杨　欣,颜　伟,朱　银. 江苏省农业种质资源智能检索系统的设计研究[J]. 江苏农业科学,2015,43(12):482 – 484.
doi:10.15889/j.issn.1002 – 1302.2015.12.147

江苏省农业种质资源智能检索系统的设计研究

杨　欣, 颜　伟, 朱　银
(江苏省农业科学院粮食作物研究所/江苏省农业种质资源保护与利用平台,江苏南京 210014)

摘要:江苏省农业种质资源智能检索系统是基于本体和语义网相关技术进行设计,实现具有简洁易用的人机交互界面的智能检索系统,系统支持通过各种自然语言形式描述的查询条件实现对种质资源内容语义层面的智能检索。主要介绍了系统架构、主要功能设计、农业种质资源本体构建等内容,并详细描述了资源语义标注、智能检索处理的相关业务流程。

关键词:江苏省;本体;种质资源;语义网;智能检索
中图分类号: S126;TP392　　**文献标志码**: A　　**文章编号**:1002 – 1302(2015)12 – 0482 – 03

农业种质资源是人类赖以生存和发展的战略性资源,是维系国家食品安全和农业可持续发展的基本保证。目前江苏省已保存农业种质资源 5.39 万份,并通过信息共享服务系统对外提供包括农作物、林木、水产、家养动物等四大类农业种质资源信息,共享特征数据超过 150 万个[1]。面对庞大的数据信息,怎样让用户精确地检索出所需数据是当前亟需解决的问题,本研究旨在设计研究一套基于本体的智能检索系统,方便实现农业种质资源的智能检索。

1　系统架构

江苏省农业种质资源智能检索系统采用 B/S(browser/server)架构,用户使用浏览器即可随时随地用自然语言描述查询条件进行种质资源相关内容的智能检索,具有简洁易用的人机交互界面。系统整体架构如图 1 所示。

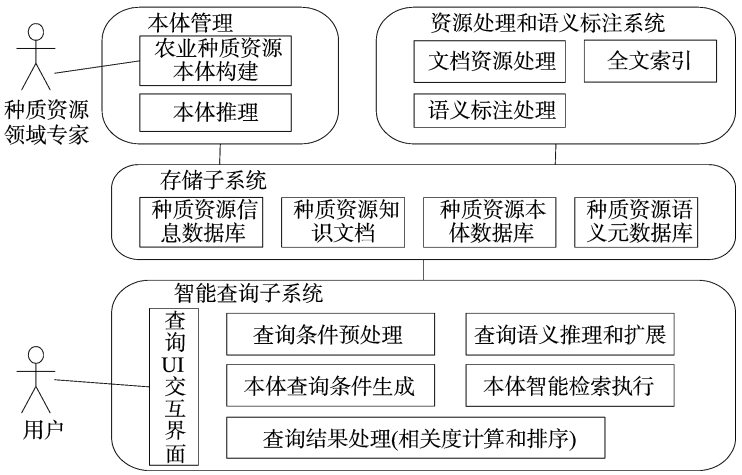

图1　系统架构

2　系统功能设计

江苏省农业种质资源智能检索系统主要包括本体管理子系统、资源处理和语义标注子系统、存储子系统、智能查询子系统。

2.1　本体管理子系统

本体是共享概念模型的明确形式化规格说明[2],它描述了客观事物的概念以及概念之间联系的领域知识。农业种质资源本体由种质资源领域专家和信息专家一起设计和构建,它实现了农业种质资源的形式化表达。

本体管理子系统实现本体构建、本体导入、本体解析、本体修改和更新、本体推理、数据持久化存储等功能。系统支持导入外部本体构建工具(如 Protégé)生成的本体描述文件,支持 W3C 的 OWL 本体描述规范。支持常见的本体推理功能,例如概念之间的上下位关系、同位关系的推导以及本体一致性检查等功能。

