# 基于本体的农业数据语义关联发现技术*

徐晓文,陈维斌,李海波
(华侨大学 计算机科学与技术学院,福建 厦门 361000)

摘 要:提出了基于本体的语义关联发现模型,通过解析构建的农业领域本体,从本体语义路径的深度广度方面计算概念间相关度,并将计算的结果扩充语义知识库。在农业领域模型中关联发现算法的应用与传统的方法相比,结果更符合领域相关性。依据关联发现模型设计了一个茶叶语义检索系统,实验验证了该提出的模型的实用性和可行性。

关键词:本体;关联发现;语义检索;相关度;检索概念
中图分类号:TP391.4 文献标识码:A 文章编号:1674-7720(2012)19-0073-03

# Ontology based semantic relevant discovery in agriculture data

Xu Xiaowen, Chen Weibin, Li Haibo
(College of Computer Science and Technology, Huaqiao University, Xiamen 361000, China)

**Abstract**: This paper proposes the ontology based semantic relevant discovery model to solve these problems. By parsing the constructed agriculture domain ontology, we calculate the correlations from the semantic distance of the depth and breadth in ontology, it improves the semantic knowledge base. Compared with the traditional methods, the experimental data in agricultural sector shows that the novel model better corresponded with domain reality. We develop an ontology based tea semantic retrieval system in agriculture on the basis of this model. The realization of the system verifies the feasibility and practicability of the relevant discovery model.

**Key words**: ontology; relevant discovery; semantic retrieval; correlations; retrieval term

传统的检索大都是基于关键字的检索,这种检索只是在字面匹配上处理,不能充分表达语义信息,用户的满意度不高。针对这些问题,国内外一些学者提出采用语义检索的方法来解决。刘群等人提出对董振东先生创建的《知网》的研究,将每个词的语义用多维义原表示,从义原相似度的角度出发解决词语间相似性和相关性[1];翟裕忠等人在语义网检索方面开展了研究工作,开发了一个面向领域的语义搜索系统,该系统采用基于图的查询机制检索出与被检概念相关联的语义对象列表[2];田萱等人设计了一种计算语义相关度的模型,采用权重的思想描述概念间的联系程度,通过在不同领域本体中的实验证明,语义相关度计算在查询扩展方面有显著优点[3]。国内外对语义检索及语义关联发现技术的研究已逐步预热。本文主要研究基于农业本体的语义关联发现技术,从领域本体的角度出发,融入关联关系发现算法,实现较普通检索更合理的语义关联检索。

## 1 语义关联发现技术相关理论

### 1.1 本体

本体(ontology)源自哲学上的一个概念,关注的是存在的本质。斯坦福大学的 Gruber 最早给出本体的定义:“本体是大多数人认同、对概念体系的明确的、形式化的规范说明”[4]。W3C 推荐的 OWL 语言(Web Ontology Language, Web 本体语言)是用户可清晰编写、机器可理解的、用于描述本体的形式化语言。

本文结合斯坦福大学提出的七步法[5]和农业情报部编制的《农业科学叙词表》,对农业领域本体的构建过程描述如下:

(1)确定研究领域为农业,根据《农表》中的叙词及关系描述,抽取类及子类;

(2)定义类间的等同、等级和相关关系;

(3)定义类的属性和属性类型等;

*基金项目:福建省重大产学研项目(2010N5008),厦门市科技计划创新项目(3502Z20110013),泉州市科技计划项目(2011G5),华侨大学基本科研业务费专项基金(JB-ZR1147)

(4)采用 OWL 语言描述本体。

### 1.2 语义相关

基于本体的概念间的语义关系主要分为 3 种:父子关系、相等关系和相关关系。对于前 2 种关系,在构建本体的时候就可以定义;对于第 3 种关系,根据关联的紧密程度,又分为直接相关和间接相关。直接相关指本体中直接定义了概念间的关系,没有经过任何其他的概念;间接相关指某两概念在本体中虽然没有直接定义关系,却通过其他概念产生了关联。如在农业本体中,“茶”和“肥料”是其中定义的相关的两概念,“茶”和“产量”也是其中定义的相关的两概念,依据间接相关,“肥料”和“产量”也是相互关联的。

