

一种基于文本知识库的推理方法研究

薛军 王儒敬

(中国科学院合肥智能机械研究所,合肥 230031)

E-mail:xuejun@mail.iim.ac.cn

摘要 首先介绍了一种新型的综合知识表示方法的基础上形成的知识库描述语言;接着提出了在智能化农业信息处理系统中,基于此描述语言的文本知识库的一种分析和推理方法;并以农业知识库为实例,分析了具体的分析方法和推理方法,同时还讨论了WEB环境下推理机的实现问题。

关键词 专家系统 知识库 语法分析 推理机

文章编号 1002-8331-(2003)21-0189-03 文献标识码 A 中图分类号 TP18

A Method of Reasoning Based on Text Knowledge-base

Xue Jun Wang Rujing

(Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031)

Abstract: In this paper we first introduce a script language based on a new synthetic method of knowledge representing. Then we present a method that can analyse and inference text knowledge-base described by above script language in Intelligent Agriculture Information System. Finally we discuss the particular program of analysis and reasoning in agriculture knowledge-base as well as the realization of this inference machine in WEB.

Keywords: Expert system, Knowledge base, Syntax analysis, Inference engine

1 概述

在新一代智能化农业信息处理系统中,如何对各种复杂农业知识进行灵活、准确、有效的表示并进行正确性与有效性校验成为系统平台开发中的一个关键问题。专家系统主要用于解决那些没有高效算法的非结构化问题,如农业领域的施肥决策和病虫害防治等,而推理则是专家系统解决问题的基本技术。因此知识表示和推理方法对专家系统解决实际问题的功能和效率起决定作用^[1,3]。

笔者在以往 863 课题的研究基础上,对基于文本的知识库的语法分析和推理解释方法做了一些改进和创新。首先是对知识表示方法的改进,采用有限知识单元表示法:在某个知识领域内,用知识单元来表示和描述该领域内的知识,知识单元按照知识的体系结构和功能模块进行划分,单元对象之间的关系表示了知识库的框架结构和知识的逻辑结构,有前提结论关系、一般特殊关系、组成部分关系等,知识结点之间构成为一个树或有向图。并在此知识表示策略的基础上形成了知识库描述语言。然后,对用此描述语言写成的文本知识库进行语法分析和检验,包括词法分析、语法分析和语法错误检查,以及知识的有效性检验。知识验证(knowledge verification)的目的是决定系统的正确性、完备性和一致性^[2]。并将检验后生成的语法树和其他相关信息按照其逻辑结构转换成中间结果,保存在中间数据库(知识库)中,作为进行下一步推理的知识库。

在推理解释时采用了对用户更为友好的交互方式:首先显示出该专家系统所有提供的功能模块的框架结构图,在用户选择了特定的功能模块后,根据中间数据库的记录,一次性向用

户提出所有需要用户提供的实变变量,用户可以根据情况提供所有或部分事实项。在推理过程中,用户可以随时中断和回溯推理过程。

采用描述语言来表达知识,并以文本的形式来存储知识具有以下优点:知识表达灵活丰富,逻辑性强,可以表示各种类型的知识,如多媒体知识,可以在知识库中调用外部自动化构件等。

2 知识表示策略

由于农业应用的复杂性和时变性,专家系统的开发需要建立包含多种丰富信息的大型农业综合知识库,这不仅要求知识表示方法能融合数据、规则、方法、模型、案例以及文本、图象、动画、超媒体、外部构件等,同时还要方便地实现知识的存储、管理以及冗余、矛盾等一致性校验。通过对框架、产生式规则、面向对象等现有知识表示方法作系统分析,引入知识单元的基本概念。

知识单元表示法是按照人类思维的习惯,将知识分成相对独立的知识单元对象,将它们从上到下、按照知识单元之间的控制关系组织成树状结构体系。知识单元按照其实现的特定功能进行划分,分为框架知识单元、规则知识单元等。在特定的知识领域内,当提供了足够多的知识描述单元,就能保证该领域知识库的准确性和完备性。在此基础上形成了文本知识库描述语言,文章完整定义了该描述语言的BNF范式,部分内容如下:

```
<知识库描述语言> ::= <知识单元序列>;
<知识单元序列> ::= <知识单元序列>; <知识单元> <知识单元>
<知识单元> ::= <框架知识单元> <控制知识单元> <规则知识单
```

