

文章编号: 1673- 0062(2010) 01- 0048- 04

基于领域本体的答疑网格系统

黄玲玲, 阳小华, 杨 凯

(南华大学 计算机科学与技术学院, 湖南 衡阳 421001)

摘 要:目前我国答疑系统隶属于某个教育平台各自开发, 导致资源重复建设; 各个系统之间也不能有效互操作导致系统间资源无法共享和统一协调管理. 本文从资源共享和易于建设的角度出发, 提出基于领域本体的答疑网格系统, 文中给出系统的体系结构, 详细阐述答疑流程, 探讨了关键技术及算法, 并分析系统特点及优势.

关键词:网格; 领域本体; 答疑网格系统; 资源请求本体; 资源节点匹配

中图分类号: TP18 TP393 **文献标识码:** A

Answering Grid System Based on Domain Ontology

HUANG Ling-ling YANG Xiao-hua YANG Kai

(School of Computer and Technology University of South China, Hengyang Hunan 421001, China)

Abstract Focusing on the current domestic situations that the exclusive development of each answering system leads to the duplicate building of resource, and that the lowly effective mutual operation between heterogeneous systems brings on poorly resources sharing and disunified management, this paper, from the starting point of resource sharing and being easy to construct, puts forward an answering grid system based on domain ontology. It advances the system architecture, elaborates the question answer flow, probes into the key implement technology and algorithms, and analyzes the advantages of the system.

Key words grid domain ontology answer grid system; require of resource ontology matching of resource node

0 引言

随着计算机及现代远程教育技术的发展, 网上答疑已成为远程教学活动中的一个重要环节. 目前国内外许多研究机构和远程教育站点都对答疑系统进行了深入的研究. 国外在智能化方面比

较出色的智能答疑系统有 Ask Jeeves 公司的 Ask Jeeves for Kids 网上答疑系统, MIT 人工智能实验室开发的 START 系统^[1]. 但是由于国外的自动答疑系统不支持汉语, 限制了这些系统在我们国内的使用^[2]. 目前我国的网上自动答疑系统还处于探索阶段, 根据文献[3]对国内网上教育平台或

收稿日期: 2009- 09- 09

作者简介: 黄玲玲(1976-), 女, 湖南邵阳人, 南华大学计算机学院讲师, 硕士研究生. 主要研究方向: 智能信息系统与知识管理.

教育网站的调研,发现国内真正具有智能性的网上答疑系统非常少.虽然有些系统也采用关键字匹配和分词搜索等技术,但是由于自然语言语义理解不足,他们实现的仅仅是简单的、机械的智能,距离基于自然语言的、基于理解的智能还有相当距离^[4].文献[5]中提出一种基于本体的问答查询系统,利用领域本体能够包含完整的某一特定领域中的概念间的关系的描述来扩展语义,使得一定程度上解决目前答疑系统中语义的理解不足的问题成为可能.但是现有的答疑系统隶属于某个教育平台各自开发,导致资源重复建设;各个系统之间不能有效互操作,造成各答疑系统间资源无法共享和统一协调管理.

网格(Grid)^[6-7]是构筑在互联网上的一组新兴技术,它将高速互联网、高性能计算机、大型数据库、传感器、远程设备等资源融为一体,为人们提供更多的资源共享、信息服务和功能交互.它让人们透明地使用计算、存储以及其他各种资源.目前网格技术已经从计算网格发展成为面向服务的知识网格.知识网格是一个智能互联环境,它能使用户或虚拟角色有效地获取、发布、共享和管理知识资源,并为用户和其他服务提供所需要的知识服务,辅助实现知识创新、协同工作、问题解决和决策支持^[8].

鉴于以上所述,本文从资源共享和易于建设的角度出发,提出基于领域本体的答疑网格系统,系统中包含多个答疑网格节点,各网格节点有计划地建设答疑资源,利用网格技术屏蔽各节点的硬件和软件异构性将分散分布的答疑网格节点连在一起构成整个答疑网格系统,以实现答疑网格内各节点答疑资源的分布建设及各节点之间资源共享和统一协调管理,同时本体的引入既在一定程度上解决目前答疑系统中语义的理解不足的问题,又为基于本体服务的资源节点搜索提供依据.这种基于网格的建设方式实现了将一个大而复杂的系统开发工作分解为许多小的工作分别建设,减小了实际开发困难.

1 答疑网格系统

1.1 系统体系结构

答疑网格系统是一种四层架构,由资源层、核心服务层、答疑业务层和答疑门户层组成,如图 1 所示.

