

信息共享联盟模型研究及其在农业领域的应用^{*})

林乐宇¹ 刘磊^{2,3} 王石^{2,3} 郑宇飞² 丰强泽^{2,3} 曹存根²

(北京市首都师范大学计算机科学联合研究院 北京 100037)¹

(中国科学院计算技术研究所智能信息处理重点实验室 北京 100080)²

(中国科学院研究生院 北京 100039)³

摘要 本文提出了一种信息共享联盟模型 UIS,对模型的结构、功能框架以及相关核心技术,包括异构数据库的标准化处理、URI 远程访问技术、快速查询算法等进行了详细讨论。UIS 的一个重要特点是:在逻辑上将各个信息资源的异构数据连接在一起,在物理上则对加盟数据按照元数据规范标准进行整合,并备份上传至联盟。该模型通过多主体协作,能够有效地整合异构、分散的数据库,实现信息资源的有效共享和有效检索。模型服务提供了多种模式的信息检索接口。对于已经共享的多媒体资源,通过引入领域专家的知识问答命题来引导用户的学习,由此实现了教与学之间的互动,提高用户的学习质量。目前 UIS 已在中国农村科技信息网中进行实验和应用,实现了包括位于北京的“中国现代农业科技信息共享示范网”、“九亿网”、“国家农业科技园区网”等多家网站在内的数据库资源的信息共享,并且在河北省进行了试点,初步建立了中国农业信息共享联盟体系。

关键词 信息共享;信息检索;异构数据库;多主体;信息联盟

The Research on the Model of Union of Information Sharing and its Application in the Agriculture

LIN Le-Yu¹ LIU Lei^{2,3} WANG Shi^{2,3} ZHENG Yu-Fei² FENG Qiang-Ze^{2,3} CAO Cun-Gen²

(Joint Institute of Computer Science, Capital Normal University, Beijing 100037)¹

(Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)²

(Graduate University of the Chinese Academy of Sciences, Beijing 100039)³

Abstract In this paper, a model of UIS (Union of Information Sharing) is proposed, which can efficiently integrate and share heterogeneous databases geographically distributed and resided in different environments with the multi-agent cooperation. We present the structure and major components of this model, including standardization of heterogeneous databases, URI remote accessing, efficient query, and natural language interface. Based on the multimedia resources which have been shared, the model service introduces expertise's knowledge question-answer proposition to guide user's study. And the statistical analysis of the users' answer in their study will also react to the expertise's further proposition. As a result, this service will reach the interaction between the teaching and learning and finally enhance the user's study efficiency and the study quality. At present, the whole system is implemented and applied in the area of agricultural information sharing, and currently contains databases from several Websites.

Keywords Information sharing, Information retrieval, Heterogeneous databases, Multi-agent; Information union

1 引言

Internet 已经成为一个庞大的、广泛的、分布式的异构数据库和应用计算平台。它给人们提供了一个非常广阔的空间,帮助人们获取所需要的各种信息、资源。然而,由于信息形式的多样性和信息存储的异构性,网络资源的有效利用率总是很低^[1]。例如,就农业信息资源而言,我国当前拥有 100 多个大型涉农信息数据库,各种小型数据库更多,运行的涉农网站约 2600 多个^[2]。这些数据库中有相当一大部分数据的信息是相同的,而存储结构和表现形式却是相异的,具有分散、异构的特征^[3~5],不利于用户的集中查询。因此,如何更加有效的组织、整合各类分散信息,使用户能高效地搜索到所需资源,已经成为当今 Internet 发展的一个研究热点。

目前 Internet 中大部分的有效资源保存在数据库之中,而体现于 Web 页面之上。人们最常用到的 Web 搜索引擎也

大都是基于静态或动态 Web 页面的关键字搜索,无法直接面向数据库。这样的搜索具有相当的通用性,却缺乏必要的针对性和专业性。特别是当用户搜索某些专业领域资源(如农业科技信息)时,搜索到的结果中往往会存在着诸多完全不相关的信息。如果能够将这些分散的、异构的数据库数据按照某种统一格式进行标准整合与共享,会极大地提高资源利用率。

然而,实现信息的整合与共享并非一件简单的工作,涉及许多必须解决的设计和技术难题:

(1)各个基层数据源信息存储于多样的环境之中,包括不同类型、结构相异的数据库、操作系统等。另一方面,根据我们的调研,资源的所有者一般不愿意对他们的资源进行标准化处理,更不愿意按照目前流行的语义 Web 思路进行改造^[6,7]。因此,文[1]所提出的信息共享思路不具有普遍性。

(2)当基层数据源信息变更时,用于检索的逻辑信息集合

^{*})本文工作得到科技部研究专项“中国农村科技信息共享研究和示范网”(课题号 2002DEA30036、2003DEA2C014)、自然科学基金(60273019、60573064、60573063 和 60496326)和国家重点基础研究发展计划 2003CB317008 和 G1999032701 的资助。

也需要在最短的时间单位内得到相应的更新,从而保证用户总是能够检索到最新的资源。

(3)需要给用户提供人性化接口,使其能够高效准确地检索到共享信息,特别是在信息量不断膨胀的过程中仍然保持较高的检索效率。

除此之外,资源整合中存在的各种异常同样会给信息共享的实现带来相当的困难。

针对上述诸多问题,本文提出了一种信息共享联盟(Union of Information Sharing,简称 UIS)模型。该模型在用户与网络资源之间建立了一个中间层。该模型具有两个显著的特点:

第一,资源的所有者不需要对他们的资源进行标准化处理,只对加盟数据按照元数据规范标准进行逻辑上的虚拟整合,相关元数据备份上传至 UIS 联盟。

第二,UIS 模型引入了多个主体(Agent)来实现相关功能。在计算机和人工智能领域中,主体可以看作是一个实体,它通过传感器感知环境,通过效应器作用于环境^[8,9]。主体具有自治性、交互性、协作性、可通信性等特征。通过多主体的协作,一方面可以有效地整合异构、分布的数据库,实现信

息资源的有效共享;另一方面为用户提供了一个统一的、多功能的信息检索接口,使用户高效而准确地获取所需资源。

2 UIS 结构设计和工作流程

UIS 模型通过一套标准的信息共享协议,使信息资源所有者能在统一的信息共享平台下方方便地加入联盟,共享自己的资源。同时,为了屏蔽用户对外部信息源的直接访问,UIS 模型提供了一个对分布式数据访问的信息共享平台,能够真正实现全国各地孤立信息资源的有效整合和共享,最终形成一个有机的全国信息联盟体系。