本文引用语义相关度来衡量概念之间的关联度。定义若两个概念没有任何关联,则其语义相关度为 0;反之若两个概念是完全相关的,则其语义相关度为 1;若两个概念存在一定的联系,但联系程度是未知的,定义其相关度的取值范围为(0,1)。

## 2 基于本体的语义关联发现技术

语义检索的目标在于关联发现,即从语义的角度发现与检索概念相关的概念。本文在传统检索模型的基础上,以语义相关度作为概念间相互关系的度量,提出了基于本体的语义关联发现模型。该模型依据语义关联发现算法发现相关概念,返回一系列满足条件的结果。

### 2.1 基于本体的语义关联发现模型

本文用一个四元组<Q、O、F、S(q,o)>表示语义关联发现模型,该模型结构如图 1 所示。

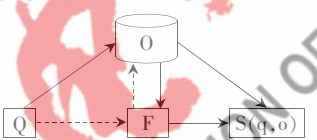


图 1 基于本体的语义关联发现模型

模型中各元素表示含义如下:

(1)Q 是查询入口,即用户输入的查询请求;

(2)O 是领域本体库,包含领域专家定义的本体中的任何信息;

(3)F 是构建 O 与 Q 关系的机制,即用于发现与检索概念相关概念的语义关联算法;

(4)S(q,o)是模型出口,是经由模型发现按相关度排序输出的相关概念。

### 2.2 基于本体的语义关联发现算法

#### 2.2.1 传统的概念相关度计算方法

传统的基于本体的概念相关度计算方法是以本体的结构信息为依据进行计算。

中国科学院软件研究所的朱礼军[6]提出,两个概念 $i$ 与 $j$ 之间的路径距离越大,则其相关度就越低,由式(1)可知:

$$d_{ij}=\frac{\alpha}{|D_{ij}|+\alpha} \tag{1}$$

其中,$d_{ij}$ 指概念 $i$ 与 $j$ 的相关度,$|D_{ij}|$表示概念 $i$ 与 $j$ 的路径距离,$\alpha$ 是当 $i$ 与 $j$ 相关度为 0.5 时的路径距离。

Rocha[7]等借鉴信息检索中常用的 IDF 方法,采用式(2)计算概念间的相关度:

$$d_{ij}=\frac{1}{\sqrt{D_{ij}}} \tag{2}$$

其中,$d_{ij}$ 是概念 $i$ 与 $j$ 的语义相关度,$D_{ij}$ 是概念 $i$ 与 $j$ 在本体层次树中的路径长度。

#### 2.2.2 本文提出的概念相关度计算方法

本文借鉴 Rocha 等人提出的概念间的长度越长,其关联程度则越弱,以及共有属性越多,相关度就越大的思想,改进了计算概念 A、B 的相关度的方法,如式(3):

$$Rel(A,B)=\frac{\sum_{i=1}^{n}W_{iAB}}{\frac{\sum_{i=1}^{n-1}W_{iA}}{\sqrt{D_{AB}}}} \tag{3}$$

其中,$Rel(A,B)$是检索概念 $A$ 与目标概念 $B$ 的语义相关度;$W_{AB}$ 代表概念 $A$、$B$ 共有的相关概念数(若相关,则增 1),$W_{iA}$ 是所有与概念 $A$ 相关的概念数,$D_{AB}$ 是检索概念 $A$ 和 $B$ 在本体定义中的路径长度(从 0 开始,若经过概念数增 1,则长度增 1)。

