

基于本体的农作物种质资源平台用户模型研究

陈丽娜^{1,2}, 方 涛¹, 潘 恺¹, 司海平³, 曹永生¹

(¹ 中国农业科学院作物科学研究所, 北京 100081; ² 商丘师范学院计算机与信息技术学院, 商丘 476000; ³ 河南农业大学信息与管理科学学院, 郑州 450002)

摘要: 针对国家农作物种质资源平台的特点, 提出了基于本体的分组分层动态用户模型构建方法。首先建立用户结点, 由用户注册信息结合领域本体构建初步的分组用户模型; 然后挖掘用户日志扩展、定期更新用户模型; 最后利用通用本体 WordNet 进行同义词、不同语种的语义扩展, 进一步优化用户模型。利用开源软件 protégé 构建了农作物种质资源核心本体, 添加个人信息映射构成用户模型, 计算用户对资源的兴趣度, 并增加了时间因素来表明用户的兴趣随时间发生变化的情况, 便于用户模型的动态更新。基于本体的分组分层动态用户模型可以给不同用户提供个性化服务, 有利于提高平台的服务质量。

关键词: 种质资源; 个性化服务; 本体; 用户模型; 分组分层

User Model Research of National Crop Germplasm Resources Infrastructure Based on Ontology

CHEN Li-na^{1,2}, FANG Wei¹, PAN Kai¹, SI Hai-ping³, CAO Yong-sheng¹

(¹ Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081;

² Department of Computer and Information Technology, Shangqiu Normal College, Shangqiu 476000;

³ College of Information and Management Science, Henan Agricultural University, Zhengzhou 450002)

Abstract: A method called dynamic multi-group and multi-level based on ontology was proposed to build user model according to national crop germplasm resources infrastructure. A user node was first created, and an initial multi-group user model was built on the basis of user registration information and domain ontology. Then site log was mined to extend and update the user model regularly. Finally synonyms extension and different languages semantic expansion was implemented using common ontology WordNet, so as to further optimize the user model. The core domain ontology of national crop germplasm resources infrastructure was built under the environment of open source software protégé. User information was added to the ontology and formed the initial user model. User interest degree to resources was calculated, and time factor was added to the user model to indicate the interest change of users, so as to dynamically update the user model. The user model of dynamic multi-group and multi-level based on ontology could provide personalized services to different users, and could help improve service quality of national crop germplasm resources infrastructure.

Key words: germplasm resources; personalized services; ontology; user model; multi-group and multi-level

国家农作物种质资源平台是国家科技基础条件平台的重要组成部分。自建立以来, 已收集了 200 多种作物、42 万份作物种质信息, 拥有近 700 个数

据库、135 万条记录、230GB 数据量, 建立了农作物种质资源技术规范体系以及作物种质资源信息系统, 为科研和教学单位人员、生物技术研究人员、育

收稿日期: 2015-01-07 修回日期: 2015-02-01 网络出版日期: 2015-08-04

URL: <http://www.cnki.net/kcms/detail/11.4996.S.20150804.1328.046.html>

基金项目: 国家农作物种质资源平台(2005DKA21001); 河南省科技厅项目(142300410395); 国家星火计划引导项目(2014GA750011)

第一作者研究方向为信息系统管理, 科学计算可视化等。E-mail: chenlina@cgris.org

通信作者: 曹永生, 研究方向为信息系统管理, 数据挖掘等。E-mail: caoyongsheng@caas.cn

种家、农技推广人员、农民以及学生等各种公益性服务,取得了良好的服务效果^[1]。但随着平台数据量的增加和用户的增多,用户对信息系统服务的要求也越来越高。如何满足用户从海量信息中快速获得有效信息,如何根据用户需求提供个性化服务是当前迫切需要解决的问题。

要想实现个性化服务,需要将用户根据需求分成不同类型,为每种类型的用户建立用户模型,用户模型的构建是个性化服务的关键。传统的用户建模主要是基于关键词,该方法无法有效提取和检索到语义间的关联信息和隐含内容,不利于信息的查全。由于本体可以表达复杂的概念关系,能很好地反映用户的兴趣特征,所以基于本体的用户建模方法可以更好地反映用户真实需求,实现信息的重用和共享^[2-3]。目前国内外已有很多基于本体的建模方法,有些用户模型是静态的,不能反映用户兴趣随着时间的不同而发生的变化;有些是针对特定领域,比如数学领域进行的用户建模;还有些用户模型在构建时没考虑到概念之间的相似性,使得所构建的用户模型与个性化推荐脱节^[4]。