从逻辑角度上讲,UIS 是在用户层与资源层之间构建的一个资源共享平台,主要由信息加盟、联盟共享和联盟服务三个模块组成,如图 1 所示。信息加盟模块首先接收资源提供者提供的各类资源,支持多种数据库和包括多媒体在内的文件格式,然后在 Web 服务器上通过信息集成主体实现对各类资源的标准化整合并上传至联盟。联盟共享模块接收并更新中央数据集数据和共享目录。最终用户可以使用多种通信终端来提交信息服务请求,通过信息服务主体以多种方式访问共享资源。

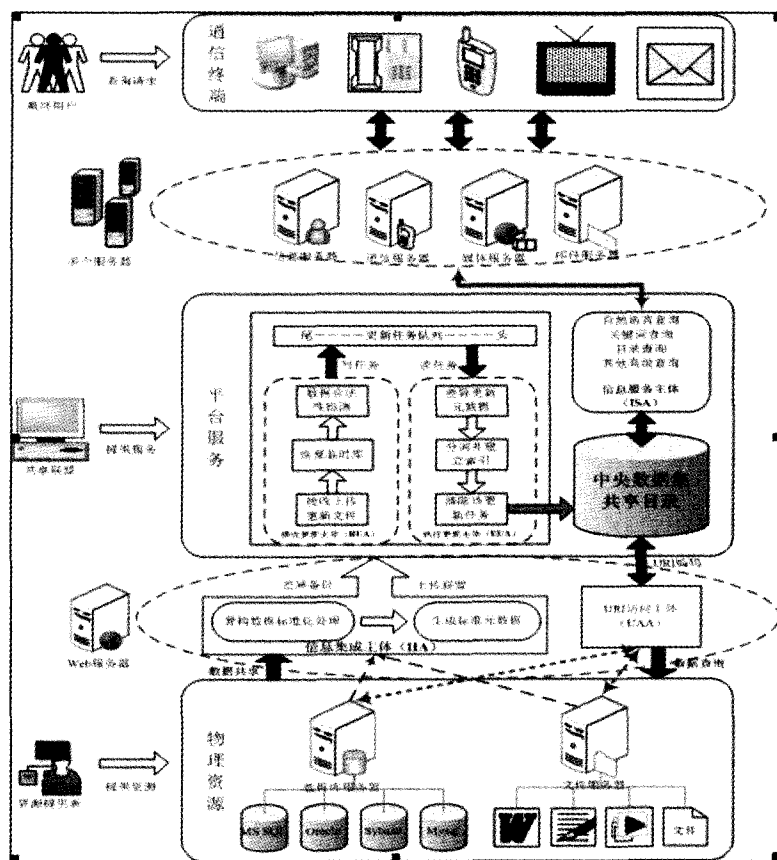


图 1 UIS 模型的结构

2.1 联盟 Agent

UIS 通过了多 Agent 的协作来有效实现异构数据的标准化处理、更新文件的收发、信息多样化查询和物理资源的远程访问。我们首先给出联盟 Agent 和其他 Agent 协作的描述。

定义 1(联盟 Agent) 联盟 Agent 是整个模型的基本构件,可抽象成一个 6 元组:

Agent(UIS) = (ID;
Type;

Interface {Domain_Interface, Communication_Interface};
Activity {Public_Activity, Exception_Activity};
Info_Blackboard;
Other_agent;
)

其中, ID 表示 Agent 的标识名称; Type 表示 Agent 的类

别;Interface 列举了 Agent 中的所有软件接口,包括 Domain_ Interface(用于启动常规信息的接口,包括用户输入的各种信息接口)和 Communication Interface(用于与其他 Agents 之间协作的全部通信接口);Activity 列举了 Agent 中的所有活动,包括 Public_Activity(描述一组由该 Agent 调度和监控的、基本或复合的计算活动)以及 Exception Activity(描述该 Agent 执行计算活动过程中可能产生的异常处理行为);Info_ Blackboard 记载该 Agent 接收的外部信息和自身产生的中间或结果信息;Other_agent 描述能通过协商参与协同工作的其它 Agent。

定义 2(UIS 多主体协作) UIS 多主体协作可抽象成一个 4 元组:

```
Agents_negotiation = ( ID;
                        Type;
                        Negotiation_Activity;
                        Agent_Member;
                        )
```

其中, ID 表示此次协作任务的标识名称; Type 表示此次协作任务的类型; Negotiation_Activity 表示需要通过多个 Agent 协作才能完成的复合任务的描述; Agent_Member 表示为完成此次协作任务需调度的 Agent 的集合。

在 UIS 模型中,我们定义了 5 种 Agent 来搭建整个模型的框架结构(如图 1 所示),包括信息集成主体(Information Integration Agent, IIA)、接收更新主体(Receiving Update Agent, RUA)、执行更新主体(Executing Update Agent, EUA)、信息服务主体(Information Service Agent, ISA)和 URI 访问主体(URI Access Agent, UAA)。同时,我们还引入了 2 种多主体协作方式:元数据更新协作(Update Metadata Cooperation, UMC)和信息检索协作(Information Retrieval Cooperation, IRC)。

本文将在后续小节结合模型结构详细介绍各主体的具体描述。

2.2 UIS 工作流程

UIS 主要工作流程分为三步:

(1)资源提供者在接受联盟章程的基础上注册加入联盟,成为信息加盟方。然后共享其提供的物理资源, IIA 主体会按照统一的模式和规范对数据信息进行加工,生成标准元数据,差异备份产生更新文件并上传至联盟共享平台。其中,资源提供者可以提供任意多个同时分布于多个数据库或文件服务器之中的物理资源。

(2)联盟共享模块中, RUA 主体和 EUA 主体共同搭建了 UMC 协作。其中, RUA 主体不断接收来自多个 IIA 主体提交的元数据更新文件,在数据合法性检测之后,向更新任务队列里写入更新任务。而 EUA 主体则在时钟的驱动下定期地被激活,读取更新任务队列中的全部任务,完成对中央数据集的逻辑数据进行更新。

(3)联盟服务模块为最终用户的信息检索与交互提供一个统一的信息服务平台。用户可以通过电脑、手机、邮件等多种终端,提交信息检索的服务请求,由 ISA 主体负责处理和返回所需信息资源。如果需要对信息的远程物理资源进行访问,则通过 IRC 主体协作(ISA 主体与 UAA 主体共同组建),由 ISA 主体与 UAA 主体建立远程通信,以 URI(统一资源标识符)为物理资源标识,实现物理资源访问远程。

3 UIS 中的信息加盟和 IIA 主体

信息加盟是 UIS 模型的基本单元。它的一个重要特点是:在逻辑上将各个信息资源的异构数据连接在一起,在物理上则对加盟数据按照元数据规范标准进行整合并备份上传至联盟。UIS 信息加盟模块具有良好的通用性和可移植性,可以处理不同类型数据库(例如:MS SQLServer, Oracle, Sybase, MySQL 等)和不同类型的文件资源(多媒体、文档、网页等),并支持跨平台操作。

