特约专稿
DOI：10.13998/j.cnki.issn1002-1248.20-1197

文本挖掘技术在农业知识服务中的应用述评

孙　坦[1]，丁　培[2]，黄永文[3*]，鲜国建[3]
（1.中国农业科学院，北京　100081；2.深圳大学 图书馆，深圳　518060；
3.中国农业科学院 农业信息研究所，北京　100081）

摘　要：［目的／意义］支撑数据密集型科学发现下科技创新生态的知识服务新业态正悄然形成。文本挖掘作为知识服务技术的核心，在知识服务新业态环境下面临挑战。本文旨在探讨在新环境下知识服务中应用文本挖掘技术的发展策略。［方法／过程］梳理文本挖掘技术框架，论证文本挖掘技术逐步成熟。结合农业领域分析文本挖掘在农业信息检索、智能问答、信息监测和知识抽取等方面应用，展示其在领域科技创新和产业应用中取得的良好应用效果。［结果／结论］结合中国国情提出以文本挖掘为主的知识服务技术发展策略：（1）基于文本挖掘技术构建专门知识服务系统；（2）重视语料库和基础知识库建设；（3）在重点领域优先开展和部署文本挖掘技术的应用。

关键词：文本挖掘；知识服务；信息抽取；知识组织
中图分类号：G302　　**文献标识码：**A　　**文章编号：**1002-1248（2021）01-0004-13

引用本文：孙坦, 丁培, 黄永文, 等. 文本挖掘技术在农业知识服务中的应用述评[J]. 农业图书情报学报, 2021, 33(1)：04-16.

Review on the Application and Development Strategies of Text Mining in Agriculture Knowledge Services

SUN Tan[1], DING Pei[2], HUANG Yongwen[3*], XIAN Guojian[3]
(1. Chinese Academy of Agricultural Sciences, Beijing 100081; 2. Shenzhen University Library, Shenzhen 518060;
3. Agricultural Information Institute of CAAS, Beijing 100081)

Abstract: [Purpose/Significance] Under the new ecological environment of scientific and technological innovation

收稿日期：2020-11-20
基金项目：国家社会科学基金项目“融合多种知识组织体系的认知搜索模式研究”（20BTQ014）；广东省哲学社会科学规划学科共建项目“支持深度知识发现的文内数据与文献关联研究”（GD18XTS07）
作者简介：孙坦（ORCID：0000-0002-8257-5064），男，博士，二级研究员，研究方向为数字信息描述与组织。丁培（ORCID：0000-0002-0045-3312），男，博士，馆员，研究方向为数字图书馆、科学数据和知识发现。鲜国建（ORCID：0000-0003-4332-1958），男，博士，研究员，研究方向为大数据融汇治理、知识组织
*通信作者：黄永文（ORCID：0000-0002-4486-3233），女，博士，副研究馆员，研究方向为知识组织与知识服务。Email：huangyongwen@caas.cn

supporting data-intensive scientific discovery, the new format of knowledge service is quietly taking shape. Text mining as the core of knowledge service is facing challenges under the environment of new knowledge service formats. This paper aims to discuss the development strategies of using text mining to carry out knowledge services in the new environment. [Method/Process]This paper sorts out the technical framework of text mining, and demonstrates that text mining is gradually maturing. Using the research and practice in the field of agriculture as a case study in such areas as information retrieval, intelligent question-answering, information monitoring and knowledge extraction, text mining has shown a good performance in scientific and technological innovation and industrial applications. [Results/Conclusions] This paper puts forward the knowledge service technology development strategies according to China's conditions: (1) constructing a specialized knowledge service system based on text mining technologies, (2) attaching importance to the construction of corpora and basic knowledge bases, and (3) giving priority to implementing the deployment of knowledge service technologies in key areas.

Keywords: text mining; knowledge services; information extraction; knowledge organization

1 引 言

开放科学大背景下，开放出版及开放获取运动的大潮推动各类知识资源及服务的开放共享化，人们可利用的开放信息资源和语料逐步增多。同时，随着人工智能技术，特别是深度学习技术不断取得突破性进展，文本挖掘技术已经成为科技文献资源开发利用的核心驱动力，以文本挖掘为核心的知识服务技术体系已经基本完善，全新的数据密集型科学发现的科技创新生态正悄然形成，而支撑新生态的知识服务呈现出了新业态，并在积极适应新的知识生态环境。

知识服务新业态表现在以下 3 个方面：①面向专门知识发现及知识服务需求，以问题解答为导向的人机交互式迭代过程，新的知识服务需要建立针对具体领域问题的专门知识服务系统；②以知识服务技术、模型、算法、工具、系统为支撑，融合知识组织与认知计算，嵌入各种计量分析、演化分析、可视化分析、协同推理在内的认知搜索、知识发现、智能推荐及智能问答服务；③新的知识服务系统和工具不是独立存在的，它们将积极适应新型的数据密集型科学发现的知识生态环境。

文本挖掘作为知识服务技术的核心，其在知识服务新业态环境下面临新的挑战。尽管国内学界、业界一直对文本挖掘领域保持着深入研究、持续追踪，但从战略出发，中国在新的科技创新业态下仍面临自主可控性的安全挑战。具体表现在支持科技创新的文本挖掘其模型、算法、工具多数非自主产权，支撑文本挖掘技术的通用语料库、基础知识库等战略基础资源和设施也未掌握在国人手中。笔者以文本挖掘技术为中心，梳理其技术框架，结合农业领域应用描绘其在知识服务新业态下的发展方向，最终结合国情实际提出文本挖掘为主的知识服务技术的发展策略。

2 文本挖掘技术框架

主流观点认为[1-5]数据挖掘是知识发现的一个步骤，其指从给定数据中抽取出隐含的、以前未知的、潜在有用的知识的过程。从广义的数据挖掘范围看，文本挖掘可看作是数据挖掘的一类，或是数据挖掘在文本数据中的应用[1,2]。因而文本挖掘又称为文本知识发现，是指从自由非结构化文本数据中发现、挖掘知识的过程。整体来看，目前文本挖掘研究主要涉及三大热门方向：①以信息检索、文本摘要、意见挖掘与情感分析为代表的文本知识发现的主要模式研究；②文本挖掘相关的技术方法研究，如自然语言处理、文本信息

