# 水稻本体实例构建研究*

李嘉锐[1]，崔运鹏[1]，张学福[1]，苏晓路[1]，郝心宁[1]，鄂志国[2]
（1. 中国农业科学院农业信息研究所，北京 100081；2. 中国水稻研究所，杭州 310006）

摘要：实例是本体的重要组成部分，它在很大程度上决定了本体的可用性。而目前本体实例构建的难度甚至超过了本体构建本身，大多实例的获取、更新和扩充依靠人工完成，既花费大量时间，又难以保证质量。文章在已完成的水稻本体概念框架基础上，利用神经网络方法进行半自动水稻实例抽取，提出水稻本体实例构建框架。统计数据表明，该方法能够有效地提高本体实例构建效率，大幅度降低手工劳动水平，提高本体实例质量，为本体实例构建和本体走向实际应用提供了思路和方法。

关键词：本体实例构建；水稻；神经网络；信息抽取
DOI：10.3772/j.issn.1673—2286.2014.11.008

## 1 引言

伴随着本体在国内的不断发展，领域本体被陆续构建出来，基于领域本体知识服务的研究成为热点。这也推动着领域本体从知识框架走向知识服务，为知识传播、知识检索等知识服务的智能化发展奠定了良好基础，对指导和揭示领域内知识关系和知识服务有着重大实践意义。

目前，一方面，已有众多领域本体被构建出来，这些本体都能够较好地反映领域内的知识概念及其关系，基本能够满足在此基础上的应用需求；另一方面，现有本体仍存在一些不足，如需构建高质量的本体并应用，则要求构建的领域本体相对比较完整，实例部分比较丰富和充分，才能在后期的开发和利用上有良好的表现。

现有的大多数领域本体都由概念、关系和少量公理和函数构成，经常出现实例部分缺失或者实例部分数量不足现象，这严重制约了领域本体的应用[1]。形成这一现象的主要原因包括：

（1）领域本体构建受到时间、人力和物力资源的约束和限制；

（2）领域本体实例获取难度较大，领域本体构建需要科研人员和领域专家不断沟通和交流，而且需要较强的专业知识；

（3）领域本体实例需要随时间不断更新和扩充。

领域本体构建工作已经进入较深层次阶段，不仅仅停留在概念层面，已经深入到实例及本体应用阶段。同时，这也表明较完整的领域本体有着更大的发展前景，发挥其知识服务作用，为上层具体应用起到坚实的支撑作用。

水稻本体构建依托于十二五科技支撑项目《面向外文科技文献的超级科技词表和本体建设》课题。截至2014年10月10日，累计收录水稻本体术语2623条，术语间关系4181个。主要由植物本体（Plant Ontology，PO）、基因本体（Gene Ontology，GO）、性状本体

* 本研究得到国家“十二五”科技支撑计划“面向外文科技文献信息的知识组织体系建设与应用研究”（编号：2011BAH10B00）资助。

（Trait Ontology，TO）、环境本体（Environment Ontology，EO）、植物学分类本体（Taxonomy Ontology，GR_tax）和序列本体（Sequence Ontology，SO）等6大部分组成。

水稻本体的术语和术语间关系在不断丰富、扩充。然而，水稻本体实例数量相对较少，英文水稻实例的构建遇到一定困难，要构建完整的水稻本体并发挥水稻本体的应用价值，亟需解决水稻品种命名实体的构建工作。

水稻品种命名实体识别的研究还不是很充分，传统的命名实体识别方法由于其受限条件较多，无法较好地适用于专业性知识较强、命名实体特征不明显、文献语料更新较快等的信息环境。为了适应农业科学领域命名实体专业性强、特征不明显、规则难总结、文献语料增长快等特点，本次水稻实例构建工作采用基于神经网络的命名实体抽取方法。

## 2 实验总体设计

### 2.1 实验设计目标

（1）水稻本体实例构建：在水稻概念术语框架构建基本完成的基础上，通过半自动方式获取水稻科技文献中水稻实例名称来构建本体的实例部分，实现水稻本体实例部分的构建，且随着水稻文献的不断更新，实现对实例部分扩充与维护，进而形成科技研究热点与水稻本体实例之间的相互关联。