3.1 IIA 主体描述

UIS 的核心功能之一——异构数据库在逻辑上的标准化处理——是通过信息集成主体 IIA 来实现的。IIA 主体的内容如下:

```
Agent IIA : Inherit Agent
{
    ID = IIA001;
    Type = IIA;
    Interface:
        Domain Interface:
            DomainInformation = GetDomainFromUser();
        Communication Interface:
            CommunicationInformation = GetCommFromAgent(RUA);
    Activity:
        Public Activity:
            MetaData = ExecuteIntegration ( DomainInformation, Exception);
            UpdateFile = ExecuteBackup ( MetaData, Exception);
            ExecuteUpload ( CommunicationInformation, Exception);
        Exception Activity:
            SolveException(Exception);
    Info_Blackboard = DomainInformation;
    Other_agent = RUA;
}
```

在 IIA 主体的 Interface 中, GetDomainFromUser() 获取用户的共享配置信息; GetCommFromAgent(RUA) 从指定的 RUA 主体中获取通信信息。在 Activity 中, ExecuteIntegration(DomainInformation, Exception) 根据用户的配置信息 DomainInformation 执行异构数据库的标准化处理并产生元数据, ExecuteBackup(MetaData, Exception) 对元数据 MetaData 进行差异备份并生成更新文件,而 ExecuteUpload(CommunicationInformation, Exception) 则是根据主体间的通信信息 CommunicationInformation 上传更新文件至联盟中央; SolveException(Exception) 则是对上述行为可能产生的各种异常 Exception 进行相关的处理。

在 IIA 的 Activity 中,异构数据库的标准化和异常处理是最为关键的核心部分。在 3.2 节和 3.3 节中,我们将详细论述。

3.2 异构数据库的标准化

由于信息加盟数据库的存储结构和数据模式各不相同,为了与中央元数据标准规范相一致,必须对异构数据库进行标准化处理,实现加盟数据库自定义数据模式到中央标准数据模式的映射和转换。IIA 主体提供了字段映射机制,使信息加盟方能自由定义加盟数据库的数据模式到标准元数据模式的映射。该映射机制属于“多对多”映射,即允许多个自定义字段对应于单个元数据字段,同时也允许一个自定义字段对应于多个元数据字段。

定义 3(元数据) 元数据(MetaData)在 UIS 模型中抽象为一个 4 元组:

```
MDT = ( Center Code;
        URI;
```

MDL {MDI1, ..., MDIn};
N;
)

其中,Center Code 是信息加盟方的中央授权码;URI 是数据的唯一标识编码;MDL 为元数据列表的集合,MDI 为元数据项;N 是元数据列表的长度。

定义 4(元数据项) 元数据项(MetaData Item)可以抽象为一个 3 元组:

MDI = (Name;
Type;
Explanation;
)

其中,Name 是元数据项的属性名称;Type 是元数据项的属性类别及取值约束;Explanation 是元数据项的属性的内涵解释;

字段映射的机制如图 2 所示。加盟数据库的数据映射分成两部分:标准元数据和非标准元数据。其中,标准元数据是组成中央数据集的逻辑数据实体,用于联盟服务模块的信息检索;而非标准元数据作为标准元数据更加详细的补充,以物理资源的形式存储于加盟数据服务器之上,通过 URI 远程访问并加载。

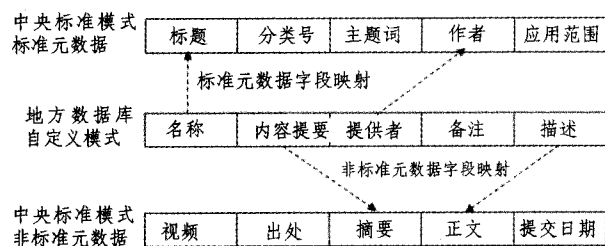


图 2 地方数据库数据到中央元数据的格式映射与转换

考虑到加盟数据库结构的显著差异以及可能存在的结构不合理等因素,我们对数据模式的映射机制做了进一步补充。例如,加盟数据库很可能因缺乏某些必要字段(如图 2 中缺“主题词”字段)而无法建立映射,这里通过设置缺省值、置空或者适当修改加盟数据库结构等方式加以解决。

3.3 异常处理

IIA 主体中考虑了诸多异常问题的处理,如字段映射匹配错误异常、元数据差异备份中断异常、数据更新文件传输失败异常等。这里以“元数据备份中断异常”为例介绍异常处理的基本思路。

元数据备份中断异常是指在 IIA 主体对加盟数据库进行元数据备份过程中,因服务器断电、网络连接中断等突发原因所造成的操作异常终止。IIA 主体中通过构造一种数据事务管理机制来处理这类异常中断问题。

元数据备份过程可分为完全备份和差异备份两个部分。信息加盟的 IIA 主体已经保存了上一次数据备份的元数据集 A(首次备份时集合 A 为空集)。因此集合 A 也就是上一次元数据的完全备份集合。而当前的备份过程则首先需要将当前用户共享数据库的最新数据按照字段映射规则进行完全备份,产生集合 B。因此集合 B 就是当前的元数据完全备份集合。然后将集合 A 和集合 B 进行对比,将两次备份的差异保存为集合 C。因此集合 C 就是用来上传至中央服务器的元数据集。最后将集合 C 的数据信息写入文件,上传至中央

服务器。信息加盟模块整个备份过程中,增加了事务管理机制,以表为单位进行操作。当一个表的操作成功结束时,该表被标识为“成功”,并对下一个表进行操作。反之,当遇到异常中断时,则该表会在系统重启检测时标识为“异常”,并撤销该表之前的全部操作,重新执行,从而保证数据操作的一致性和有效性。这个过程的执行和具体实现并不完全依赖于数据库本身提供的事务管理功能。

4 UIS 中的联盟共享、RUA 主体和 EUA 主体

在整个联盟体系中,联盟共享是 UIS 模型的核心部分,负责接收信息加盟模块提供的标准元数据,并在规定时间内单位内更新中央数据集,以供联盟服务模块进行逻辑信息检索。联盟共享模块主要通过 RUA 主体和 EUA 主体的协作来实现(如图 1 所示)。显然,更新任务队列成为 RUA 主体与 EUA 主体两个 Agent 之间的信息交流区域。

RUA 主体是连接信息加盟模块和联盟共享模块之间的接口。信息加盟模块通过 IIA 主体将差异备份的更新文件提交中央服务器,由 RUA 主体接收并恢复成临时数据库,待数据合法性检测通过之后向更新任务队列中添加任务。

EUA 主体是连接联盟共享模块和联盟服务模块之间的接口。在时钟的驱动下 EUA 主体读取更新任务队列,执行元数据和关键词索引的更新任务,更新中央数据集,从而保证联盟服务模块总能在最新的中央数据集下进行有效的信息检索。

