doi:10.6041/j.issn.1000-1298.2018.07.020

基于农业网络信息分类的热词自动提取方法

段青玲[1]　张　璐[1]　刘怡然[1]　王沙沙[2]

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 北京农信通科技有限责任公司, 北京 100081)

摘要: 热词提取对于监控和分析农业舆情具有重要意义,目前已有一定研究基础,但仍存在针对性差等问题,无法满足农业领域不同产业用户群的个性化需求,为此,提出一种基于农业网络信息分类的热词自动提取方法。首先采用多标记分类算法对文本语料进行分类,按分类类别构建语料库,然后采用基于信息熵的方法对每个类别分别提取热词候选词,最后采用基于时间变化的方法进行候选词热度计算,根据候选词热度排序结果得到热词。本文抽取农业网站上的 15 354 条文本进行实验,结果表明,热词提取准确率达到 0.9 以上,能够较高质量地提取农业热词,为不同农业用户群体发现和分析产业热点提供帮助。

关键词: 农业网络信息; 农业舆情监测; 热词; 多标记分类; 热度计算

中图分类号: TP391　　**文献标识码**: A　　**文章编号**: 1000-1298(2018)07-0160-08

Automatic Extraction Method of Hot Words Based on Agricultural Network Information Classification

DUAN Qingling[1]　ZHANG Lu[1]　LIU Yiran[1]　WANG Shasha[2]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *Agricultural Information Technology Limited Liability Company of Beijing*, *Beijing* 100081, *China*)

Abstract: With the vigorous development of the Internet, the network information grows rapidly, so does the agricultural network information. Extracting hot words from massive information is of great significance for monitoring and analyzing agricultural public opinion. Up to now, there is some research on hot words extraction, but there are still many problems such as poor pertinence. Existing hot word extraction methods cannot meet the personalized needs of users in different industries in agriculture. Therefore, a method of automatically extracting hot words based on agricultural network information classification was proposed. Firstly, the texts were classified by using the multi-label classification algorithm and multiple corpuses were built according to the classification categories. Secondly, the hot word candidates for each category were extracted by using the method based on information entropy. Thirdly, the heat of each hot word candidate was calculated by using the method based on time variation. Finally, these candidates were sorted by heat degree, and hot words were got according to the sorting results. Totally 15 354 texts from agricultural websites were extracted for the experiment, automatically obtaining the hot words in the specified time period. The experiment results showed that the accuracy was over 0.9. It proved that the proposed method can extract agricultural hot words with high quality and help different agricultural user groups find and analyze the hot spot information of the industry.

Key words: agricultural network information; agricultural public opinion monitoring; hot word; multi-label classification; heat calculation

0　引言

随着农业网站数量的迅速增长,提取农业网络热点信息对于实时监测和分析农业舆情,引导农业产业结构,维护社会稳定具有重要意义。热词是反映一个时期内热点问题的一种词汇现象,农业热词则反映了一个时期内涉农用户群体所关注的热点事件。例如,农业热词“蒜你狠”、“糖高宗”等反映了

收稿日期: 2017-12-15　修回日期: 2018-02-28

基金项目: 国家高技术研究发展计划(863 计划)项目(2013AA102306)和“十二五”国家科技支撑计划项目(2012BAD35B06)

作者简介: 段青玲(1967—),女,教授,博士生导师,主要从事智能信息处理研究,E-mail: dqling@cau.edu.cn

部分农产品价格居高不下和消费者心理不满的现象。因此，通过提取农业网站热词可以及时掌握农业行业发展动向、实时监测农业网络信息动态，利于相关部门进行正确的舆论引导和分析。

近年来，热词提取方法主要有基于规则过滤的方法[1-3]、基于热词数据库构建的方法[4-5]和基于热度权值排序的方法[6-9]。基于规则过滤的方法通过选择多个过滤特征构造判定规则，去除大量无关信息，再利用词频等设计热词抽取算法，获得热点信息，该方法中过滤特征的选择和判定规则中阈值的确定比较困难。基于热词数据库构建的方法通过构建待提取热词主题数据库，结合大数据分析技术[10-12]以及维度划分技术[13]，找出该主题下的热词，该方法提取出的热词很大程度上依赖于热词数据库的构建。基于热度权值排序的方法通过命名实体识别技术[14-15]或新词识别技术[16-17]获得热词候选词，再结合热词特点，进行候选词热度计算，热度值排在前列的候选词即为热词，但目前该方法中候选词提取和热度计算考虑的因素均具有单一性。

农业领域涵盖面广，涉及种植业、养殖业等多个产业，具有不同的用户群体和管理部门，现有的热词提取方法应用到农业中无法满足不同产业用户群的个性化需求。因此，本文将文本分类技术与热词自动提取技术相结合，针对每个农业产业类别分别提取热词，挖掘不同用户群体和农业管理部门所关注的信息热点，确保不同产业用户快速获取本产业信息动态。在热词自动提取方法中，针对目前该项研究中提取的热词词性单一等问题，提出基于信息熵的热词候选词提取方法和基于时间变化的热度计算方法。本文所述方法采用农业网站上的数据进行验证。