DOI：10.13998/j.cnki.issn1002-1248.20-1197

抽取、无监督学习、有监督学习、文本挖掘的概率方法以及针对文本流和社交媒体的挖掘；③应用研究。由于生物医学领域本身资源的开放性，以及生物医学领域本身具有非常丰富的语义关系，文本挖掘率先在生物医学领域中得到应用，此外近几年在农业领域[6]、金融领域[7]等也有大量的应用案例。

最早的文本挖掘模型是 1998 年 FELDMAN 提出的文本知识发现框架[8]，随后多位研究者总结了不同的文本挖掘通用模型。随着对文本挖掘技术研究的深入，学者们又提出了针对具体问题的多个领域文本挖掘模型。相关文本挖掘模型及流程研究对比如表 1 所示。

整体来看，文本挖掘的整个技术流程有多个关键技术节点不可缺失，即文本挖掘至少包括预处理、文本表示和编码、文本分类或聚类、信息抽取这 4 部分内容。下文对这 4 个技术点进行梳理总结。

2.1 文本预处理

自由文本的非结构化特性决定其挖掘模式不同于结构化数据，因此需要对文档或文本数据实施预处理。预处理首先要分析文本结构及内容，借助工具使其转变成纯文本内容，消除格式差异。例如对网页文档去除各种 HTML 标记、脚本，将 PDF 文档转换格式输出为 TXT 文档。随后对纯文本实施分词、过滤和归一处理。分词根据不同语言文本有所区别。英文文本内分词包括去除空格、标记、标点等，将语句还原成词和短语；中文文本没有固定分隔符，分词相对复杂，有基于规则、基于统计和基于理解的分词方法[16]。过滤即构建停用词表把停用词、半停用词过滤掉。归一，又称为词形还原，是指对一个词不同的时态表现形式实施归一，其中词干提取法是英文文本挖掘内应用最广的归一方法，通过词干提取完成文本数据归一。

2.2 文本表示和编码

文本表示和编码，即数字知识表示，该步骤将自然语言文本变成计算机可处理的数字知识表示模式。现有的文本处理或挖掘研究大都基于离散的词表示为基础的文档表示模型，尽管有研究者提出更加复杂的概念图模型[17]或概念解析文本表示模型[18]，但由于领域概念网络构建的复杂性，这类表示并未成为主流。词表示分为布尔逻辑模型、词袋模型、*N*-gram 模型等方法。早期采用布尔逻辑二值表示法[19]，利用 0 和 1 表示文档内是否出现某个词，以帮助快速检索，但结果缺乏相关性特征。*N*-gram 模型是解决不同语言文本词切分不一致问题而产生的词表示方法，主要应用在中

表 1　文本挖掘模型及流程对比

Table 1　Comparison of text mining models and workflows

提出者	文本挖掘模型及流程
FELDMAN[8]	①输入关键词标注文本集合及关键词层级结构；②标注文档实施挖掘，并通过知识表示输出给用户
TAN[9]	①文本精炼，将非结构化文本转换为结构化中间表示（文本中间表示或概念中间表示）；②知识提取，从中间表示挖掘知识模式
周雪忠等[10]	①文本的预处理；②文本模型表示；③信息或文本特征属性抽取；④文本分类和聚类；⑤结果集的数据挖掘
谌志群等[1]	①预处理，文本数据的选择、清洗、分类、特征提取等；②索引与存储；③中间表示分析，聚类、趋势分析、关联规则发现等；④后处理：知识的评价与取舍、知识的解释与知识的可视化表达
SHILPA 等[11]	①非结构化文本信息采集；②非结构化内容转化为结构化数据；③定义结构化数据模式；④分析模式；⑤抽取知识存入数据库
INZALKAR 等[12]	①文档采集；②预处理，分词、停用词、词干归一；③文本转换；④特征选择；⑤数据挖掘或模式选择；⑥评估
CHIBELUSHI 等[13]	①预处理，文本转录、文本表示；②建模，句法分析、语义标签、词汇链、基于聚类关联的模式检测；③评估，验证、最终模式提取
VISHAL 等[14]	①文档采集；②文本检索及预处理；③文本分析：信息抽取、聚类、摘要；④知识抽取
薛为民等[15]	①文本预处理；②特征提取及约减；③学习与知识模式提取；④知识模式评价

文文本表示中。词袋模型（Bag of Words）是最常见的文本表示方式方法。它在二值表示法基础上，将所有词语装进一个袋子，计算每个单词出现的次数，一段文字或一个文档即可表示为 N 维的向量。文本挖掘需要对词袋模型的维度实施降维，研究者提出信息增益 - 互信息 - 交叉熵[20]、主成分分析[21]、线性判别分析[22]、潜在语义索引[23]（LSI）、概率潜在语义索引（PLSA）[24]及主题模型[25]等不同的降维方法，目前后 3 种方法较为常用。

在词袋模型基础上，文档表示可以采用向量空间模型（VSM）、概率模型[26]和推理网络模型[27]等。其中，向量空间模型是使用最为广泛、比较成熟的文档表示模型。TF-IDF 是空间向量模型中用于特征权重计算的常见方法，有良好的性能表现。TF-IDF 基于词频和逆文档频率有效表示文档，其中 IDF 逆文档频率可以过滤掉文档中的高频通用词。其他的特征权重计算方式还有基于随机投影 Gram-Schmidt 的正交化法[28]、卡方法[29]、拉普拉斯分值法[30]、互信息方法[31]等。

传统的向量空间模型是一个高维的稀疏向量，并且无法解释不同词语之间的关系问题。在神经网络模型支持下，MIKOLOV 等[32,33]使用连续词袋法 CBOW（Continuous Bag-of-Words）和 Skip-Gram 两个模型，通过上下文内容来描绘一个词的表示形式，得到可以表示语义相关性的低维稠密向量，这种文本表示称为分布式词嵌入表示。在此基础上产生了一系列的词向量表示模型，例如 Paragraph Vector 模型[34]、Skip-Thought Vectors 模型[35]、Conv/LSTM-GRNN 模型[36]、Hierarchical Attention Networks（HAN）模型[37]等。

2.3　文本分类 & 聚类

文本分类和聚类是对文本实施浅层挖掘，识别分类信息。信息检索就是以文本分类和聚类结果为基础的文本挖掘应用。文本分类主要基于 3 类模型：逻辑模型（如决策树）、概率模型（如朴素贝叶斯）、几何模型（如支持向量机）。它们的共同特点是：预先有一个知识分类框架或者知识分类的规则，然后按照该框架和规则对每一篇文档或每一段文本逐一地进行处理和分类。邻近分类器和神经网络算法是文本分类任务中比较成熟和流行的方法。