基金项目:国家 863 高科技发展研究计划资助项目(编号:2001AA110464)

作者简介:薛军,男,硕士研究生,主要研究方向为智能信息处理系统,基于网络的推理技术。王儒敬,男,副研究员,主要研究方向为智能决策支持系统,智能化信息处理系统。

```

元><计算知识单元><外部知识单元>
<框架知识单元> ::= FKU <框架知识单元名> (<框架知识单元参数>)
<框架知识单元名> ::= 标识符
<框架知识单元参数> ::= <框架知识单元参数>, <知识单元名> <
知识单元名>
<知识单元名> ::= 标识符
.....

```

3 知识库分析系统设计

在基于多层知识单元知识表示策略下形成的知识库描述语言,具有较强的开放性与灵活性,能够表达复杂的人类经验和知识。现在考虑如何对此描述语言所表示的知识库进行语法分析以及语法错误检查,并将分析结果和相关数据信息按照原来的逻辑结构保存在数据库中,以便于后面的推理解释功能的实现。

3.1 扫描程序的设计

记号(token)定义:由扫描程序生成的逻辑单元,便于分析程序进行后面的逻辑处理,主要分为:保留字、特殊记号(如,+、<)和多字符串的记号(如标志符)。

自定义知识单元名、规则名、函数名、过程名等标识符的说明:标识符可以由汉字、字母、下划线、数字组成,必须以汉字、字母或下划线开头,字母不区分大小写,每个标志符的最大长度为 256 个字节。

记号的正则表达式:

3.1.1 数

可以是整数,实数和带有有指数的数(由 E 表示)的序列。

nat=[0-9]+

signedNat=(+|-)?nat

number=signedNat(“.”Nat)?(E signedNat)?

3.1.2 保留字和标志符

保留字=IF|THEN|INPUT|...

字母=[a-z][A-Z]+

数字=[0-9]+

汉字=[啊———](根据 ASCII 码值判断)

标识符=(字母|下划线|汉字)+(数字|字母|下划线|汉字)*

3.1.3 注释

采用 C 语言注释风格(/*.....*/)

根据以上正则表达式画出扫描程序的有穷自动机(DFA),再根据 DFA 写出扫描程序的代码实现。

3.2 语法分析和检验

采用自顶向下的递归下降语法分析算法。分析程序的任务是:以扫描程序产生的记号(Token)作为输入,来确定和分析程序的语法结构,并显式地构造出相应的分析树或语法树——作为第一遍扫描的结果。以后的分析处理将以语法树作为输入。

分析程序中的对错误的处理是较为复杂的问题,在扫描程序中如果遇到一个不符合词法的错误字符,那么只需要生成一个出错记号,并消耗掉这个字符即可。而在语法分析中,不仅要报告一个出错信息,而且必须从错误状态恢复,并继续进行分。错误恢复(Error Recover)的关键是要报告有意义的错误信

息以及在接近真正错误时继续分析下去^[9]。笔者在语法分析时采用的错误校正方法是标准的应急方式(panic mode)^[9]。

每个知识单元模块包含都有输入变量参数列表和输出变量参数列表,以及模块内部的功能实现语句,包括规则,计算语句,控制语句等。如规则知识单元的典型形式:

```

RKU 规则单元名(输入变量:类型;输出变量:类型)

```

```

{
  READ 外部变量;
  IF 前件 AND 前件 OR 前件 THEN 后件;
}

```

定义知识单元类型:

```

typedef enum{FrameUnit,RuleUnit,CalUnit,ControlUnit}UnitKind;