RUA 主体与 EUA 主体不是以“一对一”的关系存在,而是以“多对一”的关系协作完成任务,如图 3 所示。其中,RUA 主体是伴随着更新文件到达而触发,而 EUA 主体的行为触发是通过时钟驱动,与 RUA 无关;更新任务队列则成为二者的交流信息的接口,将多个 RUA 与单个 EUA 联系起来。

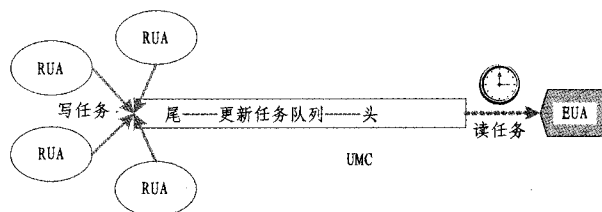


图 3 RUA 与 EUA 的“多对一”关系示意图

RUA 主体与 EUA 主体的协作 UMC 定义如下:

```
Agents_negotiation UMC
{
    ID = UMC001;
    Type = UMC;
    Negotiation_Activity;
    Activity UpdateMetaData(RUA, EUA)
    {
        // RUA 主体接收更新文件,创建更新新任务
        (1) UpdateTask = CreateTask(RUA.ReceiveFile);
        // 时钟驱动 RUA 读取更新任务执行元数据更新
        (2) ExecuteUpdate(EUA.ReadTask, Time, Exception);
    }
    Agent_Member = RUA, EUA;
}
```

UMC 主体协作需要保证中央数据集元数据更新的同步性和一致性。

中央数据集是整个联盟体系三个模块的联系纽带,其元数据更新能否保证同步和一致是整个 UIS 模型能否成功的关键所在。所谓同步更新是指,中央数据集数据能够在最短时间内得到最及时的更新。所谓一致更新是指,中央数据集

及各个副本更新的元数据要保证一致、相同。

RUA 主体和 EUA 主体是相互独立的,二者不存在任何时间先后和逻辑主次的关系。RUA 主体是实时触发,而 EUA 主体则是延时调用。UMC 主体协作通过将 RUA 主体的实时性和 EUA 主体的延时性结合在一起,来实现面向用户的更新同步性。对于中央数据集和各个副本,分别定义各自的更新任务队列来存储各个更新任务,可以有效地保证数据更新的时间有序性和更新一致性。

5 UIS 中的联盟查询服务

联盟服务模块是 UIS 模型的最终体现,也是中央数据集的信息检索接口。联盟服务模块提供的信息检索为用户提供了基于关键词的模糊查询、基于信息类别的目录查询和自然语言查询三种信息检索功能,可以帮助用户快捷而准确地检索到所需信息。

ISA 和 UAA 主体的描述以及 IRC 的定义方式均类似于 IIA 和 UMC。由于 ISA 主体提供了多项服务和功能,在此仅列举 ISA 主体的主要接口和行为。

Agent ISA :Inherit Agent

```
{
    .....
    QCI {Keywords,Category,PageSize,PageID,.....}
    Interface:
        Domain Interface:
            (1) QCInf = GetQCInf();
            (2) NLQInf = GetNLQInf();
            (3) MMRInf = GetMMRInf();
            .....
    Activity:
        Public Activity:
            QueryResult = QueryInf(QCInf,Exception);
            QueryResultNLQ = QueryInfNLQ (NLQInf,
            Exception);
            .....
        Exception Activity:
            SolveException(Exception);
            .....
}
```

在 ISA 主体的 Domain Interface 中,GetQCInf() 获取用户查询条件信息,其中查询条件结构 QCI 包括 Keywords(关键词信息)、Category(信息类别范围)、PageSize(页面列表长度)、PageID(当前显示的页面号)等相关信息;GetNLQInf() 获取用户提交的自然语言问句;GetMMRInf() 获取多媒体资源信息。在 Activity 中,QueryInf(QCInf, Exception) 执行关键词查询和目录查询行为,而 QueryInfNLQ(NLQInf, Exception) 则是执行自然语言查询的行为。

以下就 ISA 与 UAA 主体中的几个重要算法和技术进行讨论。

5.1 模糊查询的算法设计

ISA 主体提供了基于关键词的模糊查询和目录查询。这两项查询是以中央数据集的信息元数据为基础的。随着共享信息的增多,中央数据集也会随之膨胀。为了避免查询效率的潜在降低,我们提出了一种基于“关键词在指定目录范围内的精确匹配”为思想的模糊查询算法。

算法 1 共享数据的模糊查询算法

Step1. 初始化读取中央数据集的全体元数据索引信息,构造生成“双十字有序链表”的逻辑存储结构,如图 4 所示。

Step2. 按照规范的格式读取用户提交的查询任务,准确解析查询任务的各个参数,包括查询的目录、关键词列表等。

Step3. 对关键词列表进行分词处理,计算各个关键词的哈希值。

Step4. 通过“二分查找”在哈希表内查找各个关键词,横向遍历所对应的记录,按照关键词匹配系数取值,在指定的目录范围内获取记录交集。

Step5. 如果交集为空,转 Step6. 否则,转 Step7.

Step6. 如果关键词匹配为全匹配,则弱化条件,转 Step4. 否则返回空。

Step7. 将符合用户要求的单位个数的元数据方案索引信息拼成完整地 SQL 查询语句,回访数据库获取完整的元数据信息。

Step8. 按照 HTML 语法要求生成动态 Web 页面,返回给用户从而完成一次查询任务。

“双十字有序链表”的存储结构有利于信息检索,同时降低了内存的开销。如图 4 所示,链表 A 的数字是关键词对应的哈希值,而链表 B 的各个数字表示记录号为 ID 的记录所包含的各个关键词的哈希值。算法一方面可以从链表 A 起始查找包含某个关键词的全部元数据记录,另一方面也可以从链表 B 着手查找某条元数据所包含的全部关键词信息。同时,内存存储的都是元数据的索引信息,多为数字编码,这极大降低了空间开销。

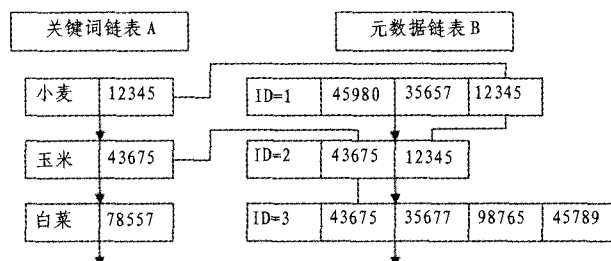


图 4 元数据查询的内存存储结构

查询算法的执行效率较高。设关键词链表 A 长度为 n , 用户提交进行查询的关键词个数是 k , 平均每个关键词横向链表长度约记为 m , 则查找的时间复杂度是 $O(mk \log n)$, 而与元数据链表 B 长度(即中央数据集记录总数)无关。