文本聚类是在没有预先定义知识框架、规则和类别的情况下，自动产生文本分类的过程。文本聚类主要有以下 3 个方法（图 1）：①层次聚类法[38]，分自顶向下、自下向上两大类；②分区聚类，典型例子是 K 均值聚类[39]，即围绕某一篇文档，将与它语义相似度最近的集合分为一类，从而通过聚类形成分类；③概率聚类和主题模型，包括概率潜在语义分析模型（PLSA）和隐含狄利克雷分布（LDA）。其中，LDA 相关研究很多，产生了监督 LDA（sLDA）、分层 LDA（hLDA）、分层弹球分配模型（HPAM）等模型变种。主题模型应用很广，例如采用基于 LDA 的本体主题模型进行自动主题标注和语义标注[40]，采用基于知识的主体模型进行上下文感知的推荐[41]，以及基于 LDA 为实体消除歧义定义更复杂的主题模型[42]等。

2.4　信息抽取

信息抽取是文本挖掘最核心的技术，负责从文本数据中抽取出结构化的文本信息并获得知识初始模式。它主要包括两方面内容，一是命名实体识别，二是关系抽取。

命名实体识别有多种方法。①基于词典的方法。如 AKHONDI 等在 ChEBI 和 HMDB 化学词表的基础上，基于 LeadMine 工具对化学物进行语法识别和抽取[43]。②基于预定规则的方法。根据预定义的语法、句法规则（人工总结[44]、基于启发式的规则学习[45]、或机器学习归纳）来抽取文档内实体。③基于统计的机器学习方法。该方法从标注过的训练文集中，让机器归纳实体识别的模式，然后基于模式在不同算法下识别新实体。

机器学习算法模型可分为 3 类：①基于分类的算法，如朴素贝叶斯、支持向量机。②基于序列的方法，如隐马尔科夫模型、条件随机场以及最大熵马尔科夫模型，代表系统包括 MnM[46]、Amilcare[47]、BioTagger-GM[48]等。③混合方法。④基于本体的实体识别方法。细分为本体构建和本体扩展两种[49]。前者识别本体中

DOI：10.13998/j.cnki.issn1002-1248.20-1197

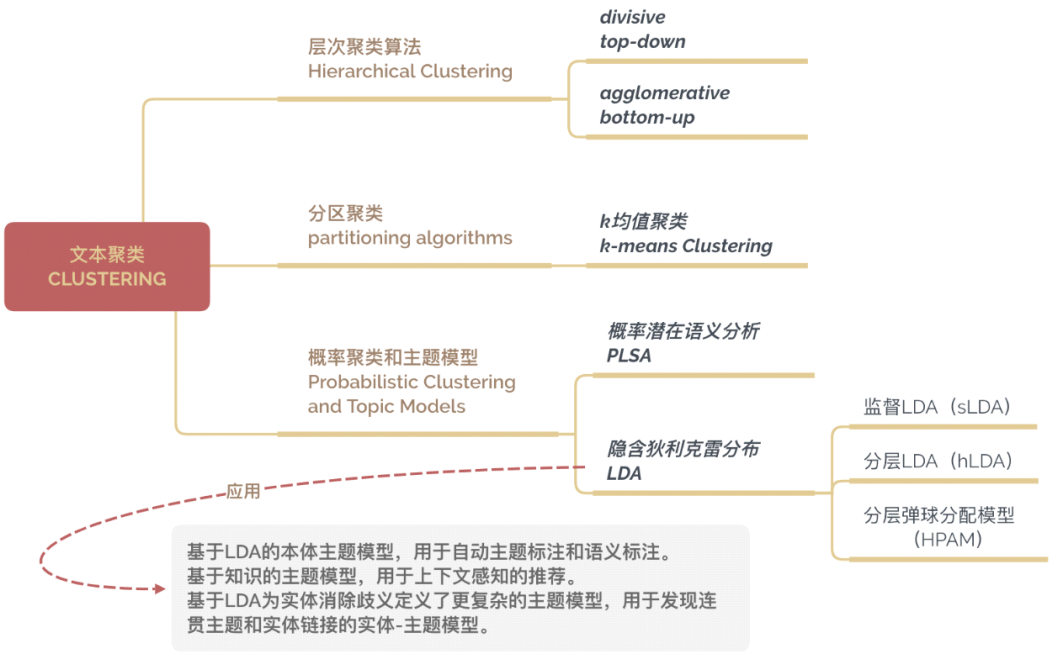

图 1　文本聚类主要方法及应用

Fig. 1　Main methods and applications of text clustering

的概念和属性，基于种子概念（领域中常见术语）和模式学习来扩展更多的概念，如 Text-To-Onto；后者则偏重于实例和属性值层次，借助本体（如叙词表）中的实例及同义词环来识别实体，如 PANKOW、OntoSyphon、Kylin、SOBA 等。⑤基于深度学习的方法。自 2018 年底谷歌发布 BERT 以后，基于 BERT 的信息抽取受到广泛关注，提出了诸多基于 BERT 的改进模型和衍生模型，如华盛顿大学提出的 RoBERTa，清华提出的 ERNIE，北京大学、腾讯和北京师范大学共同提出的 K-BERT、哈尔滨工业大学提出的 BERT-WWM 等。

关系抽取相比实体抽取更为复杂，通常要借助句法规则、上下文内容来发现关系。关系抽取方法大致分为 3 种，即基于规则、基于共现和基于分类的关系抽取。

基于规则的关系抽取借助自然语言处理研究中的句法分析和语义分析工具，基于预定义的模式和特定语法关系匹配规则对语句关系实施模式匹配。例如为获取生物分子之间的绑定和制约关系，定义动词“bind”关系模板[50]，编写动词“inhibit”模板[51]。ONO 等提出基于模式的系统，使用简单词的人工编码规则和标注词性的模式，抽取生物医学文献摘要中的特殊种类蛋白质间的交互关系[52]；PARK 等提出基于可组合分类语法的深层分析器，通过定位动词，扫描动词左右部文本，获得文本语法成分[53]；TEMKIN 等基于上下文无关文法和词典分析程序来抽取基因和蛋白质间的交互关系[54]；SEMREP[55]基于统一医学语言系统（UMLS）利用指示规则（Indicator Rules）抽取生物文献语句中的语义谓项。