（2）命名实体抽取：在大数据环境下，面对海量数据、内容复杂多样、知识专业性强的语言背景下，对命名实体抽取方法进行深入研究，进而实现命名实体的有效抽取，达到命名实体库的有效构建。

### 2.2 水稻本体实例构建设计

水稻本体实例构建框架设计（如图1）共分为语料构建、模型处理和数据分析三个部分，各部分功能描述如下：

（1）语料构建

语料构建主要是获取水稻育种与种质资源等领域的文献摘要，经过语料的预处理和清洗工作，达到语料的标准化、规范化，为模型处理提供文本数据，为实现水稻本体实例构建打下良好的数据基础。高质量语料是模型处理阶段计算命名实体之间相似度的依据和准确性的保证，因此，语料构建质量的优劣将直接影响模型处理的结果和水稻实例构建的效果。

（2）模型处理

模型处理采用神经网络的CBOW模型[2]和Skip-gram模型[2]对语料加工处理，将语料中的词汇构成N维向量空间，计算向量之间的相似性来表示词之间的关联关系，最终形成不同主题聚类。基于神经网络的信息抽取方法的优势是充分考虑词的上下文位置关系，对语料内容没有过多限制，模型处理时间花销小，适用于低频词汇的信息抽取。模型处理得到的初步聚类结果是数据分析的基础。

（3）数据分析

数据分析主要是对初步聚类结果进行筛选、分析和统计，最终得到水稻实例名称，完成水稻本体实例的构建工作。通过对初步聚类结果粗略筛选，获取有关水稻实例名称聚类较好的聚类号，并对这些类中的词进行仔细识别，从而得到水稻实例。通过对两种模型的处理结果分析统计，考察基于神经网络的信息抽取方法的有效性和抽取效果。

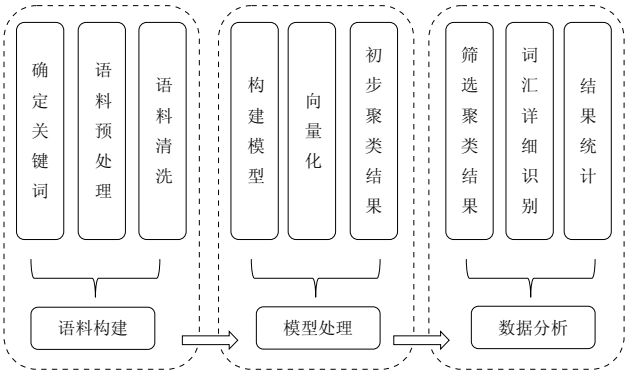


图1 水稻本体实例构建框架

## 3 水稻本体实例构建

### 3.1 语料构建

初次水稻实例构建时，以“rice”为关键词，从NSTL数据库中检索获取40000篇英文科技文献为语料，经过模型处理加工和人工筛选聚类后，发现水稻实例聚类效果并不理想。造成这种情况的主要原因是语料质量不佳，语料构建的内容涵盖了水稻领域研究的诸多方面，严重影响了模型处理结果，造成聚类主题比较分散，聚类效果欠佳。

在随后的实验时，充分总结和吸取失败的经验，调

整工作思路，在语料选取与语料质量控制上严格控制，为模型处理打下良好的数据基础。

首先，确定检索关键词。与水稻所专家进行沟通和咨询后，专家提供了符合水稻实例构建要求的一系列有关育种、种质资源方面的关键词。通过Thomson Reuters科学引文索引数据库中检索，经过对比和权衡检索到文献的质与量的关系。最终，选定了6个检索关键词“oryza+sativa+breeding”、“oryza+sativa+cultivation+technique”、“oryza+sativa+germplasm”、“rice+breeding”、“rice+cultivation+technique”和“rice+germplasm”。

选取Thomson Reuters科学引文索引数据库的Web of Science核心合集为检索相关主题词的语料源有以下两点理由：

（1）本体实例构建暂无较明确实施标准和规范。目前，大多数领域本体实例是从语料中抽取相关命名实体名称而得到的，这对构建的语料库质量有一定要求，从而保障提取的命名实体的相对规范性和专业性。

（2）使用Thomson Reuters科学引文索引数据库的Web of Science核心合集为检索数据库，该引文数据库中的收录的文献都是SCI文章，从而对检索出来的文献有了较好的质量保证。