算法可以自动弱化查询条件,以求得近似的查询结果。例如,当用户提交多个关键词而记录交集为空时,可以将“关键词匹配系数”降低。例如假设弱化系数修改为 0.75, 则 75% 的关键词匹配成功的记录将反馈给用户。该系数缺省值为 1, 即要求全部匹配。

5.2 自然语言查询

ISA 主体同时提供了自然语言查询的接口,定义了一种基于本体的查询语言,允许用户自由地进行知识查询。自然语言查询更加适合于那些普通的基层用户使用,具有直观、高效等优点。UIS 联盟服务模块增加自然语言查询,极大地拓宽了服务的适用群体,同时使得联盟共享的数据资源得到更好的利用。

用户可以通过自然语言查询,输入自然语言问句获取所需信息。如图 5 所示,ISA 主体首先对自然语言问句进行智

能分词,然后利用查询文法进行模糊匹配,找到对应句型,最后通过知识库 API 函数(KAPI),从领域知识库中找到符合用户需求的信息,生成完整的信息文本,反馈给用户,完成一次自然语言查询过程^[12]。

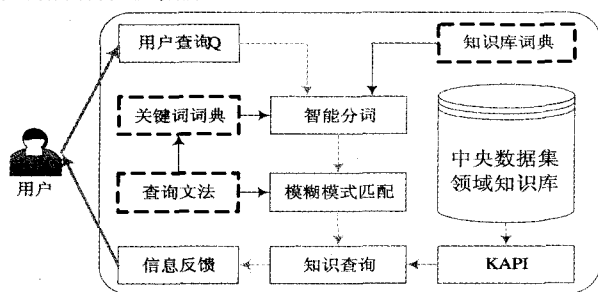


图5 自然语言查询逻辑结构图

以前的研究表明,对任何一个用户的自然语言查询,我们总可以将其拆解成对某种或某些槽的查询。因此,我们设计了一种基于概念空间和属性空间的知识查询语言,建立每个槽的各种查询形式^[10~12]。

首先,我们对概念空间中所有的槽根据其槽类进行聚类,将查询方式相似的槽聚在一起,抽出共同的查询模式,形成具有继承关系的属性树。属性树的非叶子结点表示一个提问主题,而叶子节点是概念空间中实际存在的槽,我们称之为具体槽。

常见的提问主题包括数量、时间、地点、原因、方法、来源等。一个大的提问主题又可分为几个小的提问主题,例如数量可以分为长度、面积、体积、密度、高度、宽度、气温等。

其次,我们按照自顶向下的方式对各槽进行知识查询语言设计。属性树满足知识查询语言的继承关系,上位槽的知识查询语言可以被下位槽参数继承。

5.3 盟内 URI 的远程访问

UIS 模型提供了基于元数据的信息查询功能,而更详细的资源仍保存在信息加盟方的数据库之中。ISA 主体最初提供给用户的是一种基于元数据的逻辑查询结果。如果用户希望获取更多、更详细的资源(如文档全文)时,ISA 主体可以远程通信 UAA 主体,以 URI 编码为标识访问物理资源。

URI(Unique Resource Identity,即统一资源标识符)是一组规范性编码。在整个 UIS 模型中,每一条元数据都有一个唯一的身份标识,即 URI 编码。URI 编码分段存储了元数据来源的全部标识信息,包括中央授权码,物理数据所指向的数据服务器、数据库、数据表以及对应的数据记录号信息。

URI 的编码规范如表 1 所示。

表 1 URI 编码解析

段命名	段解释	段位置	示例
CenterCode	中央授权码	1-8 位	BJICTCAS
ServerID	服务器 ID	9-11 位	001
DatabaseID	数据库 ID	12-14 位	002
TableID	共享表 ID	15-17 位	003
RecordID	记录号	18-26 位	000000010

由此规范得到的示例 URI="BJICTCAS001002003000000010"。

这里所提到的服务器 ID、数据库 ID 等都是信息加盟方共享数据时必须提供的基本信息。当用户需要获取详细资源时,ISA 主体将根据信息加盟方注册加盟时提供的 Web 服务器 IP 地址,重定位到指定的信息加盟方服务器,与 UAA 主体建立通信,根据 URI 编码进行解码,访问数据库获取指定资源。

URI 远程访问实现了逻辑信息到物理信息的转换,在保证查询结果的同时,有效地支撑了联盟的自由、自主共享机制,也最大限度地降低了中央数据集所需要的空间开销。

6 UIS 中的多媒体智能教学服务

由于共享信息中存在着大量的多媒体资源,因此有效地利用这类资源,进一步提高用户对知识的认知程度,则是 UIS 模型提供的另一个重要服务。多媒体智能教学主要是通过领域专家的命题和学习用户的答题来实现教与学之间的联系,其逻辑结构如图 6 所示。它主要由逻辑相对独立的三个层次组成:交互层、教学层和领域资源层。同时,为了更好地满足用户获取学习知识的需求,我们分别建立了知识模型、教学模型和用户模型来描述知识、教学和用户学习的状态和环境。

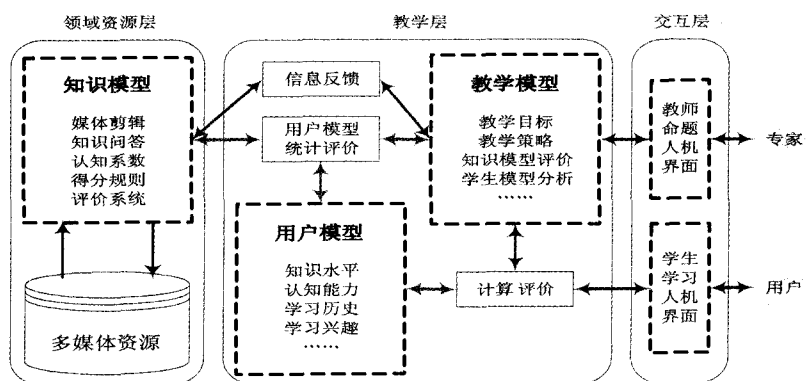


图6 智能教学的逻辑结构图

6.1 领域专家的命题设计

智能教学的智能性主要体现在教与学之间的互动上,领域专家通过增加知识命题来考核用户的学习效果,从而可以有效引导用户未来的学习。

具体来说,专家是在原有多媒体资源的基础上,增加时间

断点并根据多媒体资源片断所涉及的知识来制定知识题目,最后根据相关的算法自动地抽取适量且难度适合于用户学习的题目集合来考核用户的学习质量。在此,我们给出专家命题数据模式的形式化描述。