1　材料和方法

本文研究目标是根据农业用户的个性化需求，有针对性地为各用户提供农业网络热点信息。基于该领域现有的研究过程，本文提出基于农业网络信息分类的热词自动提取方法，流程如图 1 所示。

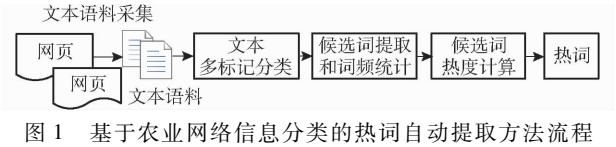


图 1　基于农业网络信息分类的热词自动提取方法流程

Fig. 1　Automatic extraction process of hot words based on agricultural network information classification

首先采用信息自动抽取技术获取网页上的农业文本信息；其次根据用户需求设定分类类别，采用多标记分类方法进行文本语料分类；再采用基于信息熵的方法针对各类别分别提取热词候选词，并进行单日词频统计；为从候选词中挑选出热词，最后提出基于时间变化的方法计算候选词热度，将热度排在前列的候选词作为该类别的热词。

1.1　文本语料采集

本文通过信息自动抽取技术从中国农业信息网、搜猪网、中国农资网、中国三农网、中国水产养殖网、中国农机网、中国棉花交易网等农业网站获取 2017 年 6 月 9—16 日的农业文本信息作为实验语料。

1.2　文本语料分类

对采集到的农业文本语料采用多标记分类算法进行分类，再针对每个类别分别提取热词。多标记分类[18-20]是指一个对象可以同时划分到多个类别中。例如，一个介绍西红柿的文本，可以同时划分到种植业和农产品市场两个类别中。对农业文本信息进行多标记分类，符合信息同时具有多个标记的实际情况，利于相关行业人员查询更为完整的信息资源。

农业文本多标记分类的处理流程主要包括：文本语料预处理和分类词库构建阶段、模型训练阶段以及测试阶段，如图 2 所示。

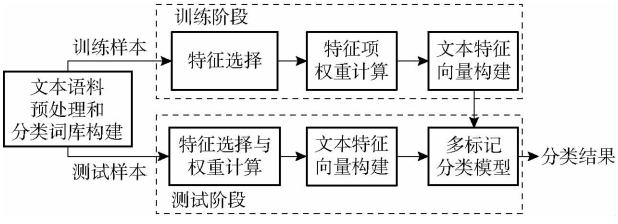


图 2　农业文本多标记分类流程

Fig. 2　Multi-label classification process for agricultural texts

1.2.1　文本语料预处理和分类词库构建

中文文本的预处理包括分词、词性标注、去停用词 3 个步骤。采用 NLPIR 汉语分词系统包（参考 http://ictclas.nlpir.org/downloads）对实验语料进行分词[21]，同时标注出词性[22]。停用词[23]主要是指使用十分广泛但实际意义不大的词，本文根据停用词表去除停用词。根据《国民经济行业分类与代码》[24]构建农业分类词库。

1.2.2　特征选择与权重计算

文本特征选择是指从文本中选取代表性的特征项来表示整个文本信息。本文采用基于分类词库的方法进行文本特征选择。首先通过计算文档频率进行特征选择，然后通过分类词库对特征集合进行扩充。文档频率 D_F（Document frequency）的计算公式[25]为

$$D_F(F_j) = \frac{T_F(F_j)}{A} \tag{1}$$

式中 $T_F(F_j)$——特征词 F_j 在语料集上的频率

A——语料集的总样本数

本文根据构建的农业分类词库对特征集合进行扩充。如:特征词“绿豆”含有关键词库中的关键词“豆”,则将“绿豆”加入到特征集合中。通过基于分类词库扩充的方法进行特征选择,避免对分类有效的低频率词语不能入选特征词的问题。图 3 为文本特征选择结果。第一条记录“首都”为特征词,“/n”表示该特征词为名词,“5.681877639268151”为文档频率。

首都/n 5.681877639268151
价值/n 3.413194097949787
村庄/n 4.254761283628006
动能/n 4.640423764439999
示范村/n 4.90868775103467
大白菜/n 4.583265350600041
等级/n 4.429114670772783
地址/n 5.394195566816371
气氛/n 4.90868775103467
龙虾/n 4.701048386256425
年收入/n 5.394195566816371