为了进一步深化平台运行服务,根据国家农作物种质资源的特点,提出了基于本体的分组分层动态用户模型构建方法,并基于国家农作物种质资源平台构建了用户模型。该模型既考虑了不同分组用户的需求,又可以根据时间动态更新,并通过用户日志和通用本体对用户模型进行扩充,向用户提供个性化服务,同时还可以实现一站式检索,进一步提高平台服务质量。

1 用户模型和本体

1.1 用户模型

用户模型是针对目标群体真实特征的勾勒,是真实用户的综合原型,它是对用户某时间段内相对稳定的个性行为和信息需求的描述。构建恰当的用户模型是个性化服务的基础和核心,它的质量直接关系到个性化服务的质量。

目前的用户建模方法分为纯手工建模、示例用户建模和自动用户建模。手工建模方法比较简单,但是全程需要手工输入,工作量大且效果完全依赖于用户模型;示例用户建模是提供与用户相关的示例来进行建模,该方法增加了额外的负担;自动用户建模是根据用户的历史行为自动构建,该方法构建的模型比较准确,但是用户历史行为难以把握,构建难度比较大^[5]。根据国家农作物种质资源平台的

特点,平台用户可以根据所访问的资源类别分为不同分组,每一分组用户常常只关注该资源的信息,包括该资源的种质信息、种业信息、科普知识、保存信息等,而且用户信息在一定时间内较少发生改变,所以这里首先根据分组构建初始的用户模型。

用户模型分为静态模型和动态模型。静态模型比较准确,但是不能够很好地反映用户的真实意图;动态模型反映了用户最近时间内的需求,但是构建难度较大。由于平台用户的需求一般比较固定,较长时间段才可能发生改变,所以可以根据用户日志对初始用户模型定期更新。这样既减少工作量,又能一定程度上反映用户需求。

用户模型的构建是个逐步求精的过程,包括显式或隐式地获取用户信息、对信息的更新、扩展等。为了使模型能满足用户更高层次的需求,可以通过用户日志和通用本体对用户模型进行扩展更新。

1.2 本体

本体的一个普遍被接受的定义是:“本体是共享概念模型的明确的形式化规范说明”,它是领域知识的归纳和形式化,使用它可以更好地描述信息以及信息之间的关系,实现信息的共享和重用。通过本体,人机之间以及机器之间就可以交流^[6-7]。本体通常包括概念、关系、函数、公理和实例,目前可以使用多种本体构建工具来构建本体。

本体在知识工程、信息检索以及自然语言处理等领域被广泛使用。近年来,在联合国粮农组织(FAO)的倡导下,出现了包括渔业、植物、食品安全以及基因等各种农业本体。但是本体的构建是一项非常大的工程,有关农作物种质资源本体还没有。所以,根据国家农作物种质资源平台的特点,构建农作物种质资源本体,生成适合平台服务的用户模型是对本体应用领域的一次尝试。

2 用户模型构建方法

基于本体的分组分层动态用户模型的构建过程主要分为3步,即构建初始用户模型,用户模型的更新和扩展。在构建用户模型前,先要建立该领域的核心本体。通常本体可以表示为树形结构,树的每一个节点表示为用户的一个兴趣点,然后重用该领域的核心本体,并根据用户的不同兴趣赋予不同的权值,构成用户本体,将用户本体映射到用户结点构成用户模型^[8]。

2.1 构建初始用户模型

国家农作物种质资源平台通过种质资源信息系统对外提供信息及实物共享服务。信息服务是指用

户通过系统查询种质信息、种业信息、科普知识、种质保存信息以及进行信息交流等。实物共享主要是根据用户的需求,无偿或者有偿地提供种质资源,实现信息共享到实物共享^[9-10]。虽然用户在平台上查询信息不需要注册,但订购信息的用户都留下了详细的个人信息以及索取资源的编号和数目,为我们详细分析用户资料,初步建立用户模型提供了便利。这里的订购信息可以认为是注册信息。