其次，语料预处理，主要是指文献下载与摘要提取工作。从Thomson Reuters科学引文索引数据库中将不同关键词检索到的文献进行批量下载，共收集到8678篇文献，经过批处理将这些文献的摘要部分进行提取，最终合并成为一个文本文档。

最后，语料清洗。通过对水稻实例构成和相关背景知识学习，发现国内水稻实例命名以拼音+数字或者拼音的方式为主，国外水稻实例命名没有明显规律可循，这也是水稻实例构建的难点之一。为了兼顾水稻领域学科自身知识特点和提高模型处理效果两点因素考虑，最后采取清除无关符号，例如“%”、“+”、“/”和“.”等标识符。同时，保留单词的原始形态（生物学领域的基因和蛋白字母拼写只有首字母大小写不同）、中横杠线和数字等标识符。

## 3.2 模型处理

在语料规范化处理过后，将文本文档输入CBOW模型和Skip-gram模型进行语料聚类处理。通过两种模型在聚类数量阈值设置为400、500和600个条件下进行语料处理加工，最后共获得6个初步聚类结果文档（如图2），每个聚类结果包括16327个词汇。

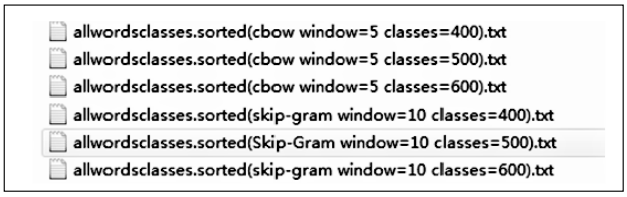


图2 初步聚类结果文档截图

每个初步聚类结果包括两项内容（如图3），第一列为聚类中的词汇，通过向量空间计算，将相似程度接近的词聚集在一个类中；第二列为类号，不同的类号表明不同聚类主题，相同聚类词汇的类号是一样的。

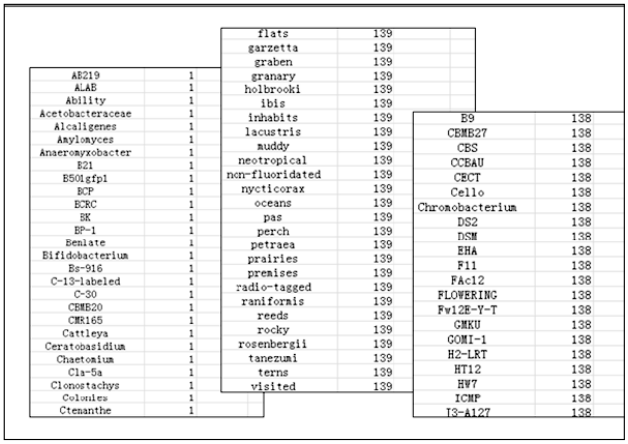


图3 初步聚类结果截图

## 3.3 数据分析

获取两个模型处理完成的初步聚类结果后，需要进行人工筛选、分析、整理和统计。首先，粗略筛选，通过人工大致浏览全部初步聚类结果，记录有关水稻实例聚类较好的类号，排除其他无关聚类结果，这些记录的类号将进行词汇的详细鉴别；其次，词汇详细识别，将记录类号中的词汇通过回溯到摘要中，判定是否为水稻实例，最终成功得到类中全部水稻实例（如图4）；最后，结果统计，将全部处理结果进行分类统计，从而可以得到详细、准确的实验结果数据（如表1和表2）。