图 3 文本特征选择结果

Fig. 3 Results of text feature selection

权重计算是指以数字形式表示特征词在文本中的重要程度。本文提出基于改进 TF - IDF(Term frequency - inverse document frequency)方法进行特征项权重计算。该方法不仅考虑特征词在整个语料集中的重要程度,而且考虑特征词在各个类别之间以及各个类别内的差异性。计算公式为

$$W(F_j,D_i)=\frac{(1+\lg t_f(F_j,D_i))P(F_j,L_k)C(F_j,L_k)\lg\dfrac{A}{T_F(F_j)}}{\sqrt{\sum_{j=1}^{c}\left[(1+\lg t_f(F_j,D_i))P(F_j,L_k)C(F_j,L_k)\lg\dfrac{A}{T_F(F_j)}\right]^2}} \tag{2}$$

其中

$$P(F_j,L_k)=\sqrt{\left(T_f(F_j,L_k)-\frac{\sum_{t=1}^{q}T_f(F_j,L_t)-T_f(F_j,L_k)}{q-1}\right)^2}\cdot\lg\left(\frac{1}{H(F_j)+0.0001}+1\right) \tag{3}$$

$$C(F_j,L_k)=\frac{T_f(F_j,L_k)}{A_k} \tag{4}$$

式中 $W(F_j,D_i)$——特征词 F_j 在文本 D_i 中的权重

$t_f(F_j,D_i)$——特征词 F_j 在文本 D_i 中的频率

c——特征词的个数

$P(F_j,L_k)$——特征词 F_j 的类间区分程度

$C(F_j,L_k)$——特征词 F_j 在类别 k 中分布的均匀程度

$T_f(F_j,L_k)$——特征词 F_j 在类别 k 上的频率

q——类别数

$H(F_j)$——特征词 F_j 的信息熵[26]

A_k——类别 k 的样本数

通过式(3)、(4)分别计算特征词对于各个类别的类间区分度和类内均匀度,计算结果如图 4 所示。图 4 中 c1 表示种植业,c2 表示种业,c3 表示畜牧业,c4 表示兽医,c5 表示渔业,c6 表示农垦,c7 表示农机,c8 表示农产品质量安全,c9 表示农村经营管理,c10 表示科教,c11 表示农产品市场。第 1 条记录中的“0.24815460223223648”表示对于种植业类中语料来说,特征词“流域”的类间区分程度,“0.01020408163265306”表示特征词“流域”在种植业中的均匀度。

c1 流域/n 0.24815460223223648 0.01020408163265306
c10 流域/n 0.24815460223223648 0.004739336492890996
c11 流域/n 0.20679550186019705 0.007662835249042145
c2 流域/n 0.24815460223223648 0.007936507936507936
c3 流域/n 1.5716458141374974 0.020833333333333332
c4 流域/n 0.0 0.0
c5 流域/n 1.116695710045064 0.03225806451612903
c6 流域/n 0.24815460223223648 0.007633587786259542
c7 流域/n 0.0 0.0
c8 流域/n 0.24815460223223648 0.006134969325153374
c9 流域/n 0.24815460223223648 0.006369426751592357

图 4 特征词类别区分度计算结果

Fig. 4 Results of feature word category distinction calculation

1.2.3 文本特征向量构建

本文特征项权重计算方法考虑了特征词的类别区分度,对处于不同类别中的同一样本分别计算权重,对同一样本进行特征项权重融合,构建文本特征向量,即

$$V_{ij}=\sum_{Y_{ik}=1}U(F_j,k) \tag{5}$$

式中 $U(F_j,k)$——对于特征词 F_j,样本 D_i 在类别 k 中的权重

$Y_{ik}=1$ 表示样本 D_i 划分到类别 k 中。

构建文本特征向量后,采用 RAKEL(Random k-labelsets)多标记分类[27]算法训练农业文本多标记分类模型。在测试阶段,通过构建文本特征向量,利用训练好的分类模型实现农业文本自动分类。最后根据多标记分类的结果,按分类类别构建语料库。

1.3 候选词提取和词频统计

农业热词反映了一个时期内农业用户群体所关注的热点事件,具有一定的概括性。从文本语料方面来说,只有体现文本重点内容的词语才有可能成为热词。因此,本文采用基于信息熵的方法提取文本关键词,将其作为农业热词候选词,包括名词、动词、形容词等。

