doi:10.3772/j.issn.2095-915x.2019.06.004

农业中文期刊论文信息自动识别与抽取模型构建及实现

开放科学（资源服务）标识码（OSID）

薛欢欢 赵瑞雪 寇远涛 鲜国建

中国农业科学院农业信息研究所 北京 100081

摘要： 面对农业领域丰富的中文期刊论文资源，为实现对农业中文期刊论文文本信息的高效利用，识别与抽取论文中信息已成为一种非常迫切的需求。通过对现有论文信息识别与抽取方法及工具进行调研，确定基于条件随机场算法以及 GROBID 工具进行农业中文期刊论文信息的识别与抽取。本文构建了农业中文期刊论文信息识别与抽取级联模型，并通过数据采集、文本预处理、特征选择、序列标注、特征模板以及模型训练及评估等一系列流程对模型进行实现与应用。实验结果表明，在进行农业中文期刊论文信息识别与抽取时，该模型在论文头信息以及引文信息抽取方面具有较好的效果，在章节标题以及段落信息的识别上仍然存在不足。

关键词： 信息抽取；条件随机场；GROBID；农业期刊论文信息

中图分类号： TP391 G35

Construction and Implementation of Automatic Identification and Extraction Model for Agricultural Chinese Journals

XUE Huanhuan ZHAO Ruixue KOU Yuantao XIAN Guojian

Agricultural Information Institution of CAAS, Beijing 100081, China

Abstract: The growing Chinese journal articles in the agricultural field has become the rich resourcos forming. In order to realize

基金项目： 中国农业科学院科技创新工程“语义知识发现系统建设与应用”（CAAS-ASTIP-2016-AII）。

作者简介： 薛欢欢（1994-），硕士研究生，研究方向：信息资源管理；赵瑞雪（1968-），博士，研究员，博士生导师，信息管理与信息系统、信息资源管理、知识组织与数字图书馆，E-mail：zhaoruixue@caas.cn；寇远涛（1982-），博士，副研究员，硕士生导师，研究方向：数字图书馆理论与技术、信息管理与信息系统；鲜国建（1982-），博士，副研究员，硕士生导师，研究方向：知识组织、关联数据、语义出版、信息系统开发。

the efficient use of the text information of agricultural Chinese journal articles, it has become a very urgent need to identify and extract information from papers. Through the investigation of existing paper information identification and extraction methods and tools, the identification and extraction of agricultural Chinese journal papers based on conditional random field algorithm and GROBID tool are developed. In this paper, the information recognition and extraction cascade model of agricultural Chinese journals is constructed, and the model is realized and applied through a series of processes such as data acquisition, text preprocessing, feature selection, sequence labeling, feature template and model training and evaluation. The experimental results show that in the information identification and extraction of agricultural Chinese journal articles, the model has a good effect on the paper head information and citation information extraction, and there are still some shortcomings in the chapter title and paragraph information identification.

Keywords: Information extraction; conditional random field; GROBID; information of agricultural journal paper

引言

随着期刊资源的开放获取活动的兴起，期刊论文全文数据的批量获取变得越来越容易，进一步推动了基于期刊论文信息的相关研究。期刊论文中的头信息和引文信息可以帮助高效的管理和组织论文，提高用户检索以及获取论文的效率。论文的全文内容中也包含丰富的信息，可以帮助研究人员获取领域研究的热点，在进行主题追踪以及学科热点预测等相关的研究中也具有重要的作用[1]。期刊论文信息碎片化和语义化的描述与组装，可以推动知识发现相关研究，促进语义知识发现系统的建设与应用。然而在进行上述研究时，首先面临的一个问题就是，如何高效、准确的对期刊论文中的信息进行识别与抽取。

农业期刊论文资源作为农业知识的载体，能够实现对农业技术以及研究成果的传播，在促进农业领域的发展中具有极其重要的作用。因此如何实现对农业领域期刊论文的高效利用、提升资源的价值一直是研究的焦点。本文主要以农业领域期刊论文为对象，通过构建农业中文期刊论文信息自动识别与抽取模型，进行论文信息的识别与抽取。

1 研究现状

论文信息识别与抽取的方法主要包括：基于规则的论文信息识别与抽取方法、基于模板的论文信息识别与抽取方法以及基于机器学习的论文信息识别与抽取方法[2]。Wei Wei 等[3]、李朝光等[4]、牛光洁等[5]、Jahongir Azimjonov 等[6]基于规则的方法对论文元数据信息进行了识别与抽取研究。Min- Yuh Day 等[7]，郭志鑫等[8]，黄泽武[9]采用模板匹配的抽取方法对期刊论文的引文元数据进行识别与抽取。欧阳辉等[10]、白光祖等[11]、李雪驹等[12]、黄勇等[13]、方龙等[14]基于机器学习方法分别来对论文元数据信息以及文本结构信息进行识别与抽取。基于规则以及基于模板的方法，在进行论文信息的抽取时具有较高的识别率，但也存在一定的缺陷。在引进新的风格的期刊论文时，需要增