用户模型可以表示为:

$$\text{UserModel} = \{ \text{Ugroup}, \text{Uinfo}, \text{Rinfo}, \text{I} \},$$

其中 Ugroup 代表用户分组。由于不同用户所关注的资源类别不同,且一般很少发生变化,比如研究人员的研究方向一般是固定的,种子公司一般也只关注几种作物,所以按资源类别分组,分为粮食作物 Crop,纤维作物 Fiber,油料作物 Oil,蔬菜 Vegetable,果树 Fruit 和其他 Other,每一类用户只关注自己所属领域的信息。

$$\text{Uinfo} = \{ \text{Uname}, \text{Uid}, \text{Udept}, \text{Uphone}, \text{Uaddr}, \text{Umail} \},$$

代表用户信息,基本和种质资源信息系统的订购信息保持一致,包括用户名、身份证号、单位、电话、地址和邮箱等。由于种质资源平台有 56 个子平台,订购用户直接通过子平台订购所需资源,所以这里对各子平台订购信息进行统计,得到各类资源的用户信息。

$\text{Rinfo} = \{ \text{C}, \text{R}, \text{S} \}$,代表资源信息,包括概念集合、概念之间的语义关系及概念相似度。该部分可以由领域本体映射而得到。这里引入概念相似度来解决用户输入的关键词和用户模型中关键词不一致的情况。通过概念相似度的计算,找到与输入关键词语义相同或者相近的关键词,返回结果。比如用户搜索的关键词为“果树”和“果树图片”,通过概念相似度的计算,都把它们映射到果树 Fruit 类上。由于本体可以看作概念树的结构,两个概念的相似度可以映射为树中两个节点的相似度,因此概念相似度可以用节点的距离来表示,计算公式为:

$$S(C_i, C_j) = \frac{2 \times \text{depth}(LGA(C_i, C_j))}{\text{depth}(C_i) + \text{depth}(C_j)},$$

其中 C_i, C_j 为任意两个概念, $LGA(C_i, C_j)$ 为 C_i, C_j 最近共同祖先, $\text{depth}(C_i)$ 表示概念 C_i 在本体树中的深度。

I 为用户兴趣度,是用户对概念 C_i 的兴趣度的集合。可以用对该概念的访问频率来计算。对平台订购信息进行统计时,可以计算出每类资源被访问的频率,初步得到用户对该资源的兴趣度,表示为:

$$I = \{ \sum_{i=1}^n I(C_i) \} = \{ \sum_{i=1}^n \text{Frequency}(C_i) \}.$$

2.2 利用用户日志更新、扩展用户模型

初始用户模型是静态的,所包括的资源种类也不完全,实际用户的需求总是随着时间而发生变化的。为了使用户模型更好地反映实际需求,需要对初始用户模型进行扩展、更新。

通常用户日志中包含了记录用户搜索行为的大量信息,是用户意图的真实写照,可以更好地反映用户的真实需求。从搜索日志中可以提取出关键词、资源 URL、对资源点击顺序、对该资源访问的绝对时间(时间和文档长度的比率)、用户对资源的点击顺序等^[11]。国家农作物种质资源平台用户日志中包含了所有访问过网站的用户信息,而不仅仅是订购过资源的用户信息。因此,可根据用户日志扩展本体,在用户模型中加入新的资源,并通过用户访问日志的时间信息来更新本体。用户兴趣度可计算为:

$$I = I_{\text{new}} \times \frac{M_{\text{date}}}{M} + I_{\text{old}} \left(1 - \frac{M_{\text{date}}}{M} \right),$$

I_{new} 是由用户日志得到的新的兴趣度, I_{old} 是旧的兴趣度, M_{date} 是用户上一次访问系统到当前的时间间隔, M 是调整参数,即用户的兴趣度由新旧兴趣度按一定的影响因子计算而得到。扩展更新后的用户模型具有动态特性,更好地体现了用户需求。