基于共现原理的关系抽取的基本原理是如果两个实体在同一个语句、段落、文章中出现时，那么两者必然存在某种关系。STAPLEY 等借助共现方法在 Medline 记录中检测基因名间的相互关系[56]。

简单同现提取的关系类型通常是未知的，通过应用一定的文本分类技术可以支持特定实体关系的提取，这是基于分类的关系抽取方法。CRAVEN 等采用贝叶斯分类器来求解同一语句中两个及以上实体间是否存在交互关系的概率[57]；DONALDSON 等利用支持向量机来抽取蛋白质相互作用关系[58]；LIU 等同样利用支持向量机分类方法，结合递归算法来抽取生物实体间的事件[59]。机器学习的方法免去了人工建立模式或者规

则所需的繁重努力，通过对一个训练集的学习自动建立分类模型来判定蛋白质之间的交互关系[60]。

3　文本挖掘在农业领域知识服务中的应用

梳理总结文本挖掘技术框架可以发现，文本挖掘技术正日渐成熟并逐步见诸领域实用。农业领域文本挖掘的热点主要分布在信息检索、信息抽取和情感分析 3 个方面。其中，信息检索研究不仅关注通常意义上的检索，还包括研究农业问答技术和文本分类。信息抽取研究最为热门，除信息抽取策略外，研究还涉及命名实体检测、监视、知识抽取、食品价格预测和农场管理等内容。文本情感分析和意见挖掘在农业领域中研究相对较弱，最不热门，主要研究方向是预测害虫的严重性、未来的食品价格以及民意挖掘。笔者结合案例分析文本挖掘在农业信息检索、智能问答、信息监测和知识抽取等方面应用。

3.1　农业信息检索服务

文本挖掘很早便应用于农业信息检索。20 世纪 60 年代，农民就可以在计算机系统内利用信息检索来识别农业文档，并帮助决策[61]。近年来，随着文本挖掘技术持续深入，传统农业信息检索呈现以下 3 方面的趋势：①农业信息检索趋向基于本体推荐语义相关术语来优化查询，或者使用“信息链接”技术作为扩展关键字搜索策略的一部分；②TF-IDF 模型在信息检索系统内广泛应用，但其无法准确表达用户查询意图，研究者倾向于借助语义或关系抽取解决该问题；③更多农业文本分类使用不预设分类的无监督学习聚类方法，并对文本聚类结果实施语义增强后，形成基础知识库，为用户提供语义更精准的信息搜索服务。

文本挖掘技术结合本体应用可以有效提高信息检索系统的发现效率。基于本体的文本挖掘能够扩展检索系统的关键词搜索策略，还可以帮助系统理解用户的检索意图，确保查询词在正确的上下文中使用，从而提高信息检索的召回率和准确率。例如，本体的类包括植物作物名称、作物描述、花期、施肥、虫害等。因此在信息检索中，由于使用词频和逆文档概率通常会忽略用户查询术语的意图，当嵌入本体之后，可以借助本体的语义关系确保查询词在正确的上下文中使用，帮助消除词语的歧义，从而有效地提高检索系统的召回率和准确性。

文本挖掘技术改进信息检索系统的案例有 PADI-Web[62]、CyberBrain 等。PADI-Web 是法国开发的一个针对非洲猪瘟、禽流感、蓝舌病等动物流行性疾病的语义搜索引擎，它的核心组件采用基于规则的信息抽取和数据挖掘技术，通过文档向量和数据融合的方法自动从 Google 新闻中收集、处理和提取英语流行病学信息，如发病新闻报道的位置、时间和主题特征（疾病宿主和疾病特征，其中疾病特征包括疾病名称、爆发病例数等），并将挖掘后形成的数据和知识提供给动物卫生局。CyberBrain 是由泰国国家电子和计算机研究中心研发的关于农业的知识服务系统，需求驱动或面向实用的本体用于从多个异构源中聚合信息，为用户提供最能满足他们需求的相关信息。该系统开发了基于面向任务本体的抽取引擎，用于从文档中提取相关信息，并将其重新组织成定义结构格式。CyberBrain 利用语义搜索技术和 PMM 模型（Problem-huMan-Method Model）实现知识搜索。基于本体和本体推理来获取、抽取和整合知识，生成的 PMM 包括疾病问题识别、能够解决该疾病问题的人类专家，以及以纠正和预防方式解决该疾病问题的方法。该系统主要面向 4 类用户，有信息需求的农民、追踪相关研究的研究者、有经营需求的中小型企业和政府智能指挥中心。

3.2　农业智能问答服务

农业智能问答系统是农业领域中文本挖掘最热门的应用，问答系统提供一般搜索引擎无法提供的农业领域的响应内容，面向具体知识问答，如农业实用技术自动问答系统[63]、AGRI-QAS 问答系统[64]等。智能问答服务通常使用本体或潜在语义索引方法辅助信息检索过程。农业领域有大量的本体可供智能问答系统使用，如 Agrovoc、中国农业主题词表、THESAGRO 等。

DOI：10.13998/j.cnki.issn1002-1248.20-1197

本体作为知识库，可以为智能问答系统提供关键词扩展，还可以构建基于本体的语料库。潜在语义索引方法将搜索词归纳为主题，然后进行主题发现，反馈主题匹配文献，提高检索精准度。

KAWAMURA 搭建了基于农业开放关联数据的植物信息问答系统[65]，能自动回答植物花期、施肥等信息。该系统使用句子级三元组（主题、动词、对象）对信息建模，主题是植物名称。问答系统知识库由预设资源和 Web 抽取信息构成，系统自动解析用户查询所用的自然语言，从句子中抽取三元组并映射为 SPARQL 查询。基于查询主题从知识库扩展动词，通过动词再扩展它的对象，进一步校准、消除用户检索过程中的语义歧义。系统设置反馈模式，向用户显示排名前三的动词，用户从中选择正确或最接近的答案，反馈结果被存储并用于完善进一步的搜索。

3.3 农业信息抽取和监测服务

信息抽取是文本挖掘的核心内容，农业领域的信息抽取涉和关键技术有基于本体的信息抽取、监督学习、无监督学习、规则发现以及半监督学习[6]。其中，基于本体的信息提取和监督学习是最常见的技术。基于本体的信息抽取主要用于命名实体识别。在实体识别中，本体用于标注训练实例，标注通常基于规则，而实例用于后续机器学习分类器的训练，识别后的命名实体可用于检测、知识抽取等任务。监督学习[66,67]方式是农业领域信息抽取应用较多的机器学习模式，基于训练数据分类学习可以取得不错的抽取效果。