doi:10.3772/j.issn.2095-915x.2019.06.004

添新的规则或模板。随着需要识别与抽取的不同风格的期刊论文增多，规则或模板制定的负担会越来越大，造成系统冗余度升高，抽取效率降低。基于机器学习的论文信息识别与抽取方法是利用机器学习的方法通过训练数据建立样本的输入与输出之间的关系来预测新数据从而实现对论文信息的识别与抽取[15]。相较于前两种，基于机器学习的抽取方法能够更好地完成对大量的异构期刊论文信息的识别与抽取，具有较强的鲁棒性，在抽取不同的风格论文上具有很大的灵活性，是论文信息抽取中最流行的一种方法。

在基于机器学习的论文信息抽取方法中，常见的抽取方法有隐马尔可夫模型(Hidden Markov Model, HMM)、支持向量机(Support Vector Machine, SVM)和条件随机场(Conditional Random Fields, CRF)。三种方法中，CRF 算法在进行论文信息识别与抽取时能有效利用上下文的特征，建立当前观察序列、当前状态、前文观察序列与前文状态之间的关系，从全局考虑，输出最优的概率模型[16]。在进行论文信息识别与抽取时，各个部分的信息不仅与自身的特征有关，还与信息的上下文的特征有关。因此，与 HMM 以及 SVM 方法相比，基于 CRF 方法进行论文信息识别与抽取效果较好，研究人员基于 CRF 算法开展了一系列的论文信息识别与抽取工作。于江德等[17]使用 CRF 算法对中文期刊论文进行信息抽取，抽取的主要信息包括论文的头信息和引文信息，并与 HMM 抽取效果对比，实验结果表明，基于 CRF 的抽取效果较于 HMM 更为准确。陆伟等[18]采用基于 CRF 方法，使用 ParsCit 工具进行论文章节标题信息的识别，在构建 CRF 抽取模型的过程中，针对章节标题的特点增加了自定义词表作为一个特征项。实验结果显示，与 ParsCit 相比，该方法对论文章节标题信息的识别有了一定的提升。王东波等[19]通过对双向长短时记忆神经网络、支持向量机和条件随机场三种模型进行对比，确定使用CRF进行论文结构信息包括章节标题、章节内容信息进行识别，实验结果表明根据章节标题和内容词汇特点加入特征项的抽取模型的 F1 值达到了 92.88%。由于 CRF 算法在论文信息识别与抽取研究中的优越性能，本文中采用 CRF 算法来进行农业中文期刊论文进行信息的识别与抽取。

随着论文信息抽取相关研究的不断发展，涌现出了一批论文信息抽取的相关工具，例如 ParsCit、Reference Tagger、Science Parse、free_cite 以及 Gene Ration Of Bibliographic Data（GROBID）等。Tkaczyk D 等，Mario Lipinski 等[20-21]在相关研究中分别对开源的期刊论文信息识别与抽取工具在头信息和引文信息的识别与抽取效果进行了对比分析，得出基于 CRF 的论文信息识别与抽取工具 GROBID 的抽取效果最好。GROBID 工具通过对 PDF 格式论文预处理得到论文文本信息和格式信息，采用精确设计的 CRF 模型来进行论文信息的识别与抽取[22]。由于 GROBID 良好的性能以及高效的抽取效果，本文采用基于开源工具 GROBID 进行农业中文期刊论文的信息识别与抽取。

2 农业中文期刊论文信息识别与抽取

农业中文期刊论文信息识别与抽取流程具

体如图 1 所示。首先基于 GROBID 工具中论文信息识别与抽取的级联模型构建农业中文期刊论文信息识别与抽取模型，包括 segmentation 模型、header 模型、reference-segmentation 模型、citation 模型以及 fulltext 模型。其次，是基于数据采集、文本预处理、特征选择、序列标注、特征模板以及模型的训练与评估一系列的流程，生成农业中文期刊论文信息识别与抽取模型。最后，根据生成的农业期刊论文信息识别与抽取的模型对期刊论文信息进行识别标注，并以符合 TEI 标准的 xml 格式对识别的信息组装输出。