```

语法树的节点数据结构以 C 语言声明如下:

```

typedef struct tagUnitNode

```

```

{
  struct tagUnitNode* child; //结点的最左孩子
  struct tagUnitNode* sibling; //结点的右兄弟
  UnitKind kind; //知识单元类型
  TreeNode* Node; //知识单元内部模块的语法树
  VarRecord* pVarInput; //输入变量列表
  VarRecord* pVarOutput; //输出变量列表
}UnitType;

```

语法分析最后生成的语法树是一棵多叉树,中间结点为知识库中各种类型的知识单元模块,叶结点为单元内部的实现语句和表达式等。

3.3 从语法树到中间数据库的转换

从语法分析得到的语法树保存在数据库中,以便于后面推理解释,转换到数据库中是为了利用数据库存储的透明性和数据管理的强大功能,提高推理的效率。另外,知识录入界面中生成的文本知识库主要显示了框架结构和逻辑结构,系统用户能够根据需要对生成的中间知识库进行修改和补充,这样有利于系统的开放性和可扩展性。

用关系数据库存储中间结果主要保存知识单元之间的逻辑关系,规则,数据项和表达式等,为每种类型的知识单元建立一个数据库表,如对于框架知识单元可考虑建立如下的表结构:

结点编号	单元名称	左孩子结点编号	右兄弟结点编号	说明
------	------	---------	---------	----

为了实现基于 WEB 的远程推理,要使用数据库来保存用户相关信息,包括用户所在地等个人信息以及用户以前的访问内容,还可以保存用户的推理步骤和推理中间结果等。

4 推理系统的实现方法

推理技术几十年来已经有了很系统的发展,从早期的基于符号逻辑的确定性推理到后来的不确定性和模糊推理。在推理路径上有正向链推理和反向链推理,如在 PROLOG 语言中,推理向后进行,即从要建立的目标开始,顺着规则链,最后瞄准事实;而正向推理是推理向前进行,从事实开始,顺着推理链搜索,最后求解出目标。正向链系统中一般用标准的蕴涵式表示规则^[2,3]。

4.1 推理策略和冲突消解

由于推理树较深,在推理策略上采用反向链推理,即从要

求解的目标出发,搜索所有符合匹配模式的规则前件,对于仍然为假设的前件则继续搜索所支持的规则,直到搜索到知识库中的事实或者用户提供的事实。如果最后没有搜索到需要的事实且用户不能提供相应数据,则推理中断。搜索先在要求解的问题知识单元内部进行,然后在相关的规则单元搜索,这样搜索效率较高。

在推理控制策略上,由知识专家给每条规则一个置信度(CF),对于所有符合匹配要求的规则根据置信度的大小来决定规则匹配的优先级。另外,系统支持不确定推理,利用概率来确定证据的可信度。对于复杂规则,即规则的前件是多个证据,给每个证据一个可信度(Degree of belief),再通过模糊逻辑来组合证据^[9]。

解释功能是用来回答用户提出的各种问题,包括与系统推理有关的问题和与系统推理无关的关于系统自身的问题。它可以对推理路线和提问的含义给出必要的解释,为用户了解推理过程以及系统维护提供方便的手段。在推理过程中记录了用户的推理路径,这样用户可以随时中断和回溯推理路径,还可以随时改变推理路径。

4.2 推理构件实现方法

推理解释模块和前面的语法分析模块都按照 COM 规范来实现,一方面是符合智能化农业信息处理系统的体系结构要求,另一方面是便于实现基于 WEB 的远程推理决策。

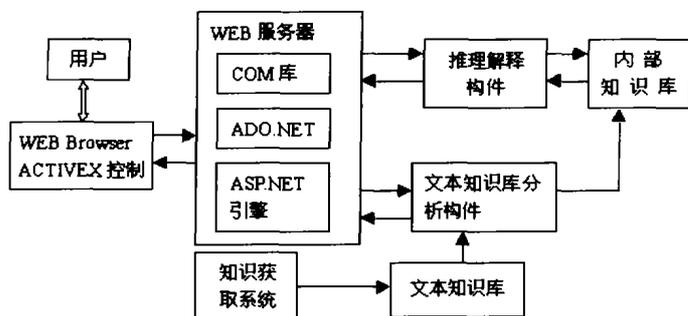


图1 智能化农业信息处理系统的控制结构

(上接 188 页)

