农作物模板自动生成的技术研究

钱跃良　　刘开瑛

中科院计算技术研究所，北京 100080　山西大学计算机科学系，太原 030006

摘要：本文主要介绍农作物信息库的建造，包括信息库的组织形式以及资料获取。并提出基于中文信息处理技术从中文文本中自动生成农作物模板的方法。

关键字：农作物种质资源　模板生成　信息抽取

Research on the Automatic Generation of Agricultural Products Templates

Qian Yueliang　　Liu Kaiying

Institute of Computing Technology
Chinese Academy of Science, Beijing 100080

Computer Science Department
Shanxi University, Taiyuan, Shanxi 030006

ABSTRACT: This paper presents an introduction of the construction of agricultural information base, including its framework and data acquisition. And, a method of automatically generating agricultural products templates from Chinese texts, based on current Chinese information processing techniques, is also given.

Keyword: agricultural species resources, template generating (TG), information extraction

1 引言

农作物是指在农业上培育、栽培的各种植物。包括：1．粮食作物：稻、麦、谷类、玉米、高粱等；2．油料作物：油菜、芝麻、花生等；3．蔬菜：黄瓜、西红柿等；4．棉花；5．烟草等。从农作物文本中自动抽取信息构造结构化信息库是模板生成（TG）问题，即在对中文文本经过文本分析和语义分析后，自动识别出某一类型的事件或关系，并抽取与这些事件或关系相关的描述参数，最后将抽取出的信息结构化表示。在模板生成和信息抽取得到的结构化信息库的基础上，可以进一步完成：信息搜索、数据挖掘、机器翻译、文本生成等操作。要从文本中抽取特定信息，必须构造信息抽取模式，而模式构造需要自然语言处理理论基础和识别技术。特别是这些模式是针对领域

的，当转向另一领域时，就需要构造新的模式。不仅粮食作物与蔬菜的模式不同，即使在粮食作物中水稻与玉米也不同。

本文介绍构造农作物种质资源信息库和农作物信息库的概况。并以信息库作为农作物信息获取的资源，探讨农作物模板自动生成的方法。

2 建造农作物信息库

2.1 目录形式组织形式

农作物信息库涉及到农业生产的方方面面，各种类型的信息资源可以按照如下的目录形式组织：

1. 农业基础科学

1.1 农业科普知识；1.2 病、虫、草、鼠害；1.3 自然灾害；1.4 生产时空；1.5 水利；1.6 气候气象；1.6 耕作区划；1.7 土壤肥料；

2. 现代农业资源

2.1 农业科技动态；2.2 农业科技成果；2.3 农业技术成就；

3. 农业知识经济

3.1 农业新闻；3.2 农业政策，法规；

4. 新技术、新产品、新方法

5. 农作物品种

5.1 作物良种；5.2 栽种良法；5.3 品种栽培；5.4 高产高效栽培模式；5.5 种植管理；5.6 种子管理；

6. 农产品加工

7. 供需信息

8. 可持续农业

2.2 从农作物种质资源信息库中获取信息

农作物种质资源即农用植物遗传资源，是生物多样性的重要组成部分，也是人类赖以生存和发展的重要物质基础。作物种质资源既包括在任何地区、任何时间所栽培或生长的植物种、半驯化种、野生种和亲缘种，还应包括人们利用采、伐、摘、挖、放牧等手段而为人所用的各种植物种。作物种质资源及其多样性，不仅为人类的衣、食等方面提供原料，为人类的健康提供营养品和药物，而且为人类幸福生存提供了良好的环境。

中国作物种质资源十分丰富。据初步统计[1]，世界上栽培植物有 1200 余种，中国就有 600 余种。近 20 年的研究项目重点是收集和保存，经整理编目的种质资源已有 160 种作物，其中中国的地方品种、稀有种及野生近缘植物，约占收集材料总数的

70%。还有小作物、水生蔬菜、热带作物、牧草、调味、香料、花卉和药用植物有待进一步收集。 现已初步形成了种质资源收集、保存、鉴定、研究、创新、利用的工作体系，建立了中国作物种质资源库。它可为人们选育所需求的新品种，开展生物技术研究提供取之不尽，用之不竭的基因来源。

例如：从国家种质库中已初步筛选出 2 万余份综合性状较好或具有某一特优性状的种质资源。这些优异材料，部分已提供给生物技术、育种利用，有些还在高寒地区、盐碱地区、干旱地区和矿山复垦区直接推广利用，有些还向世界各国提供，都已初见成效。

