究

彭 争,唐东明

(西南民族大学计算机科学与技术学院,四川 成都 610041)

摘 要:目前网络上存在着海量的农业信息,但是对于广大农民来说信息得不到有效的利用,迫切需要对信息进行集成推荐.针对网络上的农业种植方面的文本信息进行了深入研究,该系统首先利用爬虫技术自动地爬取海量农业种植信息,经清洗整理后构建数据集语料库.其次利用机器学习中 KNN 方法找到每个样本的 k 近邻对文章进行聚类,通过 TF-IDF 方法提取出关键词并构造词频矩阵,然后从文本中构建特征向量,进而对相似文档进行分类,最后将加权值经排序后的结果推荐给用户.该系统实现了对农业文本进行准确的自动分类以及自动提取出文章摘要,并对相似文章进行推荐展示的效果.

关键词:机器学习;文本分析;关联规则;个性推荐

中图分类号:TP391.1 **文献标志码**:A **文章编号**:2095-4271(2018)01-0069-06

Research on the method of agricultural planting information integration recommendation based on text classification

PENG Zheng, TANG Dong-ming

(School of Computer Science and Technology, Southwest Minzu University, Chengdu 610041, P. R. C.)

Abstract: At present, there is a huge amount of agricultural information on the network, but for the majority of farmers, the information cannot be used effectively, so it is urgent to recommend the integrated information. The text information of agricultural planting on the network is deeply studied. Firstly, the system uses crawler technology to automatically crawl the massive agricultural planting information, and after cleaning and organizing, it constructs the corpus of data set. Secondly, the KNN method is used to find the k – nearest neighbor of each sample to cluster the paper, extract the keywords and construct the word frequency matrix by TF-DF method, and then construct the feature vector from the text. Then the similar documents are classified, and the results of the weighted values are recommended to the users. The system realizes accurate automatic classification of agricultural text and automatic extraction of article abstracts and the recommendation and display of similar articles.

Key words: machine learning; text analysis; association rule; personality recommendation

中国作为农业大国,每年都积累包括作物的苗情、土情、水情、虫情、气象和灾害等,面对如此海量的数据,目前迫切需要研究解决的问题是如何充分利用数据,从而为农民提供指导性和实用性的信息.目前互联网上存在海量的线上资源,线上的农业电子资源对农民具有重要的意义.然而目前广大农民却不知道怎么查找相应的数据来解决实际生产生活中的问题.经调查,他们大多只是在百度上进行简单检索,一方面问题描述不够清晰,另一方面检索到的结果充斥着大量广告,最重要的是检索不到真正有用的信息.而

收稿日期:2017-09-04

通信作者:唐东明(1979-),男,汉族,湖北钟祥人,副研究员,博士,研究方向:模式分析、生物信息处理、移动应用. E-mail:tdm_2010@qq.com

基金项目:国家自然科学基金(61100118);西南民族大学创新型科研项目(CX2017SP272);西南民族大学专业学位研究生教育专项资助(2017YJZX005)

本研究致力于打造的服务于农民种植的应用,将网络上海量的农业信息进行汇总整理,构建知识库,使农民使用时更加精准,将相关种植信息关联在一起,具有很强的现实意义.

目前新闻文本分类是文本挖掘里面较为常见的场景,然而面对海量的信息内容常常采用人工标记新闻类别的方式,不仅消耗了大量的人力资源,同时也因为各种因素导致标注信息不准确导致的信息利用率不高[2].

本文主要通过 KNN 算法对 10 类待挖掘的文章进行处理,通过对主题权重的聚类等实现农业新闻文本的自动分类,通过分析记录用户的浏览历史记录挖掘出潜在的有价值的信息和知识进而达到个性化推荐的目的.本文设计并实现了基于文本分类算法的农业种植信息推荐系统.