的更新者(注:数据库在缺省条件下将一条记录看成一个组)。在字段分组的条件下数据库复制原理是:在对某个字段更新前先记录该记录中的键值属性值和被更新字段所在字段组中各字段的值,然后记录键值属性值和字段组中各字段的新值。在做复制操作时,以键值属性值和字段组中各字段的原值为依据,进行查找并更新。

4.4.4 使用加标记的删除方法解决数据库数据删除冲突问题

对于不会引起子实体数据一致性问题的记录,可以在第一时段加删除标记,在第二时段或以后时段对有删除标记的记录执行删除操作。因为任何人都不会对已加删除标记的记录进行修改操作,所以删除有删除标记的记录不会引起数据库数据删除冲突。所以需要在记录中加入删除标记字段。

5 结束语

对于分支机构多、地理范围广的大型企业,采用基于复制的分布式数据库系统是目前行之有效的解决方案。Oracle 的高

随着 Internet 网络的迅猛发展,广大农村基层科技推广部门通过访问 WEB 站点进行远程推理决策已成为当前的趋势,也有利于智能化农业信息处理系统的推广和应用。将知识库、数据库和模型库等数据量较大的部分放在服务器上,所有的数据库查询修改操作都放在服务器上,降低了客户端软硬件的要求。因此,该文所讨论的推理方法将基于网络的应用体系结构来实现。体系结构如图 1 所示。

5 结论与展望

由于农业知识的复杂性、地区差异大和丰富性等特点,如何简易、有效地表示农业知识并准确高效地进行推理决策始终是农业智能信息处理系统所要解决的主要问题。该文提出的知识单元基础上形成的知识库描述语言能灵活丰富地表示各种农业知识,推理方法简洁高效,和用户交互方式简单友好等,但知识库的容量受到一定限制,并且知识库如何与数据库和模型方法库有效地结合将是以后值得继续研究的问题。

(收稿日期:2002年9月)

参考文献

1. Nils J Nilsson. Artificial Intelligence: A New Synthesis[M]. Morgan Kaufmann Publishers Inc, 1998
2. R A Orchard. FuzzyCLIPS Version 6.04A[M]. Institute for Information Technology, NRC, 1998
3. Joseph Giarratano, Gary Riley. 专家系统原理与编程[M]. 北京:机械工业出版社, 2000
4. 李辉, 张宁, 许卓群等. 一种基于主特征的几何空间模糊推理方法[J]. 计算机研究与发展, 2001; 12
5. Kenneth C Loudon. 编译原理及实践[M]. 北京:机械工业出版社, 2000
6. 潘爱民. COM 原理与应用[M]. 北京:清华大学出版社, 1999
7. 李凡, 卢安, 饶勇. 智能系统中推理规则的约束优化问题求解[J]. 计算机工程与应用, 2002; 38(8): 43-45

级复制功能为此提供了强有力的支持。但是设计和实现一个分布式数据库系统特别是基于复制的分布式数据库系统是一项十分复杂的工作,在具体的工程实践中进行方案设计时要在网络连接方式、系统性能、安全性、可维护性及数据的一致性和完整性等多方面进行精心的考虑和设计。该文重点讨论了数据的一致性维护问题,所提出的数据一致性维护策略和解决各种数据冲突的方法经实践证明是合理有效的。

(收稿日期:2002年9月)

参考文献

1. 沈佩娟, 汤荷美. 数据库管理及应用开发[M]. 北京:清华大学出版社, 1995
2. 俞盘祥. Oracle 数据库系统基础[M]. 北京:清华大学出版社, 1995
3. William G Page 等. 王磊等译. Oracle 8/8i 开发使用手册[M]. 北京:机械工业出版社, 2000
4. Loney K 著. 李晓军译. Oracle8 数据库管理员手册[M]. 北京:机械工业出版社, 1999