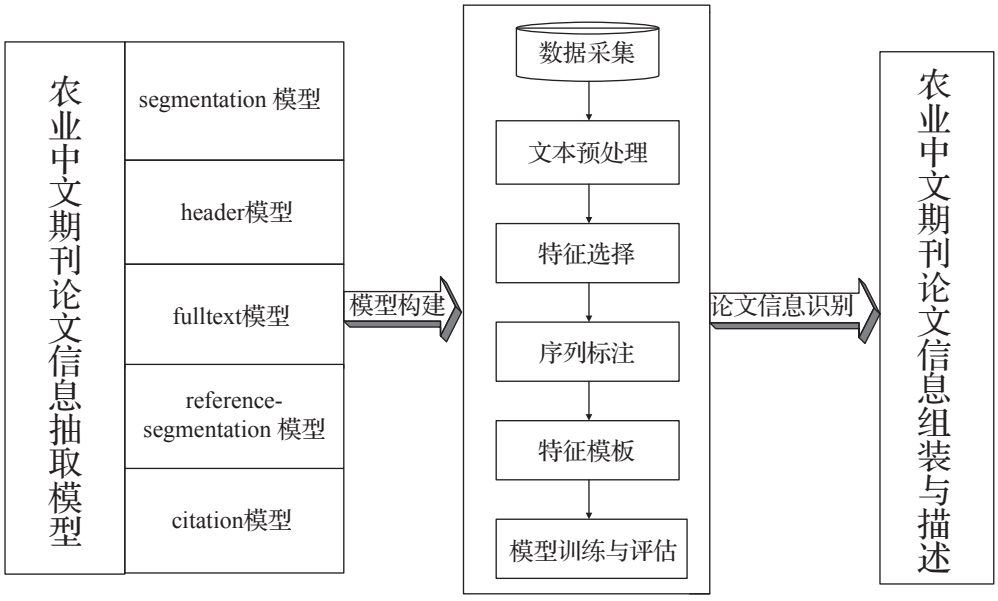

图 **1** 农业中文期刊论文信息识别与抽取流程

2.1 农业中文期刊论文信息识别与抽取级联模型

本文通过对现有的期刊论文信息抽取工具的调研，确定基于 GROBID 工具的级联抽取模型，构建农业中文期刊论文信息识别与抽取模型。图 2 为农业中文期刊论文信息识别与抽取级联模型，包括 segmentation 模型、header 模型、reference-segmentation 模型、citation 模型以及 fulltext 模型五个模型，每个模型实现不同的功能、对论文不同部分内容进行处理。

Segmentation 模型的主要是对整篇论文进行分块，识别标注出头部信息、正文信息、引文信息、页眉信息以及页脚信息，该模型标注的结果作为下级抽取模型的数据源，是整个信息抽取工作的基础。Header 模型是对期刊论文的头信息进行识别与抽取，识别的主要信息包括中英双语的论文的标题、作者、作者机构、摘要、关键词以及 DOI 信息。Reference-segmentation 模型对 segmentation 模型识别为引文的部分进行处理，主要是将整段引文数据分割为一条条引文数据。Citation 模型基于 reference-segmentation 模型处理的结果进行信息抽取工作，是对一条引文数据中的标题、作者、期刊名、出版年卷期、会议名称等信息识别和

doi:10.3772/j.issn.2095-915x.2019.06.004

抽取。Fulltext 模型的主要功能是对 segmentation 模型识别为正文的部分进行抽取工作，抽取的对象包括章节标题信息、段落信息。

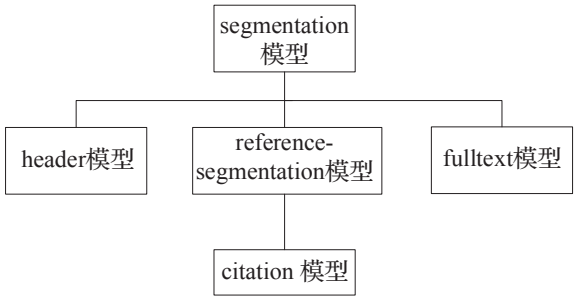

图 2 农业中文期刊论文信息抽取级联模型

2.2 数据采集及预处理

本次实验采用的数据源为 PDF 格式的 12 种农业领域的中文期刊论文数据，期刊包括《中国农业科学》、《作物学报》、《植物营养与肥料》、《农业展望》、《土壤学报》等共选取了 100 篇论文进行模型的训练与评估。

GROBID 工具中选用开源工具 pdf2xml 工具对 PDF 格式论文进行的预处理。在实验中，由于 GROBID 工具中 pdf2xml 在中文的识别上存在问题，需要安装配置中文支持包。文本预处理主要包括两部分，首先是使用 pdf2xml 对农业中文期刊论文的预处理，生成包含论文文本内容和格式内容的 xml 文档。其次是根据模型的需要，使用 lucence 分词工具对文本内容进行分词。

2.3 特征选择与序列标注

特征选择是指针对特定的抽取任务为模型选择合适的特征集。在对模型进行特征选择时要针对要抽取的内容选择合适的特征向量，这直接决定了最终模型的性能 [23]。在实验中，5 个模型识别信息的粒度以及抽取任务是不同的，因此各个模型进行特征选择工作的维度以及选取的特征项是存在差异的。例如在 segmentation 模型在对一篇文章进行分块时，相同行的文本内容属于相同的块信息。因此 segmentation 模型进行特征选择的维度是一行的数据，其他模型识别的更细粒度的信息，特征选择维度是分词后的词或字。