2.2　资源处理和语义标注子系统

种质资源信息共享服务系统内已存在大量的结构化和非结构化的数据。结构化数据主要包括数据库中保存的包括农

收稿日期:2015 – 07 – 27
基金项目:江苏省农业科技自主创新资金[编号:CX(12)5087];国家农作物种质资源平台项目(编号:NICGR2015 – 025)。
作者简介:杨　欣(1982—),女,江苏扬州人,硕士,助理研究员,主要从事农业种质资源信息系统研究。E – mail:icekeleyx@163.com。
通信作者:颜　伟,硕士,研究员。E – mail:yanw9@hotmail.com。

作物、水产资源、家养动物、林木资源四大类种质资源共性和特性描述数据,以数据表的形式存在。非结构化的数据主要包括各种知识库文档,如科技动态、科普知识等。资源处理和语义标注子系统的功能就是给上述信息进行语义分析和标注,按照设计好的规则和程序对信息进行预处理,方便实现后续的语义检索。语义标注实现了数据资源和本体知识之间的关联,系统支持自动化和人工等多种方法进行标注。

2.3　存储子系统

存储子系统是指系统中的数据持久层,以关系数据库或者文档数据库的形式存储所有的数据。农业种质资源专家构建的本体保存在种质资源本体数据库中,资源处理和语义标注的数据保存在语义元数据中。

2.4　智能查询子系统

智能查询子系统是系统的核心业务模块。查询条件预处理模块接收用户输入查询条件,进行自然语言处理,对关键词进行语义分析和本体匹配,建立词汇到本体知识的映射关系。查询语义推理和扩展功能是指对查询预处理结果进行进一步的语义扩展,例如查找本体概念的同义词或者扩展查询条件到上位概念等。对于有歧义的关键词,系统可以给出多种选择让用户进一步确认。查询条件生成模块将语义查询条件转化为系统中的实际查询语句命令,本系统语义查询引擎可基于 Apache 的 Jena 开源项目开发,系统使用 SPARQL(Simple Protocol and RDF Query Language)查询语句进行语义资源查询,SPARQL 是一种 W3C 推荐的类似 SQL 的面向 RDF、OWL 数据模型的查询语言[3]。系统支持对语义检索结果进行相关度计算和排序。UI 交互界面模块提供友好的人机交互接口,支持用户以自然语言形式输入查询条件并展示查询结果,同时还支持提供查询语句建议和示例。

3　系统关键技术

3.1　农业种质资源本体的构建

针对农业领域的本体构建国内已经有不少单位展开了研究,中国农业科学院对领域本体构建方法和应用作了详细的分析[4]。中国水稻研究所建立了一套较为完善的水稻本体系统[5]。在面向海量非结构化数据的农业领域知识获取方面,国际上也开展了基于自然语言处理、机器学习、统计推断、数据挖掘等方向的研究。

为了实现种质资源的智能检索,系统首先要实现江苏省农业种质资源本体的构建,包括农作物、水产资源、家养动物、林木资源四大类数十种小类,种质资源种类多,本体构建的工作量很大。种质资源分类的本体描述如图 2 所示。