又如：人口增加，需要增加粮食产量。粮食持续增产主要得靠提高单位面积产量。欲提高单产，灌溉、施肥及其它栽培措施固然重要，而优良品种却是首要因素。选育良种要靠作物基因的多样性。目前任何高新技术都还不能创造基因，而只能在生物体之间转移、复制或修饰基因，丰富的基因存在于多种多样的品种(包括古今中外的品种)及其野生亲缘植物中，它们统称为作物种质资源或作物基因资源。人们已认识到，育种实质上是种质资源再加工，因此保护农用植物基因资源多样性，发掘优异基因，将《作物品种资源研究》列为国家重点科技攻关项目，是实现农业可持续发展的中心策略。

2.3 动态收集新技术、新方法和最新动向

Internet 网上有 70 多个中文农业网站，及时提供最新动态，这些信息具有内容新颖、门类齐全等特点，是我们建造农作物信息库，构建模板的重要信息来源。

从专业的农作物资料以及报刊杂志中动态收集新技术、新方法和最新动向获得更实用的信息是我们建造农作物信息库的另一个重要信息来源。

下面列出我们已经入库的示例。

1）基因农业—基因育种

通过基因工程培育的转基因作物可以使粮食产量成倍增长。如杂交油菜种子“秦油二号”因其高产、稳产、适应性强、出油率高，畅销我国 15 个油菜生产省区。杂交玉米种子“陕单*号”覆盖我国 10 多个玉米主产区。

2）太空农业—太空种子

经过航天育种获得的“太空种子”，包括粮食、蔬菜、花卉等 400 多个品种，已经育成了一批高产量、优质产品。如“太空玉米”可长出版 5 种颜色，味道比普通玉米好。水稻已获得了比原品种增产 20%的高产优质品系。小麦平均亩产比原品种增产 9%。

3）精确农业

国际上指采用 3S 技术：GIS 地理信息系统，GPS 全球定位系统，RS 遥感系统而言。在时间和空间尺度上获取农作物生长发育，病虫，草害，水肥条件及众多的信息，以便进行诊断和决策。

4）设施农业

以温室，大棚，畜禽鱼场为主要内容。为绿色农业和菜篮子工程做出贡献。

5）水稻多品种间栽稻瘟病控制技术

水稻是供养全世界 1/2 人口的谷类作物，由真菌引发的稻瘟病是水稻最主要的疾病。在单种栽培的农田里，稻瘟病会像“火烧田野干草”一样蔓延。早已有科学理论提出，不同的稻瘟病攻击不同的品种水稻，水稻的品种遗传多样性可能是解决单种栽培水稻易遭受疾病侵害的一种方法，但一直缺乏实验数据。2000 年 8 月英国《自然》杂志和英国《科学》杂志同时发表文章，指出了在云南省进行的 4 万公顷大面积水稻品种多样性的间栽，控制了稻瘟病的实验获得成功的消息。水稻产量更高，收入更多。

6） 小麦《中优 9507》麵包专用新品种问世。

我国小麦的消费量、生产量、种植面积均居世界首位，但品值和效益方面是小麦弱国。我国进入 WTO 后，以美国为代表的小麦出口国将造成极大的压力。2000 年 6 月农业部组织认定农业科技跨越计划项目：小麦《中优 9507》麵包专用新品种，成功地克服了农作物品通常的“高产与优质不容”的矛盾，是我国特性最好的高产优质面包专用小麦新品种。

7）调整农业结构，推进农产品国际大循环（中国科学院国情分析研究小组 <第八号国情报告>）

目前中国的粮食价格已经高于国际价格。1999 年大米、玉米、小麦的国际价格分别为每公斤 1.9 元、0.6 元、0.73 元，而我国的价格每公斤为 2.31 元 1.42 元 1.31 元，是国际价格约 1.21—2.36 倍。我国加入 WTO 后，粮食市场将面临更大的冲击，然而我国劳动密集型的农业如蔬菜、水果及部分经济作物和养殖业，技术密集型部门如花卉、药材、部分经济作物和食品加工等，则具有较大的比较优势。因此调整农业结构，适度压缩粮食生产面积，发展劳动密集型与技术密集型产品并到国际市场上换取以粮油为代表的资源密集型产品，不仅可以取得经济利益，而且还大大减轻国内水土资源的压力，使我国农业走上可持续发展道路。