根据农业中文期刊论文中需要识别和抽取的信息的特点，从以下四个方面进行模型的特征选择，包括文本特征、格式特征、外部字典特征以及状态转移特征。文本特征是指每个数据单元的文本内容、是否为标点符号、是否包含数字、字符串长度等；格式特征主要是指字体大小、字体风格、在整个版面中所处的位置。外部字典特征是指文本内容是否属于特殊文本，例如地名、研究机构以及期刊名等。状态转移特征是指参考标注对象的上下文的状态，包括上文对应的观察值以及对应的状态值与当前状态之间的关系，该特征可以在特征模板配置。

序列标注是指根据选择的特征项对论文的内容进行处理，生成 CRF 模型可用数据。在确定模型选择的特征后，根据论文预处理后的数据生成相应的观察序列文件。通过人工标注的观察序列作为训练语料和评估语料。在 GROBID 工具中可以依据批处理模块，生成相应的观察序列文件以及论文信息标注有误的 xml 文件。xml 文件经过人工修正后，可以利用 GROBID 工具中训练模型将其余观察序列文件对应处理生成标注正确的数据文件，该数据文件可作为训练数据和评估数据。

2.4 特征模板

特征模板的作用是建立状态序列与观察序列之间的函数关系，也是特征函数选择的重要一步[24]。在 CRF 模型中，特征模板不仅能够建立起状态值与当前观察值的关系也能够建立与前文状态值和前文观察值之间的函数关系，它是影响模型效果的重要因素之一。本文在根据模型选取的特征进行序列标注时，是基于 GROBID 工具中的序列标注模块实现。生成的观察序列文件中，有些特征项的内容在农业中文期刊论文的信息识别标注中并不适用，无用特征冗余，会增加特征函数计算的复杂度，还会带来噪声，影响到模型训练的效率以及模型标注的效果[25]。

以头信息抽取为例，分别对以下三个特征模板在模型从训练时间上的消耗与模型标注效果的评价两个角度进行对比分析。特征模板 1 为 GROBID 工具 header 模型对应的特征模板。特征模板 2 为删除小写、前缀、后缀等部分在中文论文信息识别中无用的特征。特征模板 3 在特征模板 2 的基础上删除了可能在中文论文信息抽取中起作用的特征包括是否为数字等。选取相同的语料基于以上三个特征模板进行模型的训练与评估。采用了 20 篇农业中文期刊论文头信息数据进行实验。

表 1 为以上 3 个特征模板评估结果。从实验结果中，可以看出特征模板 1 和特征模板 2 较于特征模板 3 包含较多针对中文期刊论文信息识别的特征信息，所以模型的标注效果较好。特征模板 2 与特征模板 1 相比，特征模板 1 中包含有更多的冗余特征信息，模型训练时间偏长，但是模型的效果并没有提升。综合模型训练时间因素以及准确率，本文将特征模板 2 模型作为实验中 header 模型的特征模板。同时其他模型的模板参照这种方法删除部分在农业中文期刊论文进行信息识别中无作用的特征项对应的函数。

表 1 特征模板评估结果

特征模板	训练时间	准确率
特征模板1	1293.35s	99.89%
特征模板2	705.45s	99.89%
特征模板3	640.21s	97.32%

2.5 模型训练

根据模型特征模板以及训练语料采用 L-BFGS 算法对模型进行训练，计算各个特征函数的权重，最终生成模型文件。需要注意的是，在对模型训练时是存在一定的先后顺序的。首先需要对 segmentation 模型进行训练，这是由于其他模型都是基于 segmentation 模型的结果来进行下一步处理的。这与采用的级联模型的形式有关，只有上级模型能够进行准确的抽取，下级模型才能做出准确的应用。

3 实验结果及分析

3.1 模型评估与分析

3.1.1 评估指标

在对农业中文期刊论文信息识别与抽取模型的效果进行评估时，精确率以及召回率都不能代表模型整体的效果。F1 值是精确率和召回率的调和均值，是两者的均衡，可用来代表模型抽取效率。同时在对多标签模型的效果进行整体评估时，常采用各个指标的微平分来进行

doi:10.3772/j.issn.2095-915x.2019.06.004

评估。因此，本文采用机器学习中常用评估指标 F1 值以及微平分 F1 值（F1_micro）对模型效果进行评估。其计算公式如下：

$$F1=\frac{2\times precision\times recall}{precision+recall} \quad (1)$$

$$F1\_micro=\frac{2\times Micro\_precision\times Micro\_recall}{Micro\_precision+Micro\_recall} \quad (2)$$