1 系统设计

本系统构建的果农帮推荐系统将各类农业信息采集、存取、清洗、分析和可视化等进行深度集成,通过数据分析优化已有的种植方案,并将结果展示给农民.用机器学习方法进行分析研究,探索相对最优的农民种植策略问题,通过汇总的农业种植信息和灾害防治信息的集成展示可以对农业种植经行系统化指导,根据农民的兴趣来帮助农民获得更多有针对性的信息,最终设计并实现基于大数据的果蔬种植推荐系统.具体研究内容如下:

用户使用该系统过程中,如果是新用户,则需要进行注册操作,选择自己感兴趣的领域话题进行标注.如果是已注册的用户,则直接进行登录即可.当用户再次登录该系统以后,系统会根据用户的浏览记录和感兴趣的标签与数据库中处理后的关键词库进行相关匹配,运用关联规则进行推荐,从而满足用户个性化需求[2].

系统的整体架构如图 1 所示:

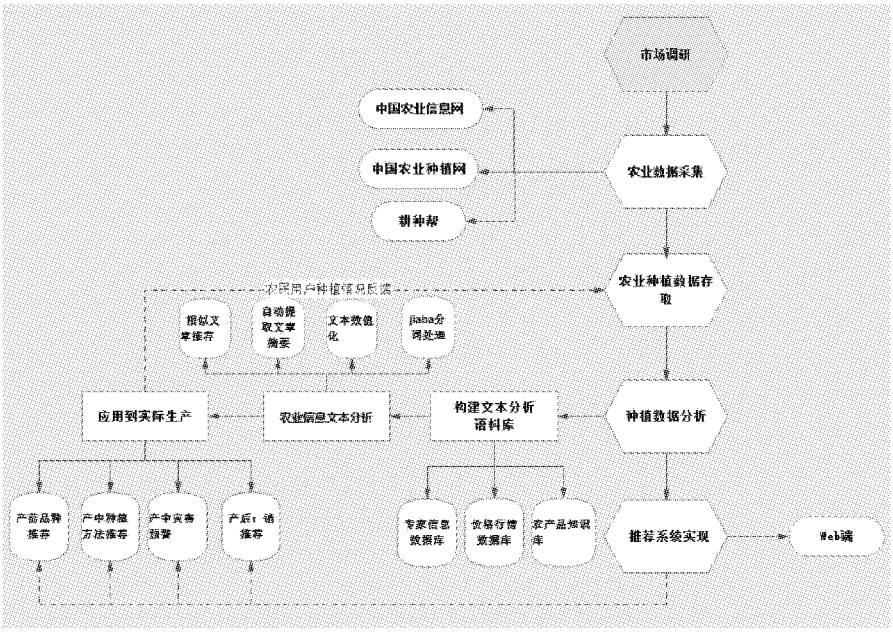


图 1 系统整体架构

Fig. 1 System flow diagram

1.1 数据库设计

该系统采用 MongoDB 进行数据库的设计与实现.MongoDB 是一个开源的,基于分布式的,面向文档存储的非关系型数据库.考虑到用户的喜好等属性会有多个记录因此将其字段设置为 array 类型方便存储,本系统主要数据库设计如表 1 和表 2 所示:

表 1 article 表

Table 1 Article Attribute Table

字段名	类型	是否可空	中文名
id	objectID	No	编号
class_name	string	Yes	类型名
content	string	Yes	内容
jiebacut_content	string	Yes	分词后内容
title	string	Yes	标题

表 2　user 表
Table 2　User Attribute Table

字段名	类型	是否可空	中文名
id	objectID	No	编号
user_name	string	Yes	登录名称
use_pwd	string	Yes	注册密码
user_love	array	Yes	用户喜好
looked_list	array	Yes	浏览记录

网络新闻文本具有数据量庞大,获取成本比较低,多样性丰富,用户自发进行发布以及信息及时性等特点[3-5].本文构建的推荐系统的服务的用户为以农业种植用户为主,兼具其他农业从业人员.考虑到农业数据种类丰富,农业信息涉及的范围非常广泛,因此数据集分类更加多样化.

对于农户来说,在种植环节,农民迫切希望了解市场的供需关系,提前对市场需求进行一定的预判以便决定种植农作物的品种和数量.在作物的生长环节,农民更关心天气信息以及灾害防治等,而作物快成熟以后,农民更希望了解市场价格趋势等问题[6].