以水稻为例,结合水稻相关国际和国家标准以及种质资源数据库中信息字段等,构建水稻种质资源及其特性的本体,相关本体片段如图 3 所示。

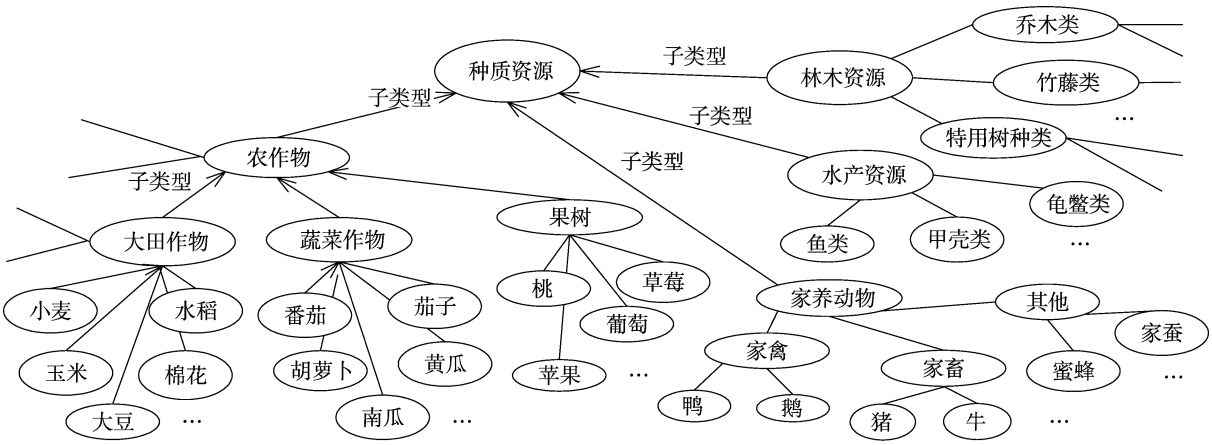

图2　农业种质资源本体片段(种质类型)

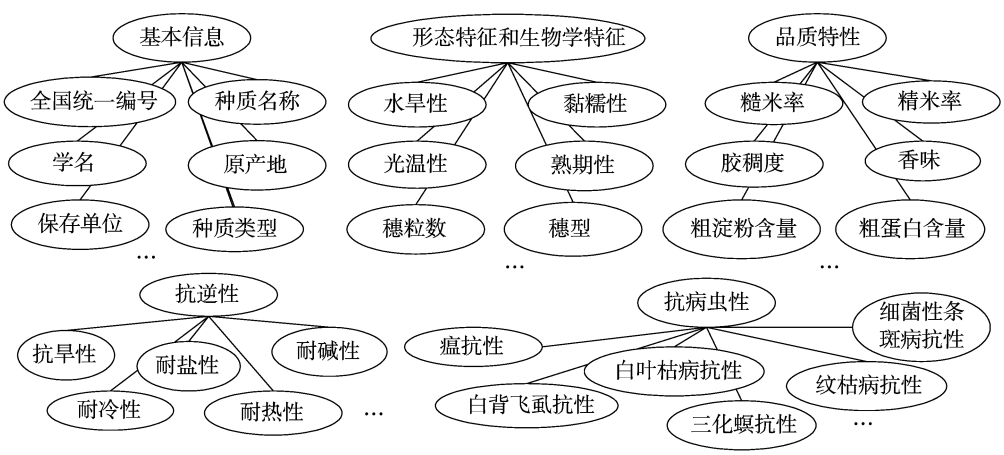

图3　水稻种质资源相关本体片段

本体构建是一个长期的过程,随着种质资源数据库的不断完善、相关专家的长期协同工作,系统中种质资源本体的定义也在不断改进和完善。

4　系统业务流程

江苏省农业种质资源智能检索系统的主要业务流程如下:

4.1　资源处理和语义标注流程

资源处理和语义标注是对各类种质资源数据进行语义化描述处理,将本体中的知识点和数据资源之间建立关联关系,最终将标注信息保存到语义元数据库中。系统中的语义标注流程如图 4 所示。

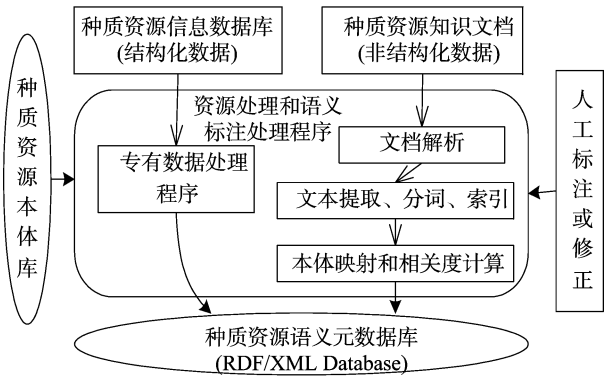

图4　种质资源资源处理和语义标注流程

种质资源信息数据库中的各种关系型数据库表,例如水稻基本信息表、共性数据表、特性数据表等,这些数据比较规整,数据列字段有着明确的定义和值约束范围。结合种质资源本体概念,可以开发专门的数据处理程序,生成最终标注结果保存到语义元数据库中。系统中还存在一些知识库形式的非结构化数据,一般经过文档解析、文本提取、分词、关键词抽取、词频统计、索引建立、本体映射和相关度计算等步骤生成标注结果。对于非结构化数据自动化标注存在准确性问题,系统中支持适当的人工干预和修正。