其中 $precision=\frac{TP}{TP+FP}$，$recall=\frac{TP}{TP+FN}$，$Micro\_precision=\frac{\sum_{k=1}^{n}TP_k}{\sum_{k=1}^{n}TP_k+\sum_{k=1}^{n}FP_k}$，$Micro\_recall=\frac{\sum_{k=1}^{n}TP_k}{\sum_{k=1}^{n}TP_k+\sum_{k=1}^{n}FN_k}$。以 header 模型中的 <abstract> 评估为例，各个指标的含义如下：

TP：header 模型标注属于 <abstract> 且实际也属于 <abstract>；

FP：header 模型标注不属于 <abstract> 但是实际属于 <abstract>；

TN：header 模型标注不属于 <abstract> 且实际也不属于 <abstract>；

FN：header 模型标注属于 <abstract> 但是实际不属于 <abstract>。

3.1.2 模型效果分析

依据选定的评估指标 F1 值对农业中文期刊论文信息识别与抽取模型中各个子模型的效果进行评估，并与 GROBID 工具中应的模型的效果进行对比。

图 3 为农业中文期刊论文信息识别与抽取模型中 segmentation 模型与 GROBID 工具中 segmentation 模型的效果对比。从图 3 中可以看出，农业中文期刊论文信息识别与抽取模型中的 segmentation 模型在对 <header> 头信息、<body> 正文信息、<page> 页码信息、<headnote> 页眉信息以及 <references> 引文信息的识别中，F1 值基本都达到了 90% 以上。与 GROBID 工具的 segmentation 模型相比，农业中文期刊论文信息抽取模型中的各个部分信息的识别上都有大幅的提升。

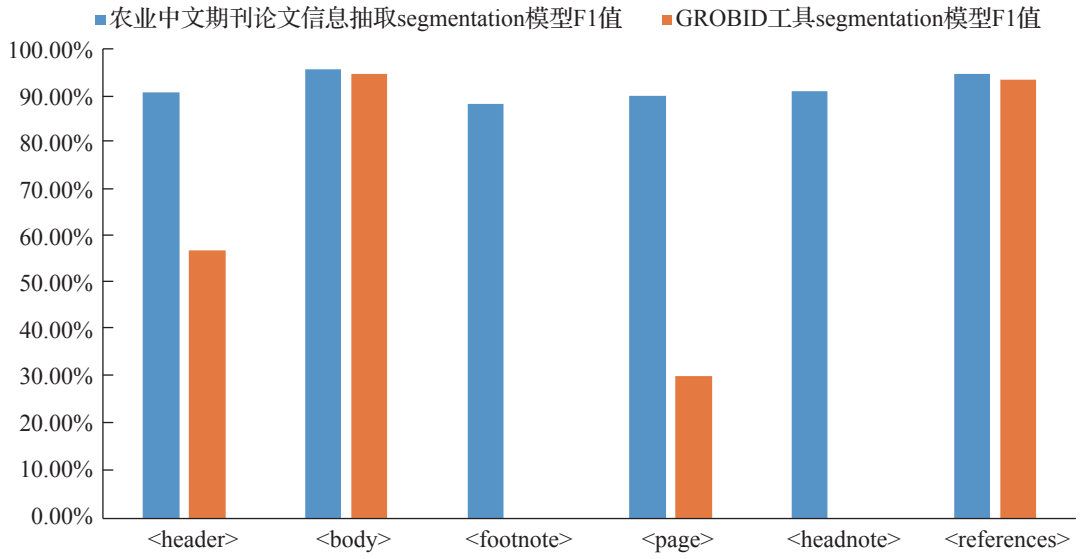

图 3 农业中文期刊论文信息抽取模型与 **GROBID** 对比 —**segmentation** 模型效果

图 4 为农业中文期刊论文信息识别与抽取模型中 header 模型与 GROBID 工具中 header 模型的效果对比。从图 4 可以看出，农业期刊论文信息识别与抽取模型中 header 模型能够高效识别出论文中的 <CNtitle> 中文标题、<CNaffliation> 中文机构、<CNabstract> 中文摘要、<CNaddress> 中文机构地址以及 <CN-keyword> 中文关键词，部分信息的 F1 值达到 90% 以上，尤其在中文标题的识别上 F1 值达到 100%。与 GROBID 工具中的 header 模型相比，农业期刊论文信息抽取模型中的 header 模型不仅实现了中英文论文头信息的识别与抽取，同时在 <author> 作者信息、<title> 英文标题、<affiliation> 英文发文机构、<abstract> 英文摘要等论文头信息的识别上都有较大提高。但本文 header 模型在 <keyword> 的识别上效果较差，会误将该部分标注英文摘要部分，仍需进一步改进。