农业类新闻文本数据具有以下特点,例如文本类形式使得结构化信息较少,一般只具有发布时间,标题,作者,内容等几个简单属性,使得进行分析时无法进行结构化检索等,只能通过自然语言处理的相关方法进行一系列的处理[6].其次,农业类文本的分类较多,涉及行业如种植、养殖、病虫害识别、市场趋势等并无统一分类规范[9].此外,农业新闻对准确性要求较高,农业新闻是农民获取信息的主要方式,若信息分类不准确会造成用户体验不好,影响推荐效果.

1.2　文本数据处理

机器学习具有很多分类方法可以应用在新闻文本的自动分类上,例如 KNN,SVM,朴素贝叶斯,决策树等,它们都有各自的优缺点.其中 KNN,SVM 等比较适合多分类场景.下本文利用爬虫技术在各大主流农业网站共收集了 10 种不同类别的农业新闻数据.判断一篇未知新闻属于哪个具体分类是一个监督分类问题,实验中有 10 类新闻数据集,每 100 篇属于一类,目标是构建一个有效的模式来判定未知新闻的类别.

本文主要进行农业类新闻的分析,因此利用现有的成熟的爬虫技术,在遵循 robosts. txt 协议的基础上,爬取主流农业类新闻网站各类原始农业新闻文本数据如下:将爬到的原始数据集存为文本文档,作为原始的数据集和语料库,如图 2 所示.

图 2　待分类源农业数据
Fig. 2　The agricultural data source

1.3　文字数值化

针对上一步收集到的原始信息,首先进行数据清洗和整理,处理过程如下:利用 python 中的 jieba 库进行分词和词频统计,利用 TF-IDF 方法进行词频统计,在处理过程中考虑到虚词标点符号等干扰项,因此进行停用词的过滤.接着利用 sklearn 库中的 KNN 进行文本信息挖掘,最后进行结果分析和评估[10].

首先利用 jieba 库进行中文分词处理.结巴(jieba)是集成在 python 中的一个工具包,可以对一段中文进行分词,代码清晰,扩展性好有三种分词模式,可

以适应不同需求. 其主要的处理思路如下:

①加载默认词典 dict. txt;

②从内存的词典中构建该句子的有向无环图;

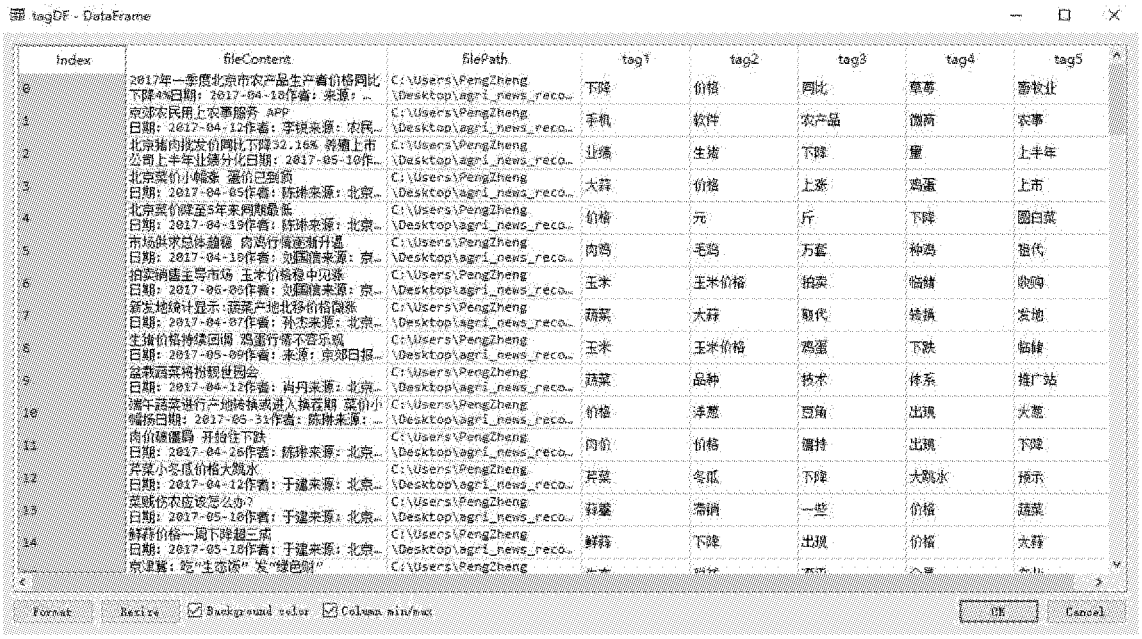
③对于词典中未收录词,使用 HMM 模型的 viterbi 算法尝试分词处理;