在已完成的水稻本体概念框架基础上，利用神经网络方法进行半自动水稻实例抽取，并成功地从8678篇

| 类号 | 水稻实例名称 |
|---|---|
| 30 | Hoshiaoba；IR65598-112-2；IR65600-87-2-2-3；IR69125-25-3-1-1； |
| 65 | Annapurna；Bharati；Damodar；Deepa；Dhauli；K-851；Khandagiri；Kusuma；LGC-407；Liaoning；Nilgiri；Paridhan-1；PDM-116；Pusa-2-21；Ratna；Rudra；Sankar；Vaghari |
| 155 | Nokkei 01530；Suweon 287；810S；Akenohoshi；Aya；Bobwhite；Gigante；IR28；KatyRR；Kaybonnet；KDML 105；Kochihibiki；LA3；Neda；Nihonmasari；Nongan；Nongken；Piaui；Sumitakara |
| 159 | Apo；Pusa Bold；BR11；BR29；C306；Chubu 111；Dhagaddeshi；Hitomebore；IR20；IR58821-23-B-1-2-1；IR-64；IR72；IR74371-46-1-1；KDML105；Kele；KMR-3R；MTU1010；N22；Nanjing35；Har-una Nijo；Norungan；RD6；SH527；Swarna；Vandana |
| 171 | Bluebonnet 50；CL161；Fukuhibiki；IR36-Shuang；Kitaibuki；MOMI；NIL28；Ouu；Takanari；WAB56-104；YTH183 |

图4 水稻品种识别结果截图

表1 CBOW模型统计结果

| 聚类数量阈值 | 待分析类目数 | 待分析类目总词汇量 | 水稻品种名称总词汇量 | 查准率 |
|---|---|---|---|---|
| 600 | 26 | 1211 | 500 | 41.3% |
| 500 | 24 | 1182 | 537 | 45.4% |
| 400 | 21 | 1292 | 527 | 40.8% |

表2 Skip-gram模型统计结果

| 聚类数量阈值 | 待分析类目数 | 待分析类目总词汇量 | 水稻品种名称总词汇量 | 查准率 |
|---|---|---|---|---|
| 600 | 31 | 1203 | 520 | 43.2% |
| 500 | 30 | 1281 | 522 | 40.7% |
| 400 | 16 | 934 | 410 | 43.9% |

注：聚类数量阈值：聚类总数量；待分析类目数：经过粗略筛选，需要进行详细分析的聚类数量；待分析类目总词汇量：需要进行详细分析聚类所包含的词汇总数；水稻品种名称总词汇量：聚类中包含水稻实例名称词汇总数；查准率：（水稻名词汇数/待分析类目总词汇量）*100%

水稻英文科技文献中获取500余个水稻实例，查准率超过40%。同时，基于神经网络的信息抽取方法能够有效地应用于专业学科的命名实体识别，提高本体实例构建效率，大幅降低手工劳动水平，提高本体实例质量。

本次水稻本体实例构建成功的主要因素有三个方面：第一，语料质量，水稻本体实例构建的语料质量直接决定着模型处理后的聚类效果，间接决定着人工识别的工作量的大小。所以，本次构建工作在语料构建方面投入较大精力，不断调整检索关键词，保证语料的高质量；第二，专家指导，由于缺乏相关专业知识背景，在构建语料与水稻实例识别等环节遇到困难，在得到了水稻领域专家指导和帮助下，顺利攻克一个个问题，保证了实验的顺利进行；第三，利用神经网络的信息抽取方法，可以在面对海量数据的情况下，较快速、准确地进行聚类，与传统的手动构建相比，大大节省了人工劳动量，提升了命名实体获取的效率。

## 4 结语

利用神经网络的命名实体抽取方法，成功地实现了水稻实例库的构建，完成了水稻本体实例的构建工作，使得水稻本体更加全面、完整。实验结果表明，基于神经网络的命名实体抽取方法在水稻本体实例构建方面具有一定的适用性。但是，也存在着一定不足：第一，由于受到个人能力和时间等因素限制，没有收集到更大规模的语料进行该方法测试，所以无法确定大规模语料处理的时间消耗和识别效果；第二，本体实验实例构建对象只是水稻实例，没能对多个实例对象进行识别效果的比较和分析。希望下一步工作可以在以上两个方面有所突破，为本体实例构建提供更多帮助，对本体实例构建和本体走向实际应用提供新的思路和方法。

## 参考文献


[1] GRUBER T R. A translation approach to portable ontology specifications, KSL 92-71 [R]. San Francisco: Knowledge Systems Laboratory of Stanford University, 1993.

[2] CBOW Model, Skip-gram Model [EB/OL]. [2014-11-21]. http://outofmemory.cn/wr/?u=http%3A%2F%2Fwww.kemaswill.com%2Fmachineearning%2F%25e8%25af%258d%25e5%2590%2591%25e9%2587%258f%25e4%25b9%258bnewlog-linear-models%2F.