4.2　智能检索处理流程

采用自然语言形式进行信息查询符合用户的日常习惯,本系统的智能检索流程如图 5 所示。

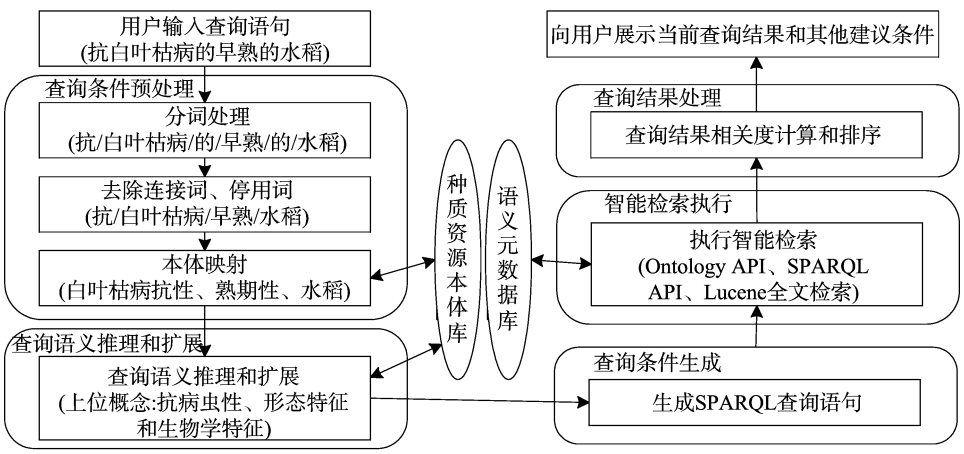

图5　种质资源智能检索处理流程

系统支持各种自然语言查询语句,例如“抗白叶枯病的早熟的水稻”、“晚熟的糯稻”、“抗病水稻”等。系统首先进行分词、去除连接词和停用词处理,获取关键词汇。然后进行本体映射处理,将相关词汇映射到种质资源本体的知识概念,例如将“抗/白叶枯病”映射到“白叶枯病抗性”的本体知识概念。接着进行查询语义推理和扩展,例如用户查询“白叶枯病抗性”还可以扩展到上位概念“抗病虫性”等,对于同位概念可直接扩展到查询条件中,增大查询结果的查全率,对于上下位概念,可以给用户提供查询建议,方便用户扩大或缩小查询范围,较为准确地检索到想要的结果。系统将扩展后的查询条件转换为 SPARQL 查询语句,利用 Jena 开源项目提供的 Ontology API、SPARQL API 和 Lucene 全文检索模块在种质资源语义元数据库中执行语义查询,并对查询结果进行相关度计算排序后,通过 UI 界面展示给用户,从而实现对种质资源信息的语义层面的智能检索。

5　总结

江苏省农业种质资源智能检索系统能在语义层面上理解用户的检索需求并实现一定的知识推理能力,提高了系统的查准率和查全率,增强了系统的易用性。本系统的建设可以使用户更方便快捷地检索到所需种质资源数据,为科学研究和农业生产过程提供丰富的决策参考信息,促进了资源信息的共享,提高了农业种质资源信息的利用效率。

参考文献:

[1]杨　欣,张　勇,林　静,等. 江苏农业种质资源信息服务系统的设计与构建[J]. 农业网络信息,2008(6):27 – 30.

[2]陈叶旺. 国家农业本体协同建构与语义检索若干技术研究[D]. 上海:复旦大学,2009.

[3]丁政建,张　路. 基于本体的语义检索研究[C]. 全国第 20 届计算机技术与应用学术会议(CACIS · 2009)暨全国第 1 届安全关键技术与应用学术会议,南宁,2009:246 – 250.

[4]李　景. 本体理论及在农业文献检索系统中的应用研究[D]. 北京:中国科学院,2004.

[5]国家水稻数据中心. 本体系统[EB/OL]. [2015 – 05 – 20]. http://www.ricedata.cn/ontology.