2.3 利用通用本体进行语义扩展

由订购信息和用户日志得到的用户模型基本能满足个性化服务的要求。但在我国农作物领域,一种作物在不同地域甚至同一地域经常会有不同叫法,比如西红柿又叫番茄,甘薯又叫红薯、番薯,玉米又叫苞米、包谷、棒子等。因为用户不同,语义搜索时所依赖的关键词也不同,因此需要进行同义词的扩展。本研究利用通用本体 WordNet 来进行同义词扩展。WordNet 是普林斯顿大学认知科学实验室开发的一部在线的英语词典。它是将单词按照意义组成一个“单词的网络”,根据不同的词性组织为同义词的集合,每一个集合表示一个基本的词汇概念,且具有强大的表达词汇关系的能力。使用它建立同义词的映射进行扩展,可以提高搜索的查全率^[12]。比如, corn 在英文 WordNet 中可以如下映射: corn-maize-indian corn-zea mays。

另外,如果想实现跨语种的一站式检索,查询不同数据库或者网站中同一概念的信息,还需要进行不同语种间词汇的映射。目前 WordNet 的版本有中文、英文和日文等^[13],比如,玉米中文、英文、日文的映射关系为:玉米-corn-トウモロコシ。

综上所述,分组分层动态用户模型的构建流程如图 1 所示:

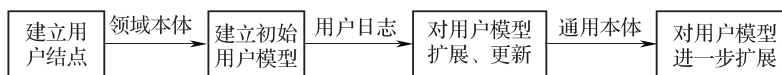


图1 用户模型构建流程

Fig.1 Process of building user model

3 国家农作物种质资源平台用户模型构建

3.1 建模工具

Protégé 软件是斯坦福大学基于 Java 语言开发的一款开源的本地开发工具,用于语义网中本体的构建。它提供了本体概念类、关系、属性和实例的构建,并且屏蔽了具体的本体描述语言,用户只需在概念层次上进行领域本体模型的构建。它提供了可扩展的独立平台环境,用于构建和编辑本体以及知识库,提供本体建设的基本功能,采用图形化界面,以树形的等级体系结构来显示本体,是目前使用最广泛的本地编辑工具之一^[14-15]。本论文采用 Protégé 4.1 构建本体,采用插件 Graphviz 来显示本体间的关系。

3.2 构建领域核心本体

根据分组分层动态用户模型构建方法,初始的

用户模型由分组信息、用户信息以及领域本体信息组成。所以,用户模型构建的前提是构建领域本体。由于领域本体是一项非常大的工程,本研究只建立领域的核心本体,开展原型实验,验证模型构建思路的可行性。后期可以对本体进行扩充。

本研究使用 2010. 1. 1 – 2012. 12. 31 两年间用户的订购信息来计算用户对概念的初始兴趣度,将资源分为 6 大类,不属于前 5 类的资源统统归属于 Other 类。各类资源的初始兴趣度分别为: $I(\text{Crop}) = 0.23$, $I(\text{Fiber}) = 0.20$, $I(\text{Oil}) = 0.16$, $I(\text{Vegetable}) = 0.12$, $I(\text{Fruit}) = 0.11$, $I(\text{Other}) = 0.08$ 。通过对系统资源进行分组,由用户信息、领域本体映射可以得到初始的用户模型。使用可视化插件 Graphviz 来显示本体间的关系,所构建的本体树形图如图 2 所示。