定义 5(命题数据模式) 命题数据模式用于记录和描述

专家的命题信息,可抽象成一个 10 元组:

```
Question(EXPERTISE)={ QuestionID;
                        TimePointID;
                        Content;
                        Difficulty;
                        OptionNumber;
                        OptionList;
                        CorrectOption;
                        Point;
                        IsMultiple;
                        SubmitDate;
                        }
```

其中,QuestionID 表示问题 ID;TimePointID 表示命题所在的时间断点 ID;Content 表示知识问答的问题;Difficulty 表示问题的难度系数;OptionNumber 表示选项个数;OptionList 表示选项列表信息,其格式为:〈AnswerList〉:=〈选择〉{~〈选择〉};CorrectOption 表示正确选项;Point 表示命题的初始分值;IsMultiple 表示是否是多项选择;SubmitDate 表示专家命题的日期、时间。

6.2 学习用户的答题设计

在智能教学过程中,用户作为学习的主体,通过多媒体资源及相关的知识命题,来获取并检验和巩固自己期待获取的知识。用户对知识认知程度的评价主要是通过回答专家的命题,从而进一步统计分析得出。在此,我们给出用户答题数据模式的形式化描述。

定义 6(答题数据模式) 答题数据模式用于记录和描述用户的答题状况,可抽象成一个 7 元组:

```
Answer(USER)={ UserID;
                 TimePointID;
                 QuestionNumber;
                 AnswerList;
                 Grade;
                 ToleranceDifficulty;
                 SubmitDate;
                 }
```

其中,UserID 表示用户 ID;TimePointID 表示回答问题所在的时间断点 ID;QuestionNumber 表示在该时间断点下的问题编号;AnswerList 表示用户的答案评价信息,其格式为:〈answer〉:=〈题目 ID〉:〈用户的选择〉:〈题目难度〉:〈是否正确〉,〈AnswerList〉:=〈answer〉{|〈answer〉};Grade 表示用户的当前答题得分;ToleranceDifficulty 表示用户的当前答题的认知能力;SubmitDate 表示用户答题的日期、时间。AnswerList 的举例如下:1:1,2:5:c|4:3;6:w|5:0;5:p|7:0;5:p,其中:如果是多项选择,则用,分隔开选择;c 表示正确,w 表示错误,s 表示部分正确(多选题);用户的选择如果为 0,表示用户没有选择答案,相应的字母为 p。

对于整个智能教学,最核心的部分则是如何合理地评价用户认知能力,即学习效果和质量的衡量。因此,我们需要制定一份理想的“考卷”和一个理性的“计算方案”。由此,我们引入一个用户认知能力系数 λ ,通过不断计算并更新 λ 来时评价用户学习效果,并据此给出一种简单实用的抽题方案。

算法 2 基于用户认知能力的抽题算法

Step1 用户进入多媒体交互学习,读取其用户模型。

Step2 如果系统没有为其建立用户模型,表示该用户是

第一次学习,建立并初始化用户模型,默认用户的认知能力 $\lambda = 5$,即中等程度,其中 $\lambda \in [1, 9]$ 。跳转至 Step4。

Step3 如果存在用户模型,读取其认知能力参数 λ 。

Step4 从知识问答信息库中,随机抽取单位数量的题目,同时保证平均难度系数约等于用户认知能力参数。

Step5 根据用户的答题结果,依据算法 3 重新计算用户的认知能力参数 λ ,并相应修改用户模型,将本次学习过程保存至数据库中以备日后统计分析。

Step6 算法结束。

知识问答信息库是由专家不断扩充和完善的,用于组成问卷的知识问答题目则是根据用户认知能力参数 λ 取值而随机抽取的,因此用户的每一次学习都将遇到不同的问卷,也都将有不同的收获。

其次,我们来定义一下用户认知能力的计算方案。

算法 3 测算用户认知系数的算法

Step1 用户 UID 开始观看视频 M。

Step2 系统依据算法 1 抽题组成问卷提交给用户答题。第 i 个时间断点答对的题目的分值之和为 S_i ,答错的题目的分值之和 $S'i$ 。

Step3 在用户 UID 观看完视频 M 后,经过判断答案,设该用户学习视频 M 的“本次得分”= $S_1 + \dots + S_k$,”本次失分”= $S'1 + \dots + S'k$,那么系统本次呈现给用户的所有选择题的“总分值”=“本次得分”+“本次失分”。

Step4 “本次成绩”可以经过以下百分制的折算,划归为百分制的形式:

$$\begin{aligned} \text{“本次成绩”} &= 100 \times \text{“本次得分”} / \text{“总分值”} \\ &= 100 \times (S_1 + \dots + S_k) / (S_1 + \dots + S_k + S'1 + \dots + S'k) \end{aligned}$$

Step5 采用以下启发式方法,重新计算该用户的认知系数:

如果 $(S_1 + \dots + S_k) \leq \text{“总分值”} / 4$,则

认知系数 = 2

否则,如果“本次成绩” ≥ 80 分,则

认知系数 = $\min(8, \text{认知系数} + 2)$

否则,如果“本次成绩” < 80 分,则

认知系数 = $\min(9, 2 + \lfloor (d_i + d_j) / 2 \rfloor)$,其中 d_i 和 d_j 表示本次问答中最难的两道题的难度系数。

Step6 算法结束。

其中,本次成绩是按照用户答对题目分值的百分比折合成分百分制模式计算产生,而认知系数则是根据本次成绩的取值进一步换算产生。例如,当用户答对的分值还不及四分之一时,则我们认为用户的认知能力较低,认知系数取值为 2。

7 UIS 的实现和应用

AUIS 的实现主要结合了 Web 技术、信息处理和数据库集成等多方面技术,基于 Java、C、PHP 等多种开发语言实现,支持 MS SQL Server、Oracle、Sybase、MySQL 数据库接口,同时能够在 Windows 系列和 Linux 操作系统下移植运行,具有较好的通用性和可移植性。

目前,AUIS 已经成功地在北京市、湖北省和浙江省试运行,取得良好的资源共享效果。AUIS 目前已经共享了科技部农村中心负责主持的“农业科技信息网”,成功地加盟了“中国现代农业科技信息共享示范网”、“九亿网”、“国家农业科技

园区网”、“中国农村技术开发中心网”等多家网站数据库的信息,制定了《中国农业科技信息共享联盟章程》等诸多规范和章程,初步建立了中国农业信息共享联盟体系(简称 AUIS)。