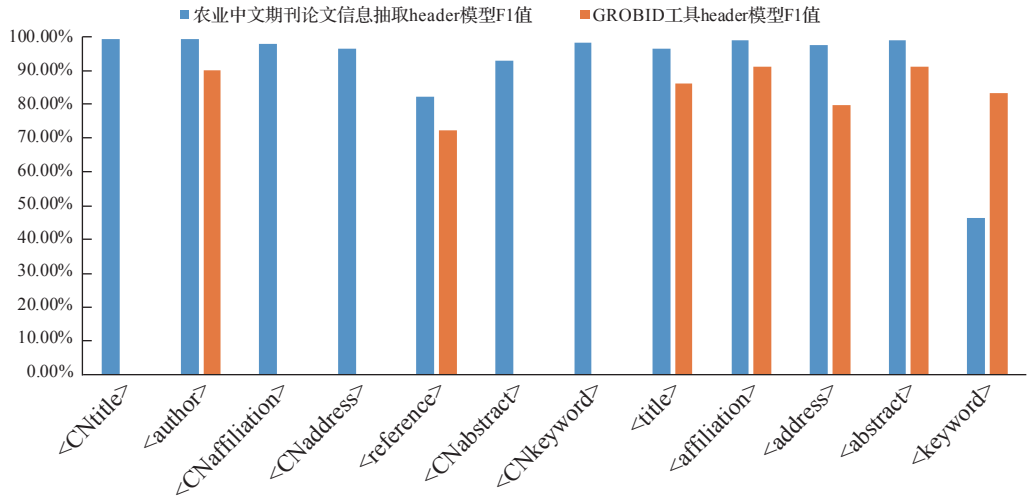

图 **4** 农业中文期刊论文信息抽取模型与 **GROBID** 对比 **—header** 模型效果

相对于 GROBID 工具的 reference-segmentation 模型，本文实验中的 reference-segmentation 模型并没有在特征模板上做调整，只是通过一些包含中文引文的训练数据对模型进行训练。实验结果显示农业期刊论文信息识别与抽取模型中 reference-segmentation 模型在对引文内容分割时，各个部分的识别上 F1 值均达到了 100%。

图 5 为农业中文期刊论文信息抽取模型中 citation 模型与 GROBID 工具中 citation 模型的效果对比。从图 5 可以看出，农业中文期刊论文信息识别与抽取模型中的 citation 模型在对引文信息各个部分信息进行识别时，F1 值均达到了 95% 以上，在部分信息的识别上 F1 值达到了 100%。与 GROBID 工具中的 citation 模型相比，农业期刊论文信息识别与抽取模型中的 citation 模型在各个部分信息的识别上均有提升。

在农业期刊论文信息识别与抽取模型中，fulltext 模型处理的对象为不包含图表信息的农业中文期刊论文，主要对论文正文内容的章节标题以及段落信息进行识别与标注。fulltext 模型在对段落的识别上 F1 值达到了 99.47%，但是在章节标题上存在一些问题，F1 值为 76.92%，需要进一步完善。

doi:10.3772/j.issn.2095-915x.2019.06.004

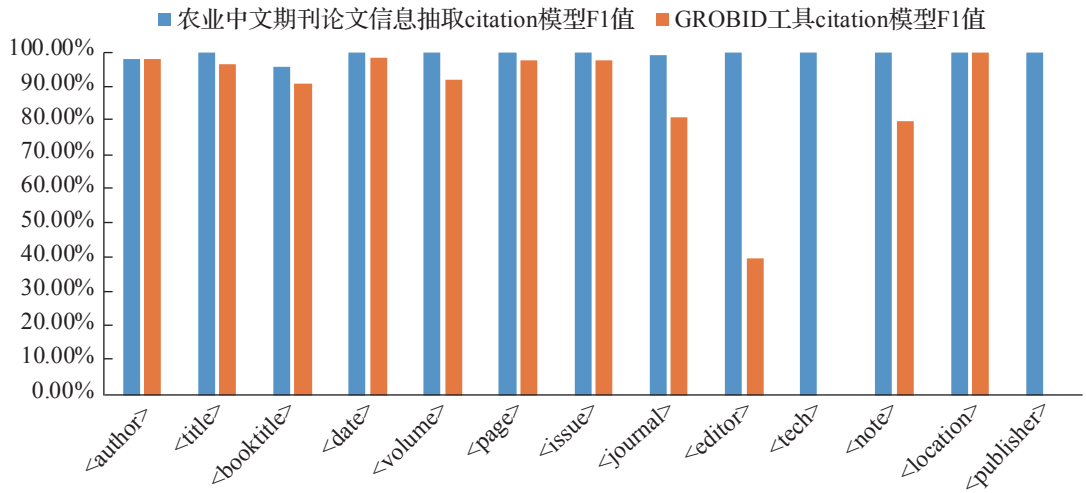

图 5 农业中文期刊论文信息抽取模型与 **GROBID** 对比 **—citation** 模型效果

整体来看，农业中文期刊论文信息识别与抽取模型中 segmentation 模型、header 模型、citation 模型以及 fulltext 模型的 F1_micro 分别 为 92.87%，96.26%，98.25%，99.5%，85.71%，尤其 reference-segmentation 模型的 F1_micro 达到了 100%。相较于 GROBID，农业中文期刊论文信息抽取模型中 segmentation 模型、header 模型、reference-segmentation 模型以及 citation 模型的信息识别效果都有一定的提升，能够实现对农业中文期刊论文头信息以及引文信息高效的识别。但是该模型中的 fulltext 模型在对章节标题的识别上效果较差，在对正文信息的识别与抽取上仍需继续优化。