农业信息抽取在食品价格预测、监测、农场管理、农业知识提取等实用领域应用广泛。食品价格预测通过对短时效文本（如推特、新闻）实施挖掘，抽取价格内容或抽取事件信息来预测短期内特定食品的价格走向[68,69]。农场管理则利用文本挖掘帮助农场相关管理决策，例如种植、收获、处理、干燥和存储[70]。

监测是农业信息抽取中热门的研究方向。通过挖掘网络文本信息，可以推断某些农业现象的演变。PADI-Web 应用信息抽取技术帮助动物疾病监测。首先检索目标相关的语料，并实施人工标注，标注内容元素包括位置、日期、病例名称、宿主和病例数量等，标注内容经过领域专家评估。然后，基于人工标注语料，借助支持向量机、深度学习等机器学习算法自动发现规则。经过前期标注和机器学习后，新输入相关文档能基于支持机器学习所建立的分类模式和知识模型实现自动、无监督的流行病学元素抽取。系统验证结果显示，其不同对象信息抽取结果介于 80% 至 96% 之间。基于准确的抽取结果结合疾病爆发模式，能发现和预警疫情。事实证明该系统会提前两到三周向世界卫生组织预警。

3.4 农业知识抽取

知识抽取是从文本中发现、抽取知识模式，完成知识建模的过程。例如计算机从海量的科技文献中总结出芽孢杆菌的调控网络是复杂知识抽取的过程。简单的知识抽取可以基于概念术语进行抽取，并结合规则来抽取相关的关系，例如构建作物与土壤的关系[71]、食物和健康间关系[72]等。如果融合半自动监测工具、多源术语抽取、语义标注、语义搜索引擎、关系抽取等工具和过程则能实现复杂的知识自动抽取。VALSAMOU[73]设计的 Alvis 知识抽取环境尝试对复杂知识实施自动抽取，Alvis 知识抽取环境如图 2 所示。首先通过 AlvisCrawler 半自动的获取全文文献，随后借助相关集成工具（如基于本体的术语抽取分类工具 ToMap，抽取关系的 AlvisRE 工具，抽取蛋白质、基因实体的 RenBio 工具）对文本语料实施分类、实体识别、术语抽取、关系抽取，最终抽取得到种子发育过程中调节网络知识，以及凝练 10 种调控关系，并提供语义知识搜索服务（AlvisIR）和在线标注生成新语料的服务（AlvisAE）。

总体而言，文本挖掘技术在农业领域应用前景非常广阔。现有研究表明，在领域知识组织体系（如本体等）和人工标注语料的支持基础上，以信息抽取为主体的文本挖掘技术可以实现较高质量的知识模式抽取并支持语义搜索、问答服务、信息监测以及预测和决策支持服务等广泛的知识服务应用。

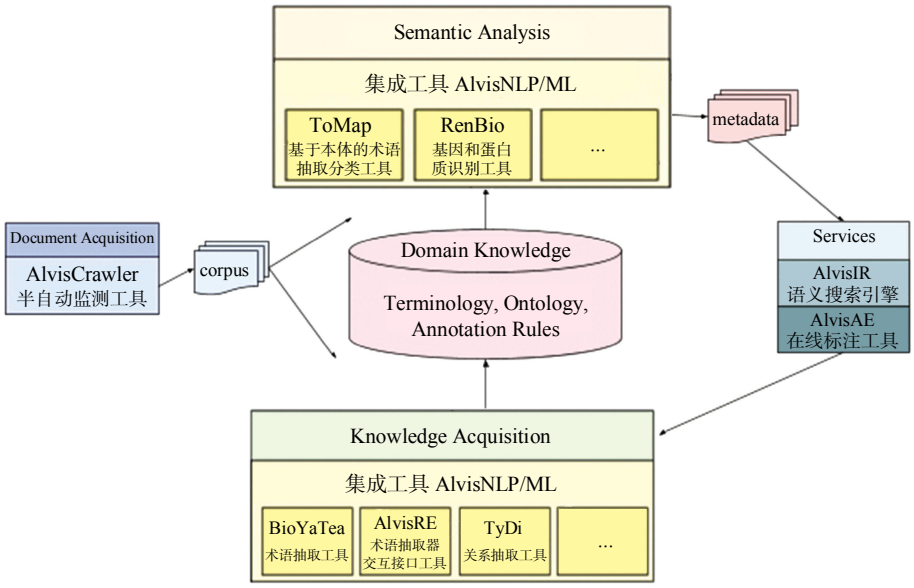

图 2　Alvis 知识抽取环境

Fig. 2　The knowledge extraction environment of the Alvis system

4　文本挖掘技术应用在知识服务中的发展策略

综合技术框架与领域应用发现，文本挖掘技术已广泛地应用在知识服务系统中。基于科技文献文本挖掘的知识服务应用，既面向科技创新，如基于信息抽取的拟南芥种子发育调控网络构建；也面向产业应用，如传染病监测、短期和大宗商品市场价格预测、农场管理，以及学术观点挖掘和情感分析。为使文本挖掘技术能在未来知识服务体系中发挥更大作用，笔者提出以下几方面的发展策略。

4.1　基于文本挖掘技术构建专门知识服务系统

当前知识发现已进入深度问题解决和个性化服务阶段，知识服务呈现专门化、智能化和交互性新业态，因此传统面向通用问题解决的知识服务系统需要改革。新型知识服务系统应针对专业问题和科学家的个人需求，知识服务系统中的文献标注、语料库构建及优化、知识组织体系嵌入以及机器学习算法与策略均要基于具体领域特征来实施，并进行个性化迭代和验证。

专业化、专门化的知识服务系统需要适应数据密集型科学发现的新型知识生态环境，以文本挖掘技术为核心的知识服务技术在其中能发挥重要作用。具体而言：①知识服务系统底层需要融合多源异构数据，并在语义知识组织框架帮助下建设融合用户问题、自动学习与进化更新的大规模语义知识库。文本挖掘中的信息抽取能协助处理大规模数据，而知识抽取能帮助机器总结发现知识模式，扩展知识库实例。②在语义知识组织体系的基础上，利用深度学习、迁移学习等方法，突破语义智能检索、检索结果的多重因子排序、智能推荐计算、潜在关系挖掘、领域自动综述等关键技术，构建文本挖掘和认知计算引擎。③基于大数据与微服务架构提供解决不同问题的应用组件，例如语义标注、语义搜索、智能推荐、智能问答等，以便研究人员根据自身需求实施数据挖掘和关联。