#### 2.2.3 语义关联发现算法

本文定义概念 $A$、$C$ 在以概念 $B$ 为条件下的相关度 $Rel(A,C)$计算如式(4):

$$Rel(A,C)=\frac{Rel(A,B)\times Rel(B,C)}{\sqrt{Rel(A,B)^2+Rel(B,C)^2}} \tag{4}$$

为了比较实验结果和清晰显示便于用户选择,需要对数据进行归一处理,方法如式(5):

$$Rel(M,i)=\frac{Rel'(M,i)}{\sum_{i=1}^{n}Rel'(M,i)} \tag{5}$$

其中,$Rel(M,i)$代表归一化后与概念 $M$ 相关的概念 $i$ 的语义相关度,$Rel'(M,i)$代表由式(3)或式(4)计算出的检索概念 $M$ 与概念 $i$ 的相关度。

语义关联发现算法处理步骤如下:

步骤(1):输入检索概念 $A$;

步骤(2):根据领域专家定义的领域知识库,得出与概念 $A$ 直接相关的概念 BList,及相关度 BValueList;

步骤(3):以检索到的概念 BList 为检索条件,继续查询知识库,得到与 BList 有关的概念 CList,及相关度 CValueList;

步骤(4):根据间接语义相关度计算方法,得出与检索概念 $A$ 有关的间接相关概念 CList 并计算修正相关度 CValueList;

步骤(5):查看是否有已知概念或参照概念,若有则输入已知概念 $D$,若无则跳入步骤(8);

步骤(6):由专家知识库,计算概念 $A$、$D$ 的相关度

大小 DValue；

步骤(7)：采用间接相关度计算方法得出在概念 $D$ 为参照的前提下，与检索概念 $A$ 相关的概念 DList，并将 DList 加入 CList 中，修正其参照后的相关度 CValueList；

步骤(8)：归一处理 BValueList、CValueList 与检索概念 $A$ 的相关度；

步骤(9)：由排序函数将相关概念按照相关度从大到小的顺序输出。

### 2.3 算法性能与实验结果比较

本文根据 1.1 节的方法构建了茶叶领域本体。选取朱礼军提出的路径距离计算方法(见式(1)，简称朱礼军法)、Rocha 提出的方法(见式(2)，简称 Rocha 法)进行实验参照对比。挑选 10 对概念，分别计算这 10 对概念的相关度，并将计算结果与传统的语义相关度计算方法比较，结果如图 2 所示。

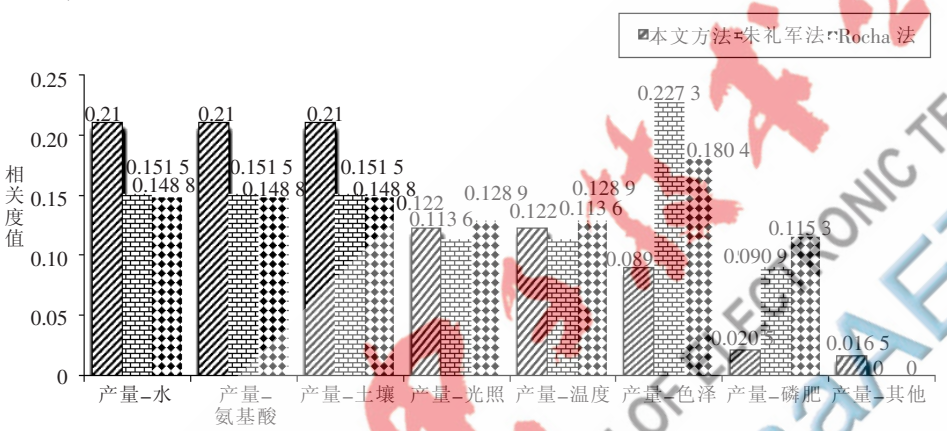


图 2 结果分析比较

通过图 2 在茶叶知识库中三种方法对以“产量”为检索目标的与其相关的一系列相关概念的对比发现，朱礼军法和 Rocha 法除了图表中列出的相关概念外，不能检索出其他的相关概念，而本文方法还能检索出 1.65%的其他概念，因此，本文方法在发现隐含资源方面，比传统方法有很大改进。