④已收录词和未收录词全部分词完毕后,使用 dp 寻找 DAG 的最大概率路径;

⑤输出分词结果.

接下来将文本中的词语转换为词频矩阵,并利用 TfidfTransformer( )方法统计每个词语的 tf-idf 权值. 再将文本转为词频矩阵,返回[(文章 idx,词语 id),词频],获取词袋模型中所有词语,遍历所有文本和获取某一文本下的词语权重[7].

经过测试实验发现,经 jieba 分词后的结果中占比较多的多为"的"等并没有实际含义的虚词,这些词汇是几乎在每篇文章中常见的停止词. 因此加载停用词词典处理,这样可以减小数据的大小,同时也使得分析的语料更有说服力. 将构建的语料库进行初步切词处理和去除停用词如图 3 所示:



Index	fileContent	filePath	tag1	tag2	tag3	tag4	tag5
0	2017年一季度北京市农产品生产者价格同比下降4%日期: 2017-04-18作者: 来源: ...	C:\Users\PengZheng\Desktop\agri_news_reco...	下降	价格	同比	草莓	畜牧业
1	京郊农民用上农事服务 APP日期: 2017-04-12作者: 李锐来源: 农民...	C:\Users\PengZheng\Desktop\agri_news_reco...	手机	软件	农产品	微商	农事
2	北京猪肉批发价同比下降32.16% 养殖上市公司上半年业绩分化日期: 2017-05-10作...	C:\Users\PengZheng\Desktop\agri_news_reco...	业绩	生猪	下降	量	上半年
3	北京菜价小幅涨 蛋价已到顶日期: 2017-04-05作者: 陈琳来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	大葬	价格	上涨	鸡蛋	上市
4	北京菜价降至5年来同期最低日期: 2017-04-19作者: 陈琳来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	价格	元	斤	下降	圆白菜
5	市场供求总体趋稳 肉鸡行情逐渐升温日期: 2017-04-18作者: 刘国信来源: 京...	C:\Users\PengZheng\Desktop\agri_news_reco...	肉鸡	毛鸡	万套	种鸡	祖代
6	拍卖销量主导市场 玉米价格稳中见涨日期: 2017-05-03作者: 刘国信来源: 京...	C:\Users\PengZheng\Desktop\agri_news_reco...	玉米	玉米价格	拍卖	临储	收购
7	新发地统计显示:蔬菜产地北移价格微涨日期: 2017-04-07作者: 孙杰来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	蔬菜	大葬	顺代	转播	发地
8	生猪价格持续回调 鸡蛋行情不容乐观日期: 2017-05-09作者: 来源: 京郊日报...	C:\Users\PengZheng\Desktop\agri_news_reco...	玉米	玉米价格	鸡蛋	下跌	临储
9	盆栽蔬菜将扮靓世园会日期: 2017-04-12作者: 肖丹来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	蔬菜	品种	技术	体系	推广站
10	端午蔬菜进行产地转换或进入换茬期 菜价小幅扬日期: 2017-05-31作者: 陈琳来源: ...	C:\Users\PengZheng\Desktop\agri_news_reco...	价格	洋葱	豆角	出现	大葱
11	肉价缓慢回 开始往下跌日期: 2017-04-26作者: 陈琳来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	肉价	价格	保持	出现	下降
12	芹菜小冬瓜价格大幅冰日期: 2017-04-12作者: 于建来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	芹菜	冬瓜	下降	大幅冰	预示
13	菜贱伤农应该怎么办?日期: 2017-05-18作者: 于建来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	蒜薹	滞销	一些	价格	蔬菜
14	鲜蒜价格一周下降超三成日期: 2017-05-18作者: 于建来源: 北京...	C:\Users\PengZheng\Desktop\agri_news_reco...	鲜蒜	下降	出现	价格	大葬
	京津冀: 吃"生态饭" 发"绿色财"	C:\Users\PengZheng					