3.2 农业中文期刊论文信息识别与抽取示例

本文构建的农业中文期刊论文信息抽取模型已经能够高效的完成对期刊论文的头信息、引文信息识别，最终基于 GROBID 工具将识别的信息以符合 TEI 标准的 xml 格式文档输出。图 6 为《中国农业科学》期刊中“北方直立穗型粳稻抗倒性的研究”论文头信息抽取结果。

```
<?xml version="1.0" ?>
<tei>
        <teiHeader>
                <fileDesc xml:id="北方直立穗型粳稻抗倒性的研究_张喜娟"/>
        </teiHeader>
        <text xml:lang="zh">
                <front>
<lb/>
        <reference>中国农业科学 2009,42(7):2305-2313<lb/> Scientia Agricultura Sinica<lb/></reference>
        <idno>doi: 10.3864/j.issn.0578-1752.2009.07.007<lb/></idno>
        <docTitle><CNtitlePart>北方直立穗型粳稻抗倒性的研究<lb/></CNtitlePart></docTitle>
        <byline><docAuthor>张喜娟，李红娇，李伟娟，徐正进，陈温福，张文忠，王嘉宇<lb/></docAuthor></byline>
        <byline><CNaffiliation>（沈阳农业大学农业部作物生理生态遗传育种重点开放实验室/辽宁省北方粳稻育种重点开放实验室</CNaffiliation></byline>
        <CNaddress>沈阳 110161）<lb/></CNaddress>
        <div type="CNabstract">摘要：【目的】培育抗倒能力强的水稻品种是实现水稻高产优质所面对的重要课题。本研究旨在探讨北方直<lb/> 立穗型粳稻抗
            倒性，为提高北方粳稻抗倒性，实现水稻高产优质提供一定的理论依据。
                                                ……
        </div>
        <CNkeyword>关键词：粳稻；抗倒性；直立穗型；弯曲穗型；茎秆形态性状；茎秆解剖结构；茎秆化学成分<lb/></CNkeyword>
        <docTitle><titlePart>The Lodging Resistance of Erect Panicle Japonica Rice<lb/> in Northern China<lb/></titlePart></docTitle>
        <byline><docAuthor>ZHANG Xi-juan, LI Hong-jiao, LI Wei-juan, XU Zheng-jin, CHEN Wen-fu, ZHANG Wen-zhong, WANG Jia-
        yu<lb/></docAuthor></byline>
        <byline><affiliation>(Key Laboratory of Crop Physiology ,Ecology, Genetics and Breeding, Ministry of Agriculture/Key Laboratory of
        Northern Japonica<lb/> Rice Breeding of Liaoning Province, Shenyang Agricultural University,</affiliation></byline>
        <address>Shenyang 110161)<lb/></address>
        <div type="abstract">Abstract: 【Objective】 In rice breeding, improvement of lodging resistance has been a main aspect for high-
        yield and<lb/> excellent-qualities. This study investigated the lodging resistance of erect panicle Japonica rice in northern China.
                                                .......
        </div>
        <keyword>Key words: japonica rice; lodging resistance; erect panicle; curved panicle; culm tissue characteristics; culm
        configuration;<lb/> chemical components of culms</keyword>

                </front>
        </text>
</tei>
```

图 6 农业中文期刊论文头信息识别与抽取结果示例

4 总结与展望

本文采用条件随机场的方法进行论文信息的识别与抽取，并基于 GROBID 工具的级联抽取模型构建农业中文期刊论文信息识别与抽取模型，实现对农业中文期刊论文信息的识别与抽取。实验结果表明农业中文期刊论文信息识别与抽取模型在对农业中文期刊论文头信息和引文信息的识别上效果较好，但在章节标题以及段落信息的识别上效果不佳，需进一步改进。

后续的研究可以从以下几个方面进行：第一，优化农业中文期刊论文信息识别与抽取模型中 fulltext 模型，提升正文章节标题信息识别与抽取效果。同时增加 fulltext 模型识别与抽取的正文信息的类别，例如图表、公式信息等。第二，与 GROBID 工具的级联抽取模型相比，农业中文期刊论文信息抽取模型缺少对信息细粒度识别的模型例如 name 模型、affliation-address 模型等。在接下来的研究中可以完善相应的抽取模型，从而实现对农业中文期刊论文更细粒度信息的识别与抽取。第三，面对越来越多农业领域的专利以及学术报告资源，也可基于该模型实现对专利以及学术报告资源中信息的识别与抽取，从而实现对其的高效利用。

参考文献

[1] 胡志刚 , 田文灿 , 孙太安 , 等 . 科技论文中学术信息的提取方法综述 [J]. 数字图书馆论坛 , 2017(10):39-47.