在此,我们仅就 AUIS 系统中的三个重要部分进行介绍。

(1) 异构数据库的标准化处理

图 7 显示了 AUIS 系统的信息加盟模块中,信息加盟方在进行元数据字段映射的过程。

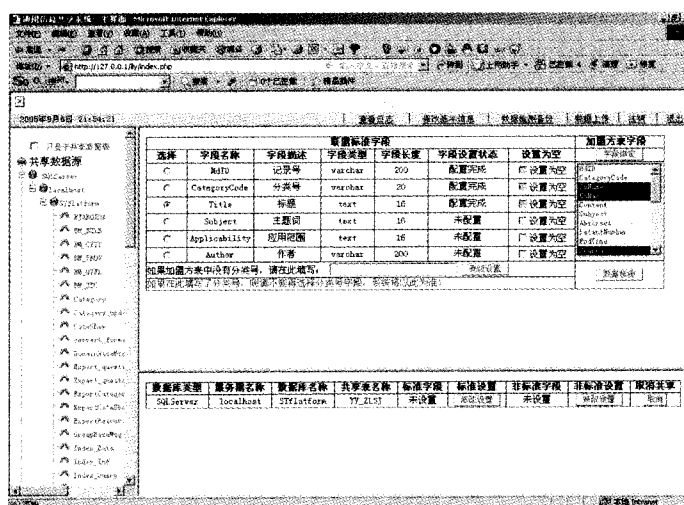


图 7 AUIS 系统信息加盟方信息加盟模块元数据字段映射界面

左侧树形列表列举了资源提供者的物理资源信息,右侧的共享列表列举了用户已经共享的资源信息。在主操作区里正在进行的的就是异构数据库标准化处理的配置过程。其中,表格列举了 AUIS 的元数据列表规范,而右侧列表框中则是用户共享表的字段列表。用户可以在此进行字段的映射配置,之后信息共享主体将会自动完成数据的标准化处理,差异备份后上传至联盟。

(2) 联盟服务中的信息检索

AUIS 系统为用户提供了基于标题、主题词等关键词检索和基于专业资源分类的目录检索服务。图 8 为用户查询资源类别为“农业科技成果”的目录检索界面。用户可以在左侧的目录树上直接按照资源类别进行查询,也可以在左上方的“搜索”栏中输入关键词进行查询。右侧为检索到的资源列表。

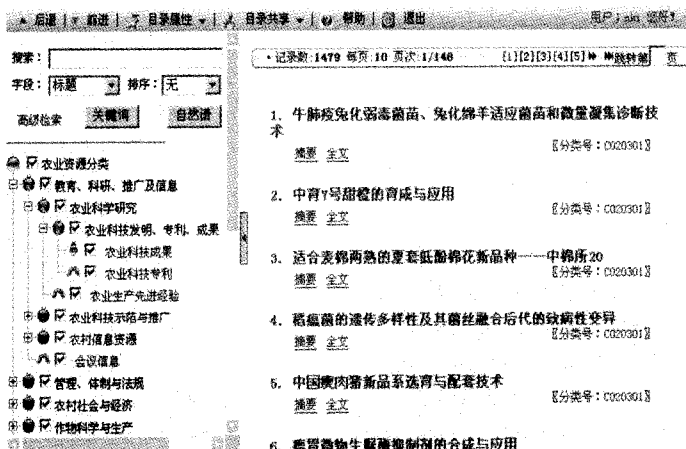


图 8 信息共享联盟查询系统在农业科技网中的应用

(3) 自然语言查询

AUIS 系统在联盟模块中提供了自然语言查询功能,如图 9 所示,用户可以提交任何与农业信息相关的自然语言问句进行信息查询。同时考虑到用户输入的多样性,如不同的语言习惯、输入不够准确或完整等,自然语言查询系统提供了支持模糊性理解和自动纠错的功能,具有较好的容模糊性和容错性。例如,用户输入“穿山甲幽神么用”,系统能够正确识别为“穿山甲有什么用”,并在中央数据集领域知识库中获取对应的答案。

通过对 1 万个问题的初步测试,自然语言查询系统取得了比较满意的结果^[12],如表 2 所示。

(下转第 23 页)