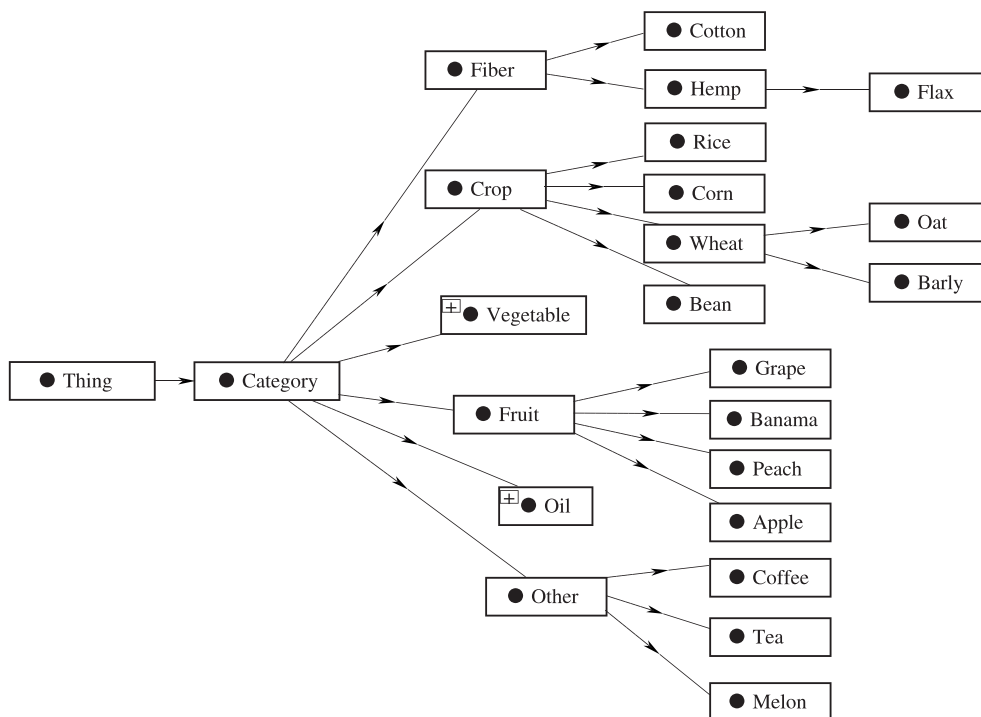


图2 核心领域本体

Fig.2 Core domain ontology

3.3 对初始用户模型扩展更新

根据初始用户模型,利用用户日志对其扩展更新。分析了国家农作物种质资源平台从 2013. 01.

01 – 2014. 12. 18 近两年间的用户日志,只关注日志中用户访问的关键词信息 and 对该资源访问的时间信息。如果在初始用户模型中没有该关键词,则加入

到用户模型中,并计算概念相似度以及兴趣度。如果初始用户模型中已经存在该关键词,则计算出新的兴趣度,由初始兴趣度和新兴趣度得到用户对该资源的最终兴趣度,这样就扩展、更新了用户模型。

通过统计工具分析,可以得到该段时间内资源关键词的访问情况,如图 3 所示(其中统计工具显示的有一些属于乱码的干扰词,已经修正),对关键词进行归类汇总。因为用户的请求千差万别,在初始用户模型中没有该关键词,所以只加入日志中访问频率比较高的关键词进行扩展。比如,这里加入板栗、桑树等(方框中显示的是扩展的资源),并计算兴趣度。对于初始用户模型中已经存在的关键词,重新计算兴趣度,比如葡萄、黄瓜、甜瓜等,使用兴趣度公式计算,可得到用户最终的兴趣度为: $I(\text{Crop}) = 0.26$, $I(\text{Fiber}) = 0.14$, $I(\text{Oil}) = 0.13$, $I(\text{Vegetable}) = 0.14$, $I(\text{Fruit}) = 0.24$, $I(\text{Other}) = 0.09$,最终得到的用户模型如图 4 所示。