观察图 2 数据，本文方法与朱礼军法和 Rocha 法计算出的茶叶“产量”与“色泽”的相关度存在明显差异，它们计算出来的这个值均是与“产量”相关的最大值，这是有悖于专家经验的，且它们在综合计算与“产量”相关的概念时，除了“色泽”(与常理相悖的误差结果)，其他结果均无明显的差异，没有关系的权重的分配。在经验知识中，茶叶的“产量”与“水”存在很大关联，从数据对比显示说明本文方法计算出的结果不仅符合专家经验，而且各个相关值权重分配清晰明显。由结果可知，该方法符合领域相关性，计算得出的结果可用于完善领域知识库。

## 3 茶叶语义检索原型系统的实现

### 3.1 系统结构模型设计

采用 MVC(Model View Controller)设计模型的思想，将系统分为三层，分别为模型层、表现层和逻辑层。模型层是数据模型，包含本体知识库和用户信息库；表现层包含用户注册登录模块和检索结果的展示模块；运行层包含本体解析模块、关键词匹配模块以及语义关联发现模块。

### 3.2 系统运行结果

采用 Java 语言实现，调用开源工具包 jena，对 owl 语言描述的本体进行语义计算，检索系统如下。输入查询请求 Q：产量，如图 3 所示。

采用排序函数 S(q,o)输出与检索概念“产量”相关的概念，如图 4 所示。得出的结果是在茶叶领域，与“产量”相关的按照从大到小的顺序排列的一些概念及对应的相关度大小。结果表明，该语义检索系统能够将语义检索融入到语义 Web 的实际应用系统中，为各自的领域活动进行指导作用。

本文从本体语义深度及广度方面，提出了基于本体的语义关联发现模型，并将此模型应用于农业领域检索，改进了语义相关度计算方法。实验数据表明，新模型得到的结果与传统的计算相关度的方法得到的结果相比，更符合人们对领域的认识，结果更合理。但是还存在一些不足，如检索效率的提高和智能问答如何实现等，这些不足有待在后续工作中得到改进。

图 3 检索界面

图 4 检索结果

参考文献

[1] 刘群，李素建.基于《知网》的词汇语义相似度计算[J].中文计算语言学，2002，7(2)：59-76.

[2] 李景，孟宪学，苏晓路. 领域本体的构建方法与应用研究[M].北京：中国农业科学技术出版社，2009.

[3] TIAN X, DU X, LI H. Computing degree of association based on different semantic relationships [C]. Database and

Expert Systems Applications of 2007. DEXA 07.18th International Workshop. IEEE Press,2007.

[4] GRUBER T R. Toward principles for the design of ontologies used for knowledge sharing [J]. International Journal of Human Computer Studies,1995,43(5):907-928.

[5] NOY N F, MCGUINNESS D L. Ontology development 101: A guide to creating your first ontology [C]. Standford Knowledge Systems Laboratory Technical report KSL-01-05 and Standford Medical Informatics Technical Seport SMI-2001-0880,March 2001.

[6] 朱礼军, 陶兰, 刘慧. 领域本体中的概念相似度计算[J]. 华南理工大学学报(自然科学版).2004,32(11):148-149.

[7] ROCHA C, SCHWABE D, ARAGAO M P. A hybrid approach for searching in the semantic web[C]. Proceedings of the 13th International Conference on World Wide Web. ACM, 2004.

(收稿日期:2012-04-17)

作者简介:

徐晓文,女,1987 年生,硕士,主要研究方向:数据库技术及应用、数据挖掘技术。

陈维斌,男,1957 年生,教授,硕士生导师,主要研究方向:数据库技术及应用、面向对象技术。

李海波,男,1972 年生,副教授,主要研究方向:服务计算技术、工作流技术,软构件技术。