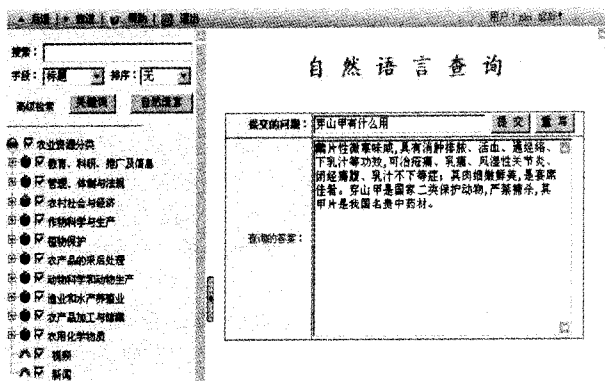


图 9 AUIS 系统中的自然语言查询过程

- ceedings of the 2006 Symposium on Applications and the Internet (SAINT'06). Phoenix, Arizona, USA: IEEE Computer Society, 2006
- 15 Katabi D, Handley M, Rohrs C. Congestion control for high bandwidth-delay product networks. ACM SIGCOMM Computer Communication Review, 2002, 32 (4): 89~102
 - 16 Zhang Y, Henderson T R. An implementation and experimental study of the explicit control protocol (XCP). In: Proceedings IEEE INFOCOM 2005. Miami, CA, USA: IEEE Society, 2005. vol 2: 1037~1048
 - 17 徐建, 李善平. 用户公平的活动队列管理. 电子学报, 2004, 32 (3): 435~440
 - 18 章森, 吴建平, 徐格. 一种改善 TCP 公平性的算法. 清华大学学报(自然科学版), 2003, 43(9): 1280~1283
 - 19 Xia Y, Subramanian L, et al. One more bit is enough. ACM SIGCOMM Computer Communication Review, 2005, 35 (4): 37~48
 - 20 Lin D, R Morris. Dynamics of random early detection. In: Proceedings of ACM SIGCOMM 1997. Cannes, France: ACM Press, 1997. 127~137
 - 21 Mahajan R, Floyd S, Wetherall D. Controlling High-Bandwidth Flows at the Congested Router: [Technical Report]. TR-01-001. AT&T Center for Internet Research at ICSI (ACIRI), April 2001
 - 22 Feng W, Kandlur D, Saha D, et al. Stochastic fair Blue: A queue management algorithm for enforcing fairness. In: Proc. IEEE INFOCOM 2001. Anchorage, Alaska: IEEE Communications Society, 2001. 1520~1529
 - 23 Pan R, Prabhakar B, Psounis K. CHOKe: A stateless active queue management scheme for approximating fair bandwidth allocation. In: Proc. of IEEE INFOCOM'2000. Tel Aviv, Israel: IEEE Communications Society, 2000, vol 2. 942~951
 - 24 Jiang Y, Hamdi M, Liu J. Self adjustable CHOKe: an active queue management algorithm for congestion control and fair bandwidth allocation. In: Proc. ISCC'03. IEEE, 2003; 1018~1025
 - 25 Chatranon G, Labrador M A, Banerjee S. BLACK: detection and preferential dropping of high bandwidth unresponsive flows. In: Proc. of ICC '03. IEEE, 2003, vol 1: 664~668
 - 26 汤德佑, 骆嘉伟, 张大方, 等. 一种提高稳定性和公平性的主动队列管理机制. 计算机研究与发展, 2005, 42(7): 1136~1142
 - 27 Ho C Y, Chan Y C, Chen Y C. A TCP-friendly stateless AQM scheme for fair bandwidth allocation. In: Proceedings of the Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services (ICAS/ ICNS 2005). IEEE Computer Society, 2005. 14~19
 - 28 Le L, Aikat J, Jeffay K, et al. Differential congestion notification: Taming the elephants. In: Proceedings of the 12th IEEE International Conference on Network Protocols, ICNP 2004. Berlin, Germany, October, 2004. 118~128
 - 29 Zhao Z, Darbha S, Reddy A L N. A Method for estimating the proportion of non-responsive traffic at a router. IEEE/ACM Transactions on Networking, 2004, 12 (4): 708~718
 - 30 Stoica I, Shenker S, Zhang H. Core-stateless fair queuing: achieving approximately fair bandwidth allocations in high speed networks. ACM SIGCOMM Computer Communication Review, 1998, 28 (4): 118~130
 - 31 刘刚, 张德运, 刘晶. 基于用户公平的随机早期检测算法改进. 微电子学与计算机, 2004, 21(6): 119~122
 - 32 邓晓衡, 陈志刚, 王建新, 等. 基于测量的 TCP 拥塞控制的公平性研究. 小型微型计算机系统, 2004, 25(11): 1894~1898
 - 33 Tam A S, Chiu D M, Lui J C S, et al. A case for TCP-friendly admission control. In: Proc. of 14th IEEE International Workshop on Quality of Service, 2006 (IWQoS'06). New Haven, CT, USA: IEEE Society, 2006. 229~238
 - 34 Chandrayana K, Kalyanaraman S. Uncooperative congestion control. ACM SIGMETRICS Performance Evaluation Review, 2004, 32 (1): 258~269
 - 35 Fan X, Chandrayana K, Arcak M, et al. A two-time-scale design for edge-based detection and rectification of uncooperative flows. IEEE/ACM Trans Networking, 2006, 14 (6): 1313~1322
 - 36 Padhye J, Firoiu V, Towsley D, et al. Modeling TCP throughput: A simple model and its empirical validation. ACM SIGCOMM Computer Communication Review, 1998, 28 (4): 303~314
 - 37 Chatranon G, Labrador M A, Banerjee S. Fairness of AQM schemes for TCP-friendly traffic. In: Proceedings of IEEE Globecom 2004. Dallas, Texas: IEEE Communications Society, 2004, vol 2. 725~731

(上接第 8 页)

表 2 自然语言查询系统测试统计结果

评价参数	测试结果
准确率	95.6%
误识率	0.28%
未识率	4.12%
处理时间	0.13 秒/问题
自由度	98.1%

(4) 多媒体智能教学

AUIS 系统提供了共享多媒体资源的智能教学服务平台。图 10 所示为领域专家为某一多媒体视频资源提供知识命题的过程。在该系统中, 专家可以任意地通过人机界面对所选多媒体资源增减时间断点, 添删任何难度及分值的问题。用户则可以在该系统下观看并回答多媒体资源中的各类问题, 从而检测自身对于知识的掌握程度。

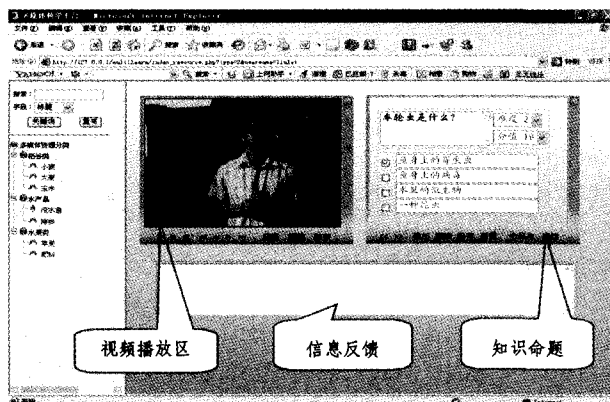


图 10 AUIS 系统中领域专家命题过程

结论和未来工作 本文介绍了一种信息共享联盟模

型——UIS, 有效地实现了分布异构数据的标准化和信息共享, 同时也实现了共享资源的多模式信息检索, 提高了网络资源的有效利用率。该模型具有诸多优点, 包括信息共享的协调和自主, 信息检索的高效、准确和多样, 模型所具有的可扩展性和可移植性等等。当然, 该模型仍然存在一些不足之处, 有待于进一步完善。例如, 模型是建立在复杂的网络环境之中, 因此数据通信的机制、协议以及系统的异常控制都决定了整个模型能否稳定执行。这些机制和协议还有待于在实践中进一步完善和规范。

由于网络环境的复杂性和资源的多样性, UIS 模型实现的需求也将随之不同。因此, 应该结合实际的调研结果, 使 UIS 模型不断完善, 适应更加复杂的需求, 具有更好的鲁棒性和健壮性。

参 考 文 献

- 1 郑宇飞, 刘磊, 丰强泽, 曹存根. 农村科技信息共享服务系统的设计和实现. 计算机科学, 2005, 32(8): 1~6
- 2 陈良玉, 等. 农村科技信息共享技术与实践. 北京: 中国农业科学技术出版社, 2004
- 3 杜青林. 中国农村经济结构战略性调整. 北京: 中国农业出版社, 2003
- 4 彭常. 关于信息传输“最后一公里”的研究. 见: 中国科协首届学术年会论文集. 中国科学技术出版社, 1999
- 5 温国泉. 谈谈我国现阶段农业信息化服务的发展模式. 科技情报开发与经济, 2002, 12(5): 33~34, 36
- 6 谢能付. 基于语义网的知识融合和同步研究: [博士学位论文]. 北京: 中国科学院计算技术研究所, 2006
- 7 张维明, 等. 语义信息模型及应用. 北京: 电子工业出版社, 2002
- 8 史忠植. 高级人工智能. 北京: 科学出版社, 1998
- 9 史忠植. 智能主体及其应用. 北京: 科学出版社, 2001
- 10 冯东辉, 曹存根. 知识界面在 NKI 中的应用. 计算机工程与应用, 2001
- 11 丰强泽. 人机交互中用户隐式知识需求分析研究: [博士学位论文]. 北京: 中国科学院计算技术研究所, 2006
- 12 Feng Q, Cao C, Si J, et al. A Uniform Human Knowledge Interface to the Multi-Domain Knowledge Bases in the National Knowledge Infrastructure. ES2002, 2002. 163~176