3 从中文文本语料库中生成农作物模板问题

模板生成一般有以下步骤。

3.1 文本采集与主题选取

从真实语料中采集相关的文本资料：专刊专著、学术论文、科普读物以及报刊新闻报道，组建农业领域的中文文本总语料库，如农作物种质资源信息库或农作物信息库。语料的采集目前主要是从大量未标注语料中通过关键字检索，选出相关的文本。关键字的确定是通过在字典等资源中查找所有有关词语，词的选择应尽量广泛，以免丢失文本。在得到检索的文本后，应进一步选择，如通过察看文章的主题，以保证文章与所抽信息的内容相关，提高模式获取的正确性。

3.2 构建领域或品种的子语料库

从中文文本总语料库分出各领域或品种的子语料库。如粮食作物、油料作物、蔬菜等领域库，又如水稻、大豆、黄瓜、西红柿等品种库。构建指定品种的模板需以专业文件为主体。

例如在粮食作物和油料作物的品种上，应以 2000 年农业部组织筛选出一批适合于当前种植结构调整的优值品种为主。这些优质品种包括五类共 180 个品种，其中早稻 38 个，小麦 26 个，玉米 65 个，油菜 39 个，大豆 20 个。

又如在栽培技术上，已经公开出版了蔬菜，果树及花卉等方面的专业小册子近百种。这些专业小册子具有主要框架结构合理，内容丰富、科学实用、通俗易懂、实用性强等特点，是构建领域或品种语料库的主要语料之一。 以黄瓜为例，可作为构建模式的基本框架如下：

（1）黄瓜的特性和生长需要的环境条件。包括根、茎、叶、花、种子和果实等。
（2）黄瓜的品种。包括长春密刺等到 25 个品种的特征和特性。
（3）黄瓜嫁接。包括嫁接用具、嫁接方法和注意事项。
（4）塑料薄膜覆盖栽培。包括春大棚、秋大棚、塑料棚等。
（5）温室黄瓜栽培。包括冬秋温室栽培、秋冬温室栽培等。
（6）露地黄瓜栽培。包括春黄瓜、夏黄瓜和秋黄瓜。
（7）黄瓜的病害和虫害防治。
（8）黄瓜的食用和药用介绍。

3.3 文本标注

对语料进行加工处理，包括词语切分、词性标注、短语标注、专业术语标注以及专有名词标注等。

3.4 句子检索

从已标注的文本中系统抽取与领域和品种的特征、特性相关的句子。目前主要通过关键字检索，将不包含关键字或其它必要信息的句子去除，得到的所有句子将作为候选模式。

3.5 标注句子的模式合并

模式合并是整个算法的核心，它包括模式间相似度的计算及合并操作的方法。开始时，检索得到的每一个标注的句子都是一个模式。计算任两个模式之间的相似度，从中选择相似度最大的两个模式，将其合并生成一个新的模式。用新生成模式替换两个旧模式后，模式总数减少一，反复执行合并操作，直到任两个模式相似度不满足阈值条件或其它终止条件。

3.6 生成模板框架

经过模式合并，并从中筛选出正确的模式，将句子分类成若干不同的栏目，生成模板框架。如基本特征、特性和生长环境、栽培技术及用途、储藏、加工和销售等。可以定期对模板框架中栏目、属性进行调整，逐步完善、实用。

采用自然语言处理技术，研究从农作物文本中生成模板和信息抽取方法以及农作物信息资料库的建设，正在试探中，许多问题尚待进一步深入研究。

参考文献

1. 刘 旭，作物种质资源与农业科技革命（中国作物种质资源信息系统），中国农业科学院作物品种资源研究所，北京，100081

2. 何信生，论中国农业国际化的障碍与对策，摘自《农业经济问题》

3. 刘 旭 董玉琛，世纪之交中国作物种质资源保护与持续利用的回顾和展望（中国作物种质资源信息系统），中国农科院作物品种资源研究所，北京 100081

4. Chikashi Nobata, Satoshi Seking, Toward Automatic Acquisition of Patterns for Information Extraction, International Conference on Computer Processing of Oriental Languages: Tokushima Japan, 1999

5. Ellen Riloff, Automatically Generating Extraction Patterns from Untagged Text, Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), pp. 1044-1049

6. Neus Catala, ESSENCE: a Portable Methodology for Building information Extraction Systems, Technsl Rrport LSI-98-54-R, Department de Llenguatges I Sistemes infomatics, Universitat Politecnica de Catalunya, 1998