图 3 农业新闻分词结果

Fig. 3 Agricultural vocabulary segmentation results

1.4 文本分类

基于 KNN(k-NearestNeighbor,简称 KNN)的分类器是一种常见的有监督学习的分类方法. K 近邻的输入为实例的特征向量,对应于特征空间的点;输出位实力的类别,可以取多类. 该方法假设给定一个训练数据集,其中的实例类别已给定. 分类时对新的实例,根据其 k 个最近邻的训练实例的类别,通过多数表决的等方式进行预测.

首先加载训练文本,并将数据集进行切分进行初步的预处理. 接下来,通过调用 fit_transform 接口进行训练样本数据,生成词语的 TF-IDF 向量空间模型. 直接调用 python 中 sklearn 库的 KNN 方法进行分类器的训练,以保证模型的最佳效果.

待训练好分类器以后,加载待预测文本数据,经行未知类别样本的预测. 经测试,该方法准确并快速的将未知文本进行了自动分类.

2 农业信息集成与推荐

当对农业新闻文本进行向量优化以后,接下来便可以对于收集到的农业类新闻文本,采用余弦相似度(cosine similiarity)计算多篇文章间的相似程度,通过计算不同的向量的差异的大小,来计算文本的相似度[11]. 相似度度量的值越小,说明个体间相似度越小,相似度的值越大说明两篇文本的差异越大. 余弦相似度计算原理如下:

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} * \sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad .$$

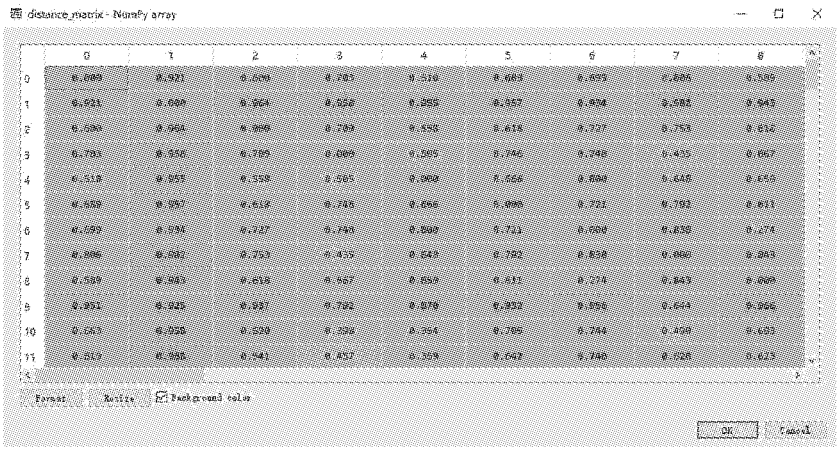


图 4　文章余弦相似度计算

Fig. 4　Article cosine similarity calculation

如图 4 所示,通过对文章进行相似的计算分析可以看出,文章本身之间相似度最高,所以对角线为 0. 由此,对每一篇文章进行向量化处理,构建出与这篇文章最相关的 5 篇文章并按照相似程度进行排序.