4.2　重视语料库和基础知识库建设

随着科学研究不断深入，研究对象的颗粒度、数量和关联复杂性愈发微观、海量和高维。为支撑科研人员快速发现知识和认知计算，语义知识库成为战略基础资源和设施。语义知识库是文本挖掘技术和知识组织融合产生的结果，同时语义知识库能给新的命名实体识别、语义相似度计算、信息抽取等文本挖掘技

DOI：10.13998/j.cnki.issn1002-1248.20-1197

术提供一定的语义数据支撑。

诚然，国家一直重视学科公共科学数据中心建设，但纵观科学研究领域，许多重要的基础知识库受国外控制，知识库资源的访问和获取无法得到完全受信的保证。如生物和医学领域不可或缺的 NT/NR 蛋白质 / 核酸数据库、UniProt 蛋白质数据库、Genbank 基因数据库等，化学领域的 SciFinder、ChemSpider 等数据库，其知识产权、数据访问、使用许可均受国外控制，在目前复杂多变的国际形势下，继续坚持自主建设基础科学知识库变得尤为重要。此外，从基础科学数据中心或平台转成支撑新型知识服务所需的基础语义知识库还有许多工作要完成。例如完善本体知识模型和构建优质的语料库。

本体知识模型非常重要，它不仅充当基础知识库中语义类别和关联的框架支撑，同时它在整个语义知识服务的检索到问答过程中发挥语义归一、语义消歧的重要作用。因此需要以科技文献为来源和核心对象，构建不同领域知识单元语义描述模型和知识属性体系，采用各类知识单元语义关联的知识组织方法，建设受控词表系统、领域本体、知识图谱等。

语料库为知识服务技术提供基础数据支撑，优质的语料库能够在信息抽取、关系抽取等文本挖掘任务中发挥巨大作用。例如上文中提到的 PADI-Web 系统，其人工构建的高质量小规模语料库保障了实体识别、关系识别精准度。可以利用文献计量学方法构建高质量小规模的初始语料集。以传统知识组织系统为基础采集并组织原始文献，基于文献质量评价体系优选抽取其中的高质量文献，形成初始种子语料集，以此为基础借助机器学习和人工筛选，生成新的更大规模的高质量语料集[74]。

4.3 在重点领域开展和部署文本挖掘技术的应用

建议优先在生命科学、医学与健康、微生物学及交叉领域、农业科学、资源环境、化学及交叉领域、边缘交叉领域等重点领域部署和开展基于文本挖掘技术为核心的知识服务。这些领域是当前科技创新活动非常频繁的领域，科技创新需求旺盛；这些领域的交叉复合使得领域知识复杂、丰富，单一领域知识表达无法全面描述；同时，这些领域在前人学者贡献下已经具备丰富的语义知识基础（如大量领域叙词表或本体、语义标注语料等）。在这些领域内，优先发展语义搜索引擎，构建具有自主知识产权的基础知识库，重点开发文献摘要与综述、知识问答与推理等知识服务应用。

文本挖掘技术在信息检索领域的应用，是知识服务基础和优先的选择。国外相关实践，如 Semantic Scholar、GoPubMed，都以语义搜索引擎为切入点推广知识服务应用。基于文本挖掘技术的语义搜索系统，不仅可以显著提高用户信息检索效率，还可以广泛应用和嵌入于后续复杂知识应用（如检索意图智能理解、领域知识画像和研究侧写、智能知识问答等）。

文献摘要与综述在学科边缘交叉日益加速情境下意义非凡。接触一个新的研究领域意味着需要补充海量相关知识，文本摘要与综述可以快速弥补跨学科研究的知识缺口。它不仅是辅助科研人员快速掌握领域知识的重要服务，也是将结构化知识重新转换为自然语言表述的知识的重要支撑。

问答与推理服务不仅是人机交互中的智能知识服务，更是智联网环境中 M2M（Machine to Machine）智能交互的重要基础。问答与推理服务不是面向科技创新，而是面向产业服务。例如智慧农场中，植物、浇水机器人、采摘机器人之间的会话场景是 M2M 的，需要知识问答和推理为其提供交互的数据基础。

5 结　语

综上所述，以文本挖掘技术为核心的知识服务技术体系日渐成熟，可以实现较高质量的知识模式抽取并支持语义搜索、语料库训练、语义知识库构建和问答服务、信息监测和预测、决策支持服务等广泛的知识服务应用，在农业等诸多领域具有可操作性。与此同时，我们也看到知识服务新业务和科技创新自主安全环境改变对新型知识服务系统提出的新挑战。美国

国家医学图书馆、英国大英图书馆等国外重要信息机构在其未来规划中提出，将继续紧密依靠人工智能、数据分析、文本挖掘等信息技术的发展，加强基于科技文献和科学数据的计算分析、知识关联等技术创新，重视未来开放科学环境中的知识服务，建设学科领域的语义知识库、提升知识发现能力、创新知识服务模式。对此，我们也应在重点领域加快核心知识服务技术的部署，重视基础知识库建设，并融合知识组织、文本挖掘、认知计算、可视化交互等技术构建专门的知识服务系统。

参考文献：

[1] 谌志群, 张国煊. 文本挖掘研究进展[J]. 模式识别与人工智能, 2005, 18(1): 65-74.
CHEN Z Q, ZHANG G X. A survey of text mining[J]. Pattern recognition and artificial intelligence, 2005, 18(1): 65-74.

[2] ALLAHYARI M, POURIYEH S, ASSEFI M, et al. A brief survey of text mining: Classification, clustering and extraction techniques[C]// KDD Bigdas, Halifax, Canada, 2017.

[3] 化柏林. 数据挖掘与知识发现关系探析[J]. 情报理论与实践, 2008, 31(4): 507-510.
HUA B L. Probe into the relationship between data mining and knowledge discovery[J]. Information studies: theory & application, 2008, 31(4): 507-510.

[4] USAMA F, GREGORY P-S, PADHRAIC S, et al. Knowledge discovery and data mining: Towards a unifying framework[C]//KDD'96: Proceedings of the second international conference on knowledge discovery and data mining, 1996: 82-88.

[5] FRAWLEY W J, PIATETSKY-SHAPIRO G, MATHEUS C J. Knowledge discovery in databases: An overview[J]. AI magazine, 1992, 13(3): 57-70.