基于信息熵的关键词提取方法在采用常用的 TF - IDF 特征外,还将词的信息熵、词的空间局部性和词所在句子的位置作为非常重要的因素加入到词的权重中[28]。

词语 w 的信息熵 $E(w)$ 表示[26]为

$$E(w) = -\sum p(x)\lg p(x) \tag{6}$$

式中　$p(x)$——字符 x 在所有字符中出现的概率

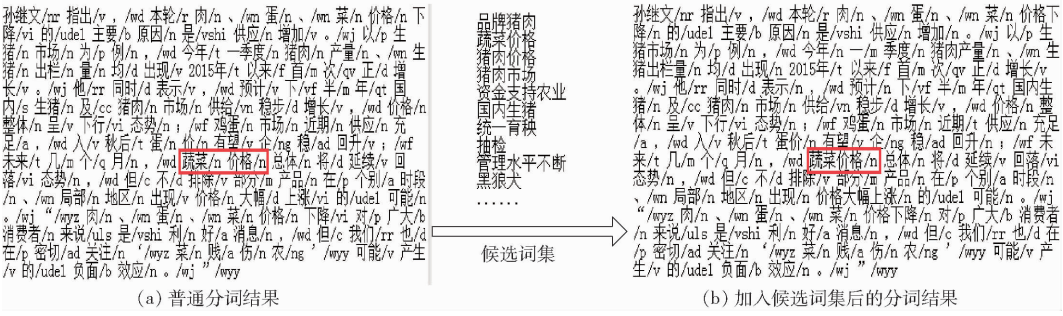
设 $L = \{l_1, l_2, \cdots, l_n\}$ 为词语 w 的左邻接集合[28]，$R = \{r_1, r_2, \cdots, r_m\}$ 为词语 w 的右邻接集合[28]，其中 l_i 和 r_j 分别为词语 w 的左邻字符串和右邻字符串。设 l_i、r_j 的出现频次分别为 n_i 和 m_j，左邻接集合和右邻集合中字符串的频次总和分别为 N_i 和 M_j。词语 w 左信息熵 $E_l(w)$ 和右信息熵 $E_r(w)$ 的公式[28]为

$$E_l(w) = -\sum_{i=1}^{n} \frac{n_i}{N_i}\lg\frac{n_i}{N_i} \tag{7}$$

$$E_r(w) = -\sum_{j=1}^{m} \frac{m_j}{M_j}\lg\frac{m_j}{M_j} \tag{8}$$

熵反映了信息的不确定性，而左信息熵和右信息熵则反映了词语 w 邻接字符串的不确定性。左信息熵和右信息熵越大，表明邻接字符串的不确定性越大，则词语 w 的使用越灵活，其成为关键词的概率也越大。

采用上述算法不仅提取出“黑狼犬”这样的单候选词，而且包括“蔬菜价格”这样的组合候选词。针对“蔬菜价格”，普通分词时会将其分为“蔬菜”和“价格”两个词，从而在对“蔬菜价格”进行词频统计时，该词词频为零。为了避免这种情况，将候选词集加入到分词词典中，则在分词时不会出现组合候选词被分割的情况。在此分词基础上按类别分别进行单日候选词词频统计。图 5 为普通分词与加入候选词集分词的结果对比。图 5a 中将“蔬菜价格”分割为“蔬菜”和“价格”两个词；图 5b 中“蔬菜价格”为一个词。



(a) 普通分词结果　　(b) 加入候选词集后的分词结果

图 5　分词结果对比

Fig. 5　Comparison of segmentation results

1.4　候选词热度计算

热词候选词具有一定的农业网络关注度，但是各候选词的受关注程度不一致，只有达到较高关注度的候选词才能称之为热词，因此本文通过计算候选词热度，挑选热度排在前列的候选词作为热词。针对农业热词单日词频高、历史波动大的特点，提出基于时间变化的热度计算方法。该方法考虑候选词的基础词频和历史波动 2 个因素，分别记为基础权值和波动权值。

基础词频是指候选词的单日词频。为了避免单日文本数不同对基础权值的影响，故进行平滑处理。基础权值 B 的计算公式[6]为

$$B = \lg(1 + \lg(1 + \lg(t_f + 1))) \tag{9}$$

式中　t_f——候选词的单日词频

历史波动采用基础权值的整体波动性、长期变化和短期变化表示。本文提出基础权值的整体波动性 V、长期变化 L 和短期变化 S 的计算方法分别为

$$V = \sqrt{\sum_{j=1}^{t}\left(B_j - \frac{1}{t}\sum_{i=1}^{t} B_i\right)^2} \tag{10}$$

$$L = \frac{B_t + 1}{\dfrac{\sum_{i=1}^{t-1} B_i}{t-1} + 1} \tag{11}$$

$$S = \frac{B_t + 1}{B_{t-1} + 1} \tag{12}$$

式中　t——实验数据周期

B_i——候选词第 i 天的基础词频

波动权值反映候选词在一段时间内的频率波动情况。本文定义波动权值 F 的计算公式为

$$F = 0.4V + 0.4L + 0.2S \tag{13}$$

候选词的基础权值和波动权值分别体现农业热词的特点。因此提出热度权值 H 的计算公式为

$$H = 0.5B + 0.5F \tag{14}$$

式(13)和式(14)中权重系数的确定通过多组实验得出。具体方法为，将权重系数以 0.1 为间距，分别对公式中涉及到的因素赋予权重，进行实验，最终得出。对于式(13)，当整体波动性 V 和长期变化 L 的权重系数分别设置为 0.4，短期变化 S 的权重系数设置为 0.2 时，实验效果最佳；对于式(14)，当基