[3] 袁璐.智能信息检索中基于本体的文本信息抽取的研究与实现[D].沈阳工业大学,2009.

[4] 李阳.英文文本中命名实体识别及关系抽取技术研究[D].华东理工大学,2012.

[5] 张素香.信息抽取中关键技术的研究[D].北京邮电大学,2007.

[6] BENGIO Y. Hierarchical Probabilistic Neural Network Language Model [EB/OL]. [2014-11-21]. http://www.iro.umontreal.ca/labs/neuro/pointeurs/hierarchical-nnlm-aistats05.pdf.

[7] MIKOLOV T, CHEN Kai, CORRADO G, et al. Efficient Estimation of Word Representations inVector Space [EB/OL]. [2014-11-21]. http://arxiv.org/pdf/1301.3781.pdf.

[8] Licstar. Deep Learning in NLP(一) 词向量和语言模型[EB/OL]. [2014-11-21]. http://licstar.net/archives/328.

**作者简介**

李嘉锐，男，1985年生，中国农业科学院农业信息研究所在读硕士研究生，研究方向：信息资源管理，E-mail：lijiaruiljr@163.com。
张学福，男，1966年生，中国农业科学院农业信息研究所研究员，研究方向：农业知识组织与可视化分析，通讯作者，E-mail：zhangxf@caas.cn。

Research on Construction of Rice Ontology Instance

LI JiaRui[1], ZHANG XueFu[1], CUI YunPeng[1], SU XiaoLu[1], HAO XinNing[1],E ZhiGuo[2]
(1. Agricultural Information Institute of CAAS, Beijing 100081, China; 2. China National Rice Research Institute, Hangzhou 310006, China)

Abstract: Instance is an important component of Ontology, which largely determines the availability of Ontology. Constructing an Ontology Instance is more difficult than Ontology construction itself in many cases. Instance acquiring, updating and expanding rely on manual operation completely. The work needs a lot of time and is hard to ensure quality. Based on the conceptual framework of Ontology, a semi-automatic extraction approach was applied by using neural network (NN) method in rice instances. The construction framework of rice ontology instance was proposed. The result shows that this approach effectively improved the efficiency of Ontology instance construction, significantly reduced the manual labor, and improved the quality of Ontology Instances. This approach provides a new idea and method for Ontology applications from Ontology Instance construction and Ontology.

Keywords: Ontology instance construction; Rice; Neural networks; Information extraction

（收稿日期：2014-11-20）

# 南京大学组建江苏省“数据工程与知识服务”重点实验室

在江苏省教育厅公布的2014年省高校重点实验室名单中，由南京大学信息管理学院牵头申报的“江苏省数据工程与知识服务重点实验室”获批立项，成为南京大学社科领域首个江苏省高校重点实验室。该实验室将依托国家重点学科情报学，联合江苏省科技情报所，集中申请单位和共建单位的研发优势，开展协同创新，解决大数据及相关服务领域的技术和应用问题，推动和引导江苏省大数据产业和科技服务领域的快速发展。

数据工程与知识服务重点实验室建设期为三年，欧洲文理科学院院士、南京大学信息管理学院叶鹰教授担任实验室主任，教育部长江学者、南京大学信息管理学院苏新宁教授为实验室学术带头人，并担任学术委员会主任，实验室专兼职研究人员近50人。实验室的主要研究方向为大数据环境下的数据整合与规划、知识关联技术、数据分析理论、知识服务技术等，将重点突出大数据环境下的知识服务，实验室将联合有关企事业单位，建立示范性相关数据平台、技术平台、服务平台，促进有关应用和技术的转化和推广。

数据工程与知识服务重点实验室为开放实验室（http://deks.nju.edu.cn）每年还将提供数据和设备平台，设立开放课题供国内外有关学者共同参与数据工程与知识服务的研究。2015年度设立的开放课题研究方向有，（1）大数据整合与规划：政策与战略；（2）大数据知识关联：技术与应用；（3）大数据分析：理论、模型与应用；（4）大数据的行业影响：金融、医疗等；（5）数据开放、数据交易与隐私保护；（6）数据工程、数据资源融合。