[6] DRURY B M, ROCHE M. A survey of the applications of text mining for agriculture[J]. Computers and electronics in agriculture, 2019, 163: 104864.

[7] KUMAR B S, RAVI V. A survey of the applications of text mining in financial domain[J]. Knowledge based systems, 2016, 114(15): 128-147.

[8] FELDMAN R, DAGAN I, HIRSH H. Mining text using keyword distributions[J]. Journal of intelligent information systems, 1998, 10(3): 281-300.

[9] TAN A H. Text mining: the state of the art and challenges[J]. Proceedings of the PAKDD workshop on knowledge discovery from advanced databases, 1999: 65-70.

[10] 周雪忠, 吴朝晖. 文本知识发现: 基于信息抽取的文本挖掘[J]. 计算机科学, 2003, 30(1): 63-66.
ZHOU X Z, WU C H. Knowledge discovery in text: A survey[J]. Computer science, 2003, 30(1): 63-66.

[11] SHILPA D, PEERZADA H A. Text mining: Techniques and its application [J/OL]. International journal of engineering & technology innovations, 2014: 22-25.

[12] INZALKAR S M, SHARMA J. A survey on text mining-techniques and application[J/OL]. International journal of research in science & engineering, 2014: 488-495.

[13] CHIBELUSHI C, SHARP B, SALTER A. A text mining approach to tracking elements of decision making: a pilot study[C]//International workshop on natural language understanding & cognitive science, DBLP, 2004.

[14] VISHAL G, LEHAL G S. A survey of text mining techniques and applications[J]. Journal of emerging technologies in web intelligence, 2009, 1(1): 60-76.

[15] 薛为民, 陆玉昌. 文本挖掘技术研究[J]. 北京联合大学学报(自然科学版), 2005, 19(4): 59-63.
XUE W M, LU Y C. Research on text data mining[J]. Journal of Beijing union university (natural sciences), 2005, 19(4): 59-63.

[16] 胡静, 蒋外文, 朱华. Web 文本挖掘中数据预处理技术研究[J]. 现代计算机(专业版), 3: 48-51.
HU J, JIANG W W, ZHU H. Research on data preprocessing techniques in web text mining[J]. Modern computer, 3: 48-51.

[17] Manuel M-Y-G, Gelbukh A, Aurelio L-L. Text mining at detail level using conceptual graphs[M]// Conceptual structures: Integration and interfaces, Springer Berlin Heidelberg, 2002.

[18] BING L, JIANG S, LAM W, et al. Adaptive concept resolution for document representation and its applications in text mining[J]. Knowledge-Based systems, 2015, 74(1): 1-13.

DOI：10.13998/j.cnki.issn1002-1248.20-1197

[19] ARMSTRONG R. WebWacher: a learning apprentice for the world wide web [C]//AAAI spring symposium on information gathering from heterogeneous, distributed environments, 1995.

[20] MLADENIC D, GROBELNIK M. Feature selection for unbalanced class distribution and naive bayes [C]//Proceedings of the sixteenth international conference on machine learning (ICML 1999), Bled, Slovenia, DBLP, 1999.

[21] JOLLIFFE I T. Principal component analysis[J]. Journal of marketing research, 2002, 87(4): 513.

[22] MARTINEZ A M, KAK A C. PCA versus LDA[J]. IEEE transactions on pattern analysis & machine intelligence, 2002, 23(2): 228–233.

[23] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the association for information ENCE & technology, 2010, 41(6): 391–407.

[24] HOFMANN T. Probabilistic latent semantic indexing[C]//International ACM SIGIR conference on research & development in information retrieval, ACM, 1999.

[25] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003, 3: 993–1022.

[26] MANNING C D, RAGHAVAN P, HINRICH S. Introduction to information retrieval[M]. 北京: 人民邮电出版社, 2010.
MANNING C D, RAGHAVAN P, HINRICH S. Introduction to information retrieval[M]. Beijing: Posts & telecom press, 2010.

[27] TURTLE H R, CROFT W B. Inference networks for document retrieval[C]//SIGIR'90, 13th international conference on research and development in information retrieval, Brussels, Belgium, Proceedings, ACM, PUB27, New York, USA, 1990.

[28] WANG D, ZHANG H, LIU R, et al. Unsupervised feature selection through Gram–Schmidt orthogonalization–a word co–occurrence perspective[J]. Neurocomputing, 2016, 173(3):845–854.

[29] YANG Y, PEDERSEN J O. A comparative study on feature selection in text categorization[C]//Fourteenth international conference on machine learning, 1997: 412–420.

[30] BENABDESLENI K, ELGHAZEL H, HINDAWI M. Ensemble constrained Laplacian score for efficient and robust semi–supervised feature selection[J]. Knowledge and information systems, 2016, 49 (3): 1161–1185.

[31] VILA M, BARDERA A, FEIXAS M, et al. Tsallis mutual information for document classification[J]. Entropy, 2011, 13(9): 1694–1707.

[32] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer science, 2013.

[33] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013: 3111–3119.

[34] QUOC V L, MIKOLOV T. Distributed representations of sentences and documents[J]. Computer science, 2014.

[35] KIROS R, ZHU Y, SALAKHUTDINOV R R, et al. Skip–Thought vectors[J]. Advances in neural information processing systems, 2015, 28.

[36] TANG D, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]//Proceedings of the 2015 conference on empirical methods in natural language processing, 2015.

[37] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification [C]//Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies, 2016: 1480–1489.

[38] PETER W. Recent trends in hierarchic document clustering: A critical review[J]. Information processing & management, 1988, 24 (5): 577–597.

[39] KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An efficient k–means clustering algorithm: Analysis and implementation[J]. IEEE transactions on pattern analysis & machine intelligence, 2002, 24 (7): 881–892.

[40] ALLAHYARI M, KOCHUT K. Automatic topic labeling using Ontology–Based topic models[C]//IEEE international conference on machine learning & applications, IEEE, 2016.

[41] ALLAHYARI M, KOCHUT K. Semantic Context–Aware recommendation via topic models leveraging linked open data[C]//In international conference on web information systems engineering, Springer, 2016: 263–277.

[42] PRITHVIRAJ S. Collective context–aware topic models for entity disambiguation [C]//In proceedings of the 21st international conference on world wide Web, ACM, 2012: 729–738.

[43] AKHONDI S A, HETTNE K M, EELKE VAN D H. Recognition of chemical entities: Combining dictionary-based and grammar-based approaches[J]. Journal of cheminformatics, 2015, 7(1).