础权值 B 和波动权值 F 的权重系数均设置为 0.5 时,实验效果最佳。

对于实验语料中种植业类的热词候选词“花球”和“水利”,通过上述热度计算方法得出“花球”的热度权值为 0.680 039 128 501 011 3,“水利”的热度权值为 0.281 568 874 380 245 75,在该类候选词中,“花球”的热度排第 4 位,“水利”的热度排第 3 437 位,因此,“花球”属于热词,“水利”不是热词。另一方面,从“花球”和“水利”的频率变化图也可以得出相同的答案,如图 6 所示,“水利”一词每天的频率都很高,但是前后期波动不大;而“花球”一词前期频率一直为零,后期突然上升,前后期波动很大,所以说“花球”是热词,“水利”不是热词。

对候选词进行热度计算后,按照权值进行热度排序,得到各类别在指定时间段内的热词。

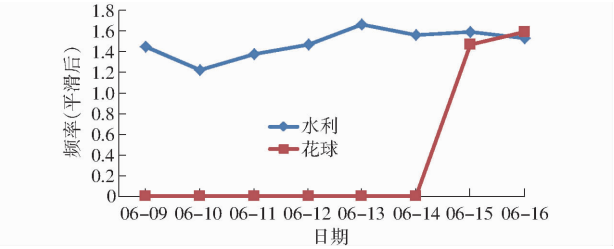


图 6 候选词频率变化曲线

Fig. 6 Frequency variation curves of candidate words

2 结果与分析

本文获取了中国农业信息网、搜猪网、中国农资网、中国三农网、中国水产养殖网、中国农机网、中国棉花交易网等农业网站 2017 年 6 月 9—16 日的农业信息,共 15 354 条。采用本文所述方法自动提取 6 月 16 日的农业热词。表 1 为单日实验文本数统计表。

表 1 实验语料统计

Tab. 1 Experimental corpus statistics

日期	6 月 9 日	6 月 10 日	6 月 11 日	6 月 12 日	6 月 13 日	6 月 14 日	6 月 15 日	6 月 16 日
文本数	2 119	587	507	2 446	2 536	2 471	2 471	2 217

本文编码环境为 MyEclipse 2014。文本分词和农业热词候选词提取通过调用 NLPIR 汉语分词系统包中的分词函数和关键词提取函数实现。

针对热词提取的结果,采用前 N 个返回结果的准确率和二元偏好值来衡量本文热词提取方法的性能[29-30]。

前 N 个返回结果的准确率($P@N$)即计算在返回的前 N 个最优结果的准确率,计算公式为

$$P@N = \frac{T}{N} \tag{15}$$

式中 T——返回的前 N 个结果中正确结果的个数

二元偏好值 B_{pref}(Binary preference)用以评价返回结果中,正确结果与非正确结果的相对位置,主要体现方法能否将热词在非热词之前返回。计算公式为

$$B_{pref} = \frac{1}{R}\sum_{r}\left(1 - \frac{|b|}{R}\right) \tag{16}$$

式中 R——对每个主题,已判定结果中正确结果个数

r——正确结果项

b——前 R 个不正确结果集合的子集

$|b|$——当前正确结果项之前不正确的结果个数

2.1 文本语料分类结果

用户可以根据需求设定分类类别,对文本语料分类。本文设定为 11 个类别:种植业、种业、畜牧业、兽医、渔业、农垦、农机、农产品质量安全、农村经营管理、科教和农产品市场。这 11 个类别的内容范围不是完全独立的,各类别之间有些信息是共同涉及的。为了保证各类别信息的完整性,采用多标记分类算法对 8 d 的语料分别进行分类。分类结果如表 2 所示。

由文献[31]中的多标记分类评价指标对分类结果进行评价,得出准确率为 92.54%,汉明损失为 0.056 4,一类错误为 0.063 6,覆盖率为 1.597 3,排序损失为 0.046 5。

2.2 候选词提取与词频统计结果

实验将 6 月 9—16 日作为一个周期,通过分析此周期内热词候选词的频率变化情况来提取 6 月 16 日的农业热词。首先从 6 月 16 日的 11 个类别的文本语料分别提取热词候选词;然后将候选词集加入到分词词典中,对 8 d 的文本语料分别进行分词;最后进行候选词单日词频统计。各类别候选词个数统计结果如表 3 所示,部分候选词单日词频统计结果如表 4 所示。热词候选词单日词频统计结果是为计算候选词热度做准备的。