由此可见,通过一段时间内的用户日志,动态更新了用户的兴趣度。其中粮食作物 Crop、蔬菜 Vegetable 以及水果 Fruit 类用户兴趣度有了较大的提高。

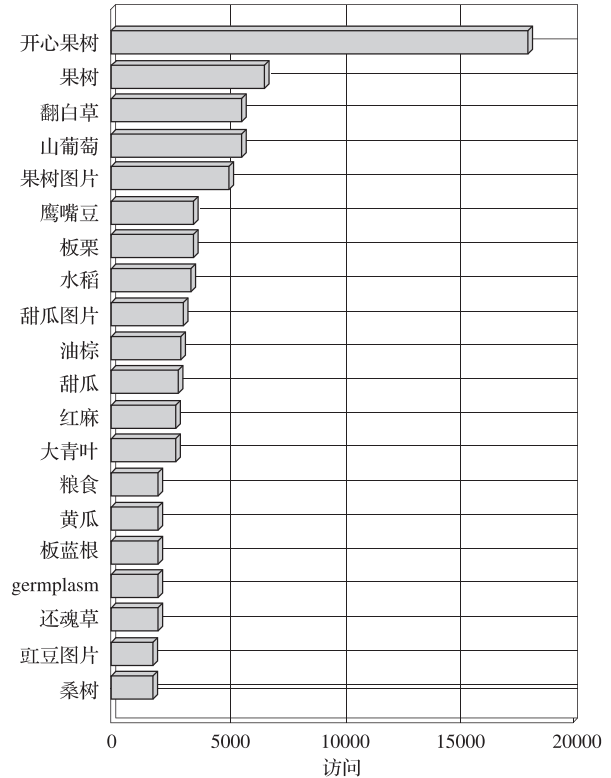


图 3 用户日志关键词搜索图

Fig.3 Keyword search map of users log

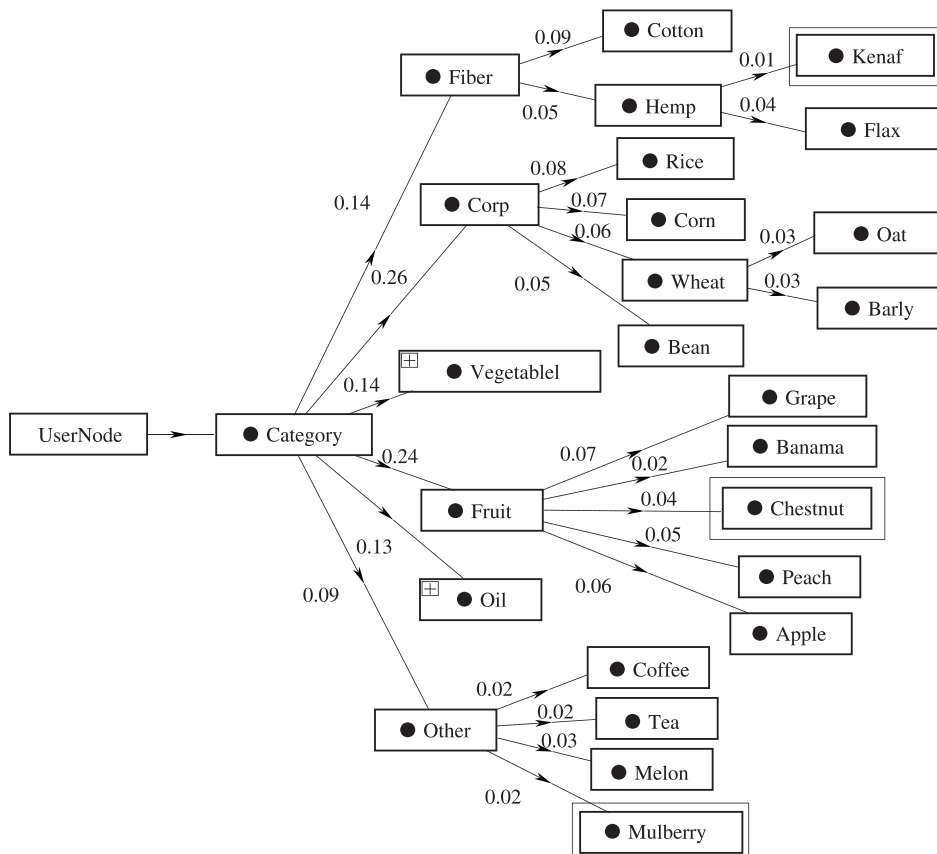


图 4 用户模型树形结构图

Fig.4 Tree structure chart of user model

最后再使用 WordNet 进行同义词和不同语种间的语义扩展。具体方法是将用户模型中的关键词分别输入到不同语种的 WordNet 中,找到它们在该词典中的同义词关系,加入到用户本体中,用户模型也因此进行了扩展。

4 结论

本研究针对国家农作物种质资源平台用户个性化服务的需求,提出了基于本体的分组分层动态用户模型。该模型先基于不同群体用户进行分组,由注册信息建立初始用户模型,然后根据用户日志对该模型进行扩展更新,最后再利用通用本体进行同义词、不同语言的扩展,满足了个性化服务的要求。

在构建该用户模型时考虑了 3 个关键因素,第一是引入概念相似度,计算概念相似度找到最接近的关键词,据此返回最接近的结果,有效解决了用户输入关键词和资源不一致的矛盾;第二是通过用户日志定期更新用户模型,既考虑了用户原有兴趣度,又考虑了用户的兴趣随时间变化的因素,定期更新的方式避免了经常更新造成的计算工作量大的问题,又符合平台用户的特点;最后是使用通用本体 WordNet 对用户模型进行扩展,包括同义词和不同语义的扩展,使得用户模型能满足更高层次的需求。基于本体的分组分层动态用户模型构建是对国家农作物种质资源平台服务模式的有效尝试,为进一步深化平台服务进行了有益的探索。

参考文献

[1] 曹永生,方涛. 国家农作物种质资源平台的建立和应用[J].