3　系统运行结果

该系统可以实现农业新闻的自动爬取,利用 python 的 scrapy 框架进行对主流农业网站的信息爬取. 爬取到的数据经过清洗整理后存储在 mongodb 数据库中. 网站的主界面在不登陆情况下,显示各个分类后的新闻模块;用户注册后,浏览不同类别的文章后,在历史浏览中会显示浏览记录,方便用户今后查阅. 爬取到的信息通过机器学习算法自动生成摘要展现在文章列表中,将文章最精华的部分展示给用户减少用户的信息处理时间. 通过记录过用户浏览过的历史记录和偏好,在用户点击一定数量的文章以后,后台通过算法自动推算出用户最感兴趣的内容,将结果按照相关程度进行排序最后展示在“我的推荐”一栏. 用户登录后会根据用户浏览记录及感兴趣标签等信息显示个性化推荐后的新闻以及浏览过的农业新闻记录,用户可以浏览最近天气以及种植相关信息. 系统在实际运行中,可极大方便用户的检索时间,用户可以最快速的定位到自己感兴趣的话题和内容,该系统对农业的发展具有一定的促进作用. 系统的最终运行部分界面如图 5 所示:

图5 系统最终实现界面

Fig. 5 System interface

4 总结与展望

本文将机器学习算法应用于传统的农业领域,实现了对种植信息的集成和挖掘,文本的自动分类以及个性化推荐等相关功能,在一定程度上满足了农民用户的现实需求. 随着大数据和移动计算时代的来临,往往使用单一数据源的静态历史数据方法的推荐系统[10],无法满足用户的需求因为用户在不同领域具有不同的兴趣[14-17],并没有考虑到用户的兴趣也随着时间会发生变化,今后可以考虑用户的选择受当前所处的地点,时间,周围相关人员等众多因素的影响所以更为智能推荐系统可以利用大数据和移动计算技术来增强"跨域"感知能力,构建推荐平台.

参考文献

[1]何洁. 基于Web使用数据挖掘的个性化推荐系统设计[J]. 数字技术与应用,2012(07):141-142.

[2]游兰,彭庆喜,王时绘. 基于Web使用挖掘的个性化站点研究[J]. 江汉大学学报(自然科学版),2005(03):51-54.

[3]姜楠,赵杏,狄查美玲,等. 移动农业信息推荐系统设计[J]. 大连民族大学学报,2016,18(05):505-508.

[4]陈龙飞,赵雪. 信息推荐技术与农资网站个性化推荐技术综述[J]. 河北科技师范学院学报,2013(04):46-51.

[5]张峰,茶正早,罗微,等. 面向中低端手机的移动农业应用软件研究——以香蕉小助手为例[J]. 安徽农业科学,2009(18):8806-8808.

[6]贾宝红,王晓蓉,马雪,等. 天津市农业信息推送服务系统设计与实现[J]. 山西农业科学,2015,43(10):1329-1332+1362.

[7]牛秀萍. 基于隐马尔科夫模型词性标注的研究[D]. 太原理工大学,2013.

[8]姜丽红,徐博艺,席俊红. 基于案例推理的过滤算法及智能信息推荐系统[J]. 清华大学学报(自然科学版),2006(S1):1074-1077.

[9]吴润方. 农业专家系统应用综述[J]. 科技广场,2016,172(03):179-181.

[10]刘建国,周涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009(01):1-15.

[11]张小彬. 中文Web文本分类关键技术研究与实现[D]. 西安电子科技大学,2011.

[12]马建斌,李滢,滕桂法,等. KNN和SVM算法在中文文本自动分类技术上的比较研究[J]. 河北农业大学学报,2008(03):120-123.

[13]郭平,刘波,沈岳. 农业云大数据自组织推送关键技术综述[J]. 软件,2013(03):1-6.

[14]赵璞,朱孟帅,秦波,等. 农业APP研究进展及展望[J]. 农业展望,2016(02):59-64.

[15]赵秋云,魏乐,舒红平,等. 农业信息化应用软件开发平台设计与实现[J]. 农机化研究,2015(11):230-235.

[16]付娟妮. 基于信息用户的新闻推荐系统特点及构建[J]. 企业科技与发展,2013(15):39-40.

[17]李春子,叶颖泽,贺立源. 提高我国农业网站建设质量的方法探讨[J]. 高等农业教育,2009(09):93-95.

(责任编辑:张阳,付强,李建忠,罗敏;英文编辑:周序林)