2.3 候选词热度计算结果

对 11 个类别中的热词候选词分别计算热度。表 5 列出了各类别中部分候选词的热度计算结果。

表 2　多标记分类结果

Tab. 2　Results of multi-label classification

类别	6 月 9 日	6 月 10 日	6 月 11 日	6 月 12 日	6 月 13 日	6 月 14 日	6 月 15 日	6 月 16 日
种植业	866	245	220	103	1 084	1 064	1 049	926
种业	99	19	24	81	104	123	121	97
畜牧业	393	135	112	506	483	466	486	427
兽医	116	36	30	138	106	136	107	116
渔业	131	49	34	163	158	135	126	132
农垦	120	27	26	163	192	154	165	127
农机	132	16	16	153	163	119	121	94
农产品质量安全	444	124	84	495	497	491	515	493
农村经营管理	833	232	208	911	986	964	954	838
科教	494	118	99	554	595	608	576	485
农产品市场	708	226	187	823	815	771	839	755

表 3　候选词个数统计结果

Tab. 3　Statistics of number of candidate words

类别	种植业	种业	畜牧业	兽医	渔业	农垦	农机	农产品质量安全	农村经营管理	科教	农产品市场
候选词	4 733	704	2 341	812	823	809	577	2 744	4 333	2 854	3 836

表 4　种植业部分候选词词频统计结果

Tab. 4　Statistics of word frequency of part candidate words in crop farming

候选词	6 月 9 日	6 月 10 日	6 月 11 日	6 月 12 日	6 月 13 日	6 月 14 日	6 月 15 日	6 月 16 日
统一育秧	0	0	0	0	0	0	0	4
茶叶基地	1	2	0	4	1	0	0	7
食用油	20	6	4	19	21	79	27	22
椒农	0	0	0	0	0	0	4	16
插秧	30	0	33	21	52	17	22	16
全县蔬菜	0	0	0	1	4	2	4	5
生态园	10	6	1	9	8	10	7	3
辣椒产业	2	1	0	0	4	4	4	24
茶苗	2	0	0	0	1	3	0	12
玉米淀粉	5	0	19	3	6	1	2	37

表 5　候选词热度计算结果

Tab. 5　Heat calculation results of candidate words

类别	热度排行第 1		热度排行第 2		热度排行第 3		热度排行第 4	
	候选词	热度	候选词	热度	候选词	热度	候选词	热度
种植业	黑狼犬	0. 788	乐视	0. 757	食品生产许可	0. 698	花球	0. 680
种业	花球	0. 809	罂粟	0. 779	黑狼犬	0. 759	王志永	0. 755
畜牧业	黑狼犬	0. 774	乐视	0. 699	执业兽医资格	0. 66	工作犬	0. 652
兽医	执业兽医资格	0. 794	解冻	0. 768	考试	0. 767	安倍	0. 764
渔业	整改	0. 859	噪声扰民	0. 826	苫盖	0. 806	金砖国家农业	0. 79
农垦	整改	0. 863	噪声扰民	0. 826	苫盖	0. 791	装修垃圾	0. 751
农机	王志永	0. 828	职业教育	0. 818	项目申报	0. 792	文化	0. 776
农产品质量安全	噪声扰民	0. 797	中国奶业	0. 755	苫盖	0. 734	武山	0. 725
农村经营管理	食品生产许可	0. 797	乐视	0. 781	沙化土地	0. 766	樟坑华侨新村	0. 743
科教	沙化土地	0. 783	武山	0. 781	执业兽医资格	0. 779	中国奶业	0. 735
农产品市场	黑狼犬	0. 806	乐视	0. 777	食品生产许可	0. 773	武山	0. 765

根据热词的定义，人工为每个类别标注出 100 个热词，采用前 N 个返回结果的准确率和二元偏好值对本文提出的热词提取方法的性能进行评价。对于 P@ N，主要考虑 P@ 4、P@ 20、P@ 50、P@ 100 这

4 个指标。对于 B_{pref} 的计算,选取 $R=100$。表 6 是本文热词提取方法返回结果的评价。

表 6 热词返回结果评价

Tab.6 Evaluation of returned results of hot words

类别	P@4	P@20	P@50	P@100	B_{pref}
种植业	1	0.95	0.92	0.81	0.768 0
种业	1	1	0.92	0.80	0.764 0
畜牧业	1	1	0.94	0.83	0.802 2
兽医	1	1	0.90	0.82	0.771 7
渔业	1	0.90	0.88	0.81	0.763 1
农垦	1	1	0.96	0.86	0.820 6
农机	1	1	0.92	0.82	0.782 1
农产品质量安全	1	0.95	0.86	0.80	0.743 2
农村经营管理	1	1	0.96	0.85	0.815 7
科教	1	0.95	0.88	0.85	0.796 5
农产品市场	1	0.90	0.92	0.82	0.765 0

注:P@4 指前 4 个返回结果的准确率。

本文针对不同农业产业分别提取热词,得出各类别的热点关注,便于不同农业产业的用户群监测本产业动向。由表 6 可知,各类别中 P@4 的值均为 1,表明在各类别中热度排在前 4 的候选词均是该类别的热词,验证了本文热词提取方法的正确性。P@4 的结果优于 P@20,P@20 的结果优于 P@50,P@50 的结果优于 P@100,表明返回结果数越少,准确率越高。B_{pref} 表明在热词返回结果中,非热词在热词之前出现的次数情况。该值越大,表明非热词在热词之前出现的次数越少,算法效果越好。由表 6 可知,各类别中 B_{pref} 值稳定在 0.8 左右,实验效果较好。