- 生物多样性,2010,18(5):455-460
- [2] Razmerita L. An ontology-based framework for modeling user behavior-A case study in knowledge management [J]. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2011, 41(4): 772-783
- [3] Viviani M, Bennani N, Egyed-Zsigmond E. A survey on user modeling in multi-application environments[C]// Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services (CENTRIC). Nice, France: IEEE, 2010: 111-116
- [4] 李艳. 基于兴趣学习的 Web 内容推荐及其优化研究[D]. 武汉: 华中科技大学, 2012
- [5] 符红光, 刘莉, 钟秀琴, 等. 基于 WordNet 和 Wikipedia 的平面几何本体的构建[J]. 电子科技大学学报, 2014, 43(4): 575-580
- [6] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods[J]. Data Knowl Eng, 1998, 25(1): 161-197
- [7] Jiang X, Tan A H. Learning and inferencing in user ontology for personalized Semantic Web search [J]. Inform Sci, 2009, 179(16): 2794-2808
- [8] Pretschner A, Gauch S. Ontology based personalized search[C]// Tools with Artificial Intelligence. Chicago, Illinois: IEEE, 1999: 391-398
- [9] 司海平, 方涛, 唐鹏, 等. 基于 SOA 的农作物种质资源调查信息系统研究[J], 植物遗传资源学报, 2010, 11(5): 517-521
- [10] 司海平, 刘俊辉, 马新明, 等. 农作物种质资源调查数据标准制定与共享[J], 植物遗传资源学报, 2012, 13(5): 704-708
- [11] Mimi A, Bahloul S N. Protect user anonymity in query log[C]// Machine and Web Intelligence (ICMWI). Algiers, Algeria: IEEE, 2010: 421-425
- [12] Chen S, Du Y, Peng Q. Extracting Query Expansion Terms Based on User's Search Behavior[C]// Computational Intelligence and Design. Changsha, China: IEEE, 2009, 1: 268-271
- [13] Chen R C, Bau C T, Yeh C J. Merging domain ontologies based on the WordNet system and Fuzzy Formal Concept Analysis techniques[J]. Appl Soft Comput, 2011, 11(2): 1908-1923
- [14] Wanberg C R, Kammeyer-Mueller J, Marchese M. Mentor and protégé predictors and outcomes of mentoring in a formal mentoring program[J]. J Vocat Behav, 2006, 69(3): 410-423
- [15] 高小龙, 朱信忠, 赵建民, 等. 电影本体的构建与一致性分析[J]. 计算机应用, 2014, 34(8): 2192-2196, 2201

欢迎订阅 2016 年《果树学报》

《果树学报》是中国农业科学院郑州果树研究所主办的国家级学术期刊,分别被有关权威期刊评价机构评为中国精品科技期刊、中国农林水产类权威学术期刊、中文园艺学核心期刊、中国科技核心期刊,已被中国科学引文数据库来源期刊(核心库)、美国化学文摘(CA)、日本科学技术振兴机构数据库(JST)、英国 CABI 等 20 余种国内外重要数据库收录。

本刊栏目设置有种质资源·遗传育种·分子生物学、栽培·生理·生态、植物保护·果品质量与安全、贮藏·加工、专论与综述、技术与方法、新品种选育报告等。读者对象为果树学科的科研人员、高等农业院校师生及基层果树管理技术人员。

本刊从 2016 年起改为月刊,定价 20 元,全年 240 元。邮发代号:36-93,国际代号 BM/1107。欢迎投稿,欢迎订阅。

地址:中国农业科学院郑州果树研究所

邮编:450009

电话:0371-63387308(兼传真)

E-mail: guoshuxuebao@caas.cn; chinagxb@163.com

网址: http://gskk.cbpt.cnki.net