[2] 龚立群 , 马宝英 , 常晓荣 . 科技文献元数据自动抽取研究述评 [J]. 计算机系统应用 , 2013, 22(3):11-15.

[3] Wei W, King I, Lee H M. Bibliographic Attributes Extraction with Layer-upon-Layer Tagging[C]. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. IEEE Computer Society, 2007.

[4] 李朝光 , 张铭 , 邓志鸿 , 等 . 论文元数据信息的自动抽取 [J]. 计算机工程与应用 , 2002(21):189-191.

[5] 牛永洁 , 薛苏琴 . 基于 PDFBox 抽取学术论文信息的实现 [J]. 计算机技术与发展 , 2014, 24(12):61-63+68.

[6] Azimjonov J, Alikhanov J. Rule Based Metadata Extraction Framework from Academic Articles[J]. arXiv:1807.09009 [cs.IR], 2018.

[7] Day M Y, Tsai R T H, Sung C L, et al. Reference Metadata Extraction Using a Hierarchical Knowledge representation framework[J]. Decision Support Systems, 2007(43):152-167.

[8] 郭志鑫 , 金海 , 陈汉华 . SemreX 中基于语义的文档参考文献元数据信息提取 [J]. 计算机研究与发展 , 2006, 43(8):1368-1374.

[9] 黄泽武 . 基于语义的科技文献共享平台的信息抽取系统 [D]. 武汉 : 华中科技大学 , 2007.

[10] 欧阳辉 , 禄乐滨 . 基于 SVM 的论文元数据抽取方法研究 [J]. 电子设计工程 , 2010, 18(5):4-7.

[11] 白光祖 , 何远标 , 马建霞 , 等 . 利用小样本量机器学习实现学术文摘结构的自动识别 [J]. 现代图书情报技术 , 2014(Z1):34-40.

[12] 李雪驹 , 王智广 , 鲁强 . 一种规则与 SVM 结合的论文抽取方法 [J]. 计算机技术与发展 , 2017, 27(10):24-29.

[13] 黄永 , 陆伟 , 程齐凯 . 学术文本的结构功能识别——基于章节内容的识别 [J]. 情报学报 , 2016, 35(3):293-300.

[14] 方龙 , 李信 , 黄永 , 陆伟 . 学术文本的结构功能识别——在关键词自动抽取中的应用 [J]. 情报学报 , 2017, 36(6):599-605.

[15] 张秀秀 , 马建霞 . PDF 科技论文语义元数据的自动抽取研究 [J]. 现代图书情报技术 , 2009(2):102-106.

doi:10.3772/j.issn.2095-915x.2019.06.004

[16] 王昊 , 邓三鸿 . HMM 和 CRFs 在信息抽取应用中的比较研究 [J]. 现代图书情报技术 , 2007(12):57-63.

[17] 于江德 , 樊孝忠 , 尹继豪 . 基于条件随机场的中文科研论文信息抽取 [J]. 华南理工大学学报，2007, 35(9):90-94

[18] 陆伟 , 黄永 , 程齐凯 . 学术文本的结构功能识别——功能框架及基于章节标题的识别 [J]. 情报学报 , 2014, 33(9):979-985.

[19] 王东波 , 高瑞卿 , 叶文豪 , 等 . 不同特征下的学术文本结构功能自动识别研究 [J]. 情报学报 , 2018, 37(10):997-1008.

[20] Tkaczyk D , Collins A , Sheridan P , et al. Evaluation and Comparison of Open Source Bibliographic Reference Parsers: A Business Use Case[J]. arXiv:1802.01168v3 [cs.DL], 2018.

[21] Lipinski M, Yao K, Breitinger C, et al. Evaluation of header metadata extraction approaches and tools for scientific PDF documents[C]. ACM/IEEE-CS Joint Conference on Digital Libraries. ACM, 2013: 385-386.

[22] Lopez P, Romary L. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID[J]. Proceedings of International Workshop on Semantic Evaluation, 2010:248-251.

[23] 朱道辉 , 肖基毅 , 程阳 , 等 . 基于长距离依赖条件随机域的文本信息抽取 [J]. 计算机应用与软件 , 2011, 28(5):203-205.

[24] 薛俊欣 . 条件随机场模型研究及应用 [D]. 济南 : 山东大学 , 2014.

[25] 曾佳妮 . 基于条件随机场的中文短文本分类算法研究 [D]. 上海 : 上海交通大学 , 2013.