从表 5 中各类别的热词可以看出,有些热词在多个类别中出现。如热词“食品生产许可”在种植业、农村经营管理和农产品市场 3 个类别中都出现。这是由于采用多标记分类算法对文本语料进行分类,保留了文本的多样性特点,保证了用户查找行业信息的完整性。种植业、农村经营管理和农产品市场 3 个产业的用户在查找本产业信息时都可以通过热词“食品生产许可”获取到与之相关的新闻。

3 结论

(1)将农业文本分类技术与热词提取技术结合,提出了一种基于农业网络信息分类的热词自动提取方法。根据用户需求,设定分类类别,针对每个类别分别提取热词,挖掘出不同用户群体和农业管理部门所关注的信息热点,实现不同产业用户群快速获取产业动态,进行分析和决策的功能。采用前 N 个返回结果的准确率对实验结果进行评价,当 N 取值为 20 时,其准确率达到 0.9 以上。

(2)针对目前研究中提取的热词词性单一等问题,提出基于信息熵的热词候选词提取方法;结合热词特点,提出基于时间变化的候选词热度计算方法,有效衡量了热词候选词的受关注程度,实现农业热词的准确提取。

参 考 文 献

1 汪洋,帅建梅,陈志刚. 基于海量信息过滤的微博热词抽取方法[J]. 计算机系统应用,2012,21(11):131 - 136.
WANG Yang, SHUAI Jianmei, CHEN Zhigang. Hot word extraction for microblog based on massive data filtering[J]. Computer Systems & Applications,2012,21(11):131 - 136. (in Chinese)

2 郝晓玲,茅嘉惠,于秀艳. 微博热词抽取及话题发现研究[J]. 情报杂志,2015,34(6):109 - 113,157.
HAO X L, MAO J H, YU X Y. Micro-blogging hot words extraction and topic detection[J]. Journal of Intelligence,2015,34(6):109 - 113,157. (in Chinese)

3 郭冲. 基于新闻标题的网络热词发现算法[J]. 计算机与现代化,2013(3):58 - 62,66.
GUO Chong. Algorithm of network hot word detection based on news title[J]. Computer and Modernization,2013(3):58 - 62,66. (in Chinese)

4 WU D, TANG X J. Preliminary analysis of Baidu hot words[C]// Proceedings of the 11th Youth Conference on Systems Science and Management Science. Wuhan: Wuhan University of Science and Engineering Press, 2011:478 - 483.

5 喻国明. 基于语料库方法的舆论热词数据库的构建——以 2011—2013 年全国两会舆情中心词和关联词的发现与分析为例[J]. 新闻与写作,2014(1):54 - 60.

6 李渝勤,孙丽华. 面向互联网舆情的热词分析技术[J]. 中文信息学报,2011,25(1):48 - 53,59.
LI Yuqin, SUN Lihua. Hot-word detection for internet public sentiment[J]. Journal of Chinese Information Processing, 2011, 25(1):48 - 53,59. (in Chinese)

7 耿升华. 新词识别和热词排名方法研究[D]. 重庆:重庆大学,2013.
GENG Shenghua. Newword recognition and hot word ranking methods[D]. Chongqing:Chongqing University,2013. (in Chinese)

8 王馨,王煜,王亮. 基于新词发现的网络新闻热点排名[J]. 图书情报工作,2015(6):68 - 74.

9 唐蓉青. 基于微博热词挖掘的新闻话题提取研究[D]. 长沙:湖南大学,2014.
TANG Rongqing. The study on the extraction of the news topic based on web mining of micro-blog hot words[D]. Changsha: Hunan University,2014. (in Chinese)

10 陈世敏. 大数据分析与高速数据更新[J]. 计算机研究与发展,2015,52(2):333 - 342.

CHEN Shimin. Bigdata analysis and data velocity[J]. Journal of Computer Research and Development,2015,52(2):333 – 342. (in Chinese)

11 程学旗,靳小龙,王元卓,等. 大数据系统和分析技术综述[J]. 软件学报,2014,25(9):1889 – 1908.
CHENG X Q, JIN X L, WANG Y Z, et al. Survey on big data system and analytic technology[J]. Journal of Software,2014, 25(9):1889 – 1908. (in Chinese)

12 LAZER D, KENNEDY R, KING G, et al. Big data. The parable of google flu: traps in big data analysis[J]. Science,2014, 343(6176):1203.