[44] CIRAVEGNA F, DINGLI A, IRIA J, et al. Multi-Strategy definition of annotation services in MELITA[C]//ISWC 2003 international semantic web conference, 2003: 97-107.

[45] CIRAVEGNA F, CHAPMAN S, DINGLI A, et al. Learning to harvest information for the semantic web[C]//The semantic web: Research and applications, Springer Berlin Heidelberg, 2004: 312-326.

[46] VARGAS-VERA M, MOTTA E, DOMINGUE J, et al. MnM: Ontology driven semi-automatic and automatic support for semantic markup[C]// International conference on knowledge engineering and knowledge management, Springer Berlin Heidelberg, 2002.

[47] CIRAVEGNA F, DINGLI A, PETRELLI D, et al. User-System cooperation in document annotation based on information extraction[C]//Knowledge engineering and knowledge management, ontologies and the semantic web, Siguenza, Spain, 2002.

[48] MANABU T, ZHANGZHI H, WU C H, et al. BioTagger-GM: A gene/protein name recognition system[J]. Journal of the American medical informatics association, 2009(2): 247-255.

[49] 丁培. 科学文献与科学数据细粒度语义关联研究[J]. 图书馆论坛, 2016(7): 24-33.
DING P. Semantic association between scientific data and scientific literature at fine-grained level[J]. Library tribune, 2016(7): 24-33.

[50] RINDFLESCH T C, PHD L H, ARONSON A R. Mining molecular binding terminology from biomedical text[C]//Proceedings of the AMIA-A99 annual symposium, 1999.

[51] PUSTEJOVSKY J, CASTAFIO J, ZHANG J, et al. Robust relational parsing over biomedical literature: Extracting inhibit relations[J]. Pacific symposium on biocomputing pacific symposium on biocomputing, 2002, 16(9): 362-373.

[52] ONO T, HISHIGAKI H, TANIGAMI A, et al. Automatic extraction of information on protein-protein interactions from the biological literature[J]. Bioinformatics, 2001, 17(2): 155-161.

[53] PARK J C, KIM H S, KIM J J. Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar [C]//Proceedings of the pacific symposium on bio computing, Hawaii, USA, 2001: 396-407.

[54] TEMKIN J M, GILDER M R. Extraction of protein interaction information from unstructured text using a context-free grammar[J]. Bioinformatics, 2003, 19(16): 2046-2053.

[55] SemRep[EB/OL]. [2020-05-21]. http://semrep.nlm.nih.gov/.

[56] STAPLEY B, BENOIT G. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in Medline abstracts [C]//Proceedings of the pacific symposium on biocomputing, Hawaii, USA, 2000: 529-540.

[57] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C]//Proceedings of the 7th international conference on intelligent systems for molecular biology. Heidelberg, Germany, 1999: 77-86.

[58] IAN, DONALDSON, JOEL, et al. PreBIND and textomy-mining the biomedical literature for protein-protein interactions using a support vector machine[J]. BMC bioinformatics, 2003: 234-239.

[59] LIU X, BORDES A, GRANDVALET Y. Extracting biomedical events from pairs of text entities[J]. BMC bioinformatics, 2015, 16 (10): S8-S8.

[60] 封二英, 牛耘, 魏欧. 基于大规模文本的蛋白质交互关系自动提取[J]. 计算机应用, 2012(32): 147-150.
FENG E Y, NIU Y, WEI O. Extraction of protein-protein interactions by searching large scale text[J]. Journal of computer applications, 2012(32): 147-150.

[61] EISGRUBER L M. Micro-and macro-analytic potential of agricultural information systems[J]. American journal of agricultural economics, 1967, 49.

[62] ARSEVSKA E, VALENTIN S, RABATEL J, et al. Web monitoring of emerging animal infectious diseases integrated in the French animal health epidemic intelligence system[J]. PLoS ONE, 2018, 13(8).

[63] 王婷, 崔运鹏, 王健, 等. 认知计算及其在农业领域的应用研究[J]. 农业图书情报, 2019, 31(4): 4-18.
WANG T, CUI Y P, WANG J, et al. Cognitive computing and applications in agriculture[J]. Agricultural library and information, 2019, 31(4): 4-18.

[64] GAIKWAD S, ASODEKAR R, GADIA S, et al. AGRI-QAS ques-

DOI：10.13998/j.cnki.issn1002-1248.20-1197

tion-answering system for agriculture domain[C]//International conference on advances in computing. IEEE, 2015: 1474-1478.

[65] KAWAMURA T. Question-Answering for agricultural open data[M]. Transactions on Large-Scale data and Knowledge-Centered systems XVI. Springer Berlin Heidelberg, 2014.

[66] ARSEVSKA E, ROCHE M, HENDRIKX P, et al. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web[J]. Computers & electronics in agriculture, 2016, 123: 104-115.

[67] MALARKODI C S, LEX E, DEVI S L. Named entity recognition for the agricultural domain[J]. Research in computing science, 2016, 117(1): 121-132.

[68] KIM J, CHA M, LEE J G. Nowcasting commodity prices using social media[J]. PeerJ computer ENCE, 2017, 3(262): E126.

[69] SUNANDAN C, ASHWIN V, SRIKANTH J, et al. Predicting socio-economic indicators using news events [C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016: 1455-1464.

[70] LIAO W T, RODRIGUEZ L F, DIESNER J, et al. Improving farm management optimization: Application of text data analysis and semantic networks[C]//2015 ASABE annual international meeting, American society of agricultural and biological engineers, 2015.

[71] CHATTERJEE N, KAUSHIK N, BANSAL B. Inter-Subdomain relation extraction for agriculture domain[J]. IETE technical review, 2019, 36(2): 157-163.

[72] MICHAEL WIEG D K. Towards the detection of reliable food-health relationships[J]. NAACL, 2013.

[73] VALSAMOU D. Information extraction for the seed development regulatory networks of Arabidopsis thaliana[D]. Universite? Paris-Saclay, 2017.

[74] 孙坦, 刘峥, 崔运鹏, 等. 融合知识组织与认知计算的新一代开放知识服务架构探析[J]. 中国图书馆学报, 2019, 45(3): 38-48.
SUN T, LIU Z, CUI Y P, et al. Analysis and design of a new generation of open knowledge service system integrating knowledge organization and cognitive computing[J]. Journal of library science in China, 2019, 45(3): 38-48.