13 周芳芳,李俊材,黄伟,等. 基于维度扩展的 Radviz 可视化聚类分析方法[J]. 软件学报,2016,27(5):1127 – 1139.
ZHOU F F, LI J C, HUANG W, et al. Extending dimensions in Radviz for visual clustering analysis[J]. Journal of Software, 2016,27(5):1127 – 1139. (in Chinese)

14 李想,魏小红,贾璐,等. 基于条件随机场的农作物病虫害及农药命名实体识别[J/OL]. 农业机械学报,2017,48(增刊): 178 – 185. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 2017s029&flag = 1. DOI:10.6041/j.issn. 1000-1298.2017.S0.029.
LI Xiang, WEI Xiaohong, JIA Lu, et al. Recognition of crops, diseases, and pesticides named entities in Chinese based on conditional random fields[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017,48(Supp.):178 – 185. (in Chinese)

15 尹存燕,黄书剑,戴新宇,等. 中英命名实体识别及对齐中的中文分词优化[J]. 电子学报,2015,43(8):1481 – 1487.
YIN Cunyan, HUANG Shujian, DAI Xinyu, et al. Optimization of Chinese word segmentation in named entity recognition and word alignment[J]. Acta Electronica Sinica,2015,43(8):1481 – 1487. (in Chinese)

16 孙立远,周亚东,管晓宏. 利用信息传播特性的中文网络新词发现方法[J]. 西安交通大学学报,2015,49(12):59 – 64.
SUN Liyuan, ZHOU Yadong, GUAN Xiaohong. A method of discovering new Chinese words from Internet based on information propagation[J]. Journal of Xi'an Jiaotong University,2015,49(12):59 – 64. (in Chinese)

17 李钝,曹元大,万月亮. Internet 中的新词识别[J]. 北京邮电大学学报,2008,31(1):26 – 29.
LI Dun, CAO Yuanda, WAN Yueliang. Internet-oriented new words identification[J]. Journal of Beijing University of Posts and Telecommunications,2008,31(1):26 – 29. (in Chinese)

18 HE Z, WU J, LI T. Label correlation mixture model: a supervised generative approach to multilabel spoken document categorization[J]. IEEE Transactions on Emerging Topics in Computing,2015,3(2):235 – 245.

19 LIM Hyunki, LEE Jaesung, KIM Dae-Won. Optimization approach for feature selection in multi-label classification[J]. Pattern Recognition Letters, 2017,89:25 – 30.

20 何志芬,杨明,刘会东. 多标记分类和标记相关性的联合学习[J]. 软件学报,2014,25(9):1967 – 1981.
HE Z F, YANG M, LIU H D. Joint learning of multi-label classification and label correlations[J]. Journal of Software,2014, 25(9):1967 – 1981. (in Chinese)

21 韩冬煦,常宝宝. 中文分词模型的领域适应性方法[J]. 计算机学报,2015,38(2):272 – 281.
HAN Dongxu, CHANG Baobao. Approaches to domain adaptive Chinese segmentation model [J]. Chinese Journal of Computers, 2015,38(2):272 – 281. (in Chinese)

22 袁里驰. 基于改进的隐马尔科夫模型的词性标注方法[J]. 中南大学学报:自然科学版,2012,43(8):3053 – 3057.
YUAN Lichi. A part-of-speech tagging method based on improved hidden Markov model[J]. Journal of Central South University: Science and Technology,2012,43(8):3053 – 3057. (in Chinese)

23 熊文新,宋柔. 信息检索用户查询语句的停用词过滤[J]. 计算机工程,2007,33(6):195 – 197.
XIONG Wenxin, SONG Rou. Removal of stop word in users' request for information retrieval[J]. Computer Engineering,2007, 33(6):195 – 197. (in Chinese)

24 GB/T 4754—2011 国民经济行业分类与代码[S]. 2011.

25 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/OL]. 农业机械学报,2015,46(增刊): 174 – 179. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no = 2015S029&flag = 1&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2015.S0.029.
WEI Fangfang, DUAN Qingling, XIAO Xiaoyan, et al. Classification technique of Chinese agricultural text information based on SVM [J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2015,46(Supp.):174 – 179. (in Chinese)

26 SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal,1948, 27(3):379 – 423.

27 TSOUMAKAS G, KATAKIS I, VIAHAVAS I. Random k-labelsets for multilabel classification[J]. IEEE Transactions on Knowledge & Data Engineering,2011,23(7):1079 – 1089.

28 张华平. 大数据搜索与挖掘[M]. 北京:科学出版社, 2014.

29 BAEZA-YATES R, RIBEIRO-NETO B. Modern information retrieval the concepts and technology behind search[M]. Beijing: China Machine Press,2011.

30 BUCKLEY C, VOORHEES E M. Retrieval evaluation with incomplete information[C]// ACM SIGIR 2004 Proceedings, 2004:25 – 32.

31 ZHANG M. An improved multi-label lazy learning approach[J]. Journal of Computer Research & Development,2012,49(11): 2271 – 2282.