最底层是资源层:资源节点的答疑资源包括问题-答案资源库和本体库,问题-答案库是该

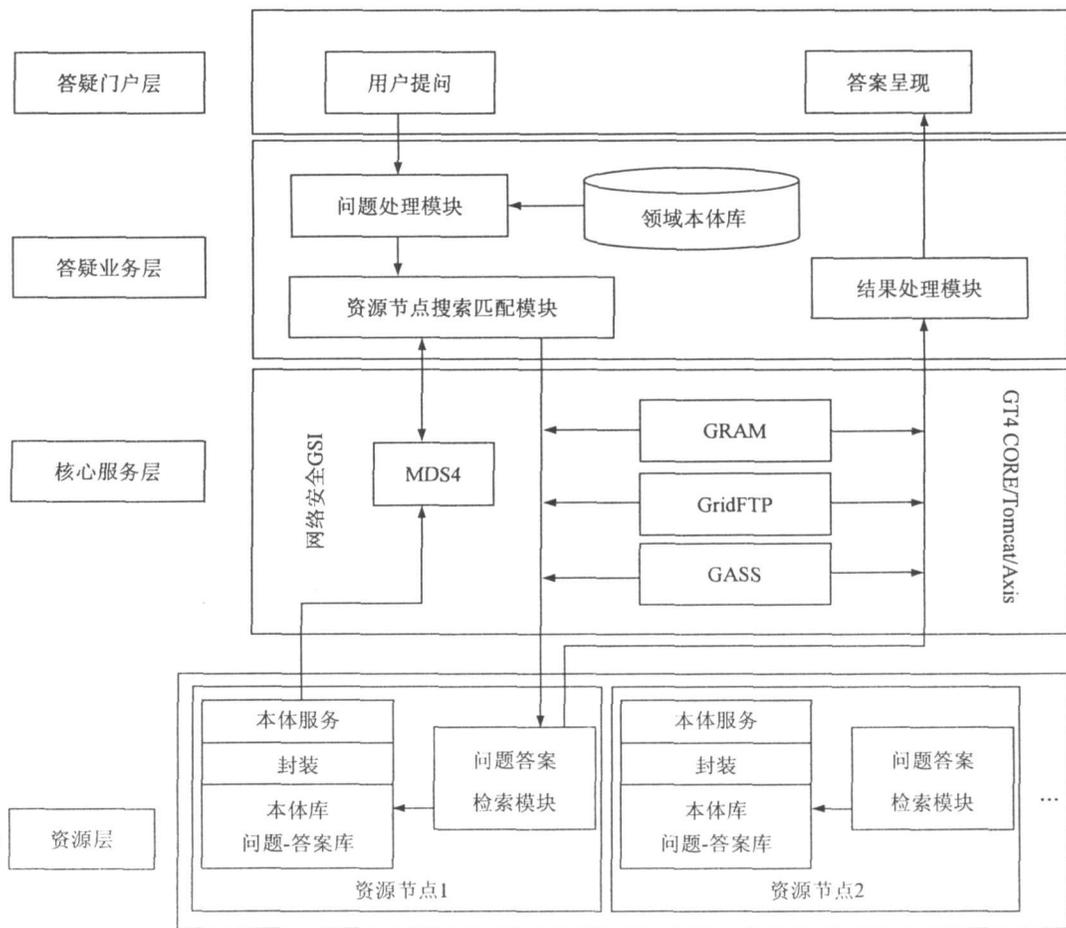
节点需要提供共享的真正资源实体,本体库是问题-答案库资源所包含所有本体的描述.资源层将本体库本体遵照 OGSA 规范描述封装成本体服务注册到 GT4 的信息服务中心 MDS4 本体服务对外提供该资源节点的问题-答案库资源的一个索引,用户通过在信息服务中心 MDS4 查询某节点注册的本体服务来确定某节点是否具有回答某问题的能力,即某节点注册的本体服务中能匹配用户提问所包含的本体,则表示该节点可能回答用户提问.问题答案检索模块提供对节点问题-答案库的问题搜索匹配,进而得到对应的答案.

第二层是核心服务层:核心服务层中包括网格应用所需要的基本服务,本文在 GT4 CORE / Tom cat/Axis 上构筑运行时环境,以较为成熟的网格中间件 Gbbus 为基础部署 GRAM, MDS4 OGSA - DA I GridFTP, GSI 等核心服务为基础构建底层支撑环境,提供资源管理、信息管理、安全认证等网格的核心支撑功能.

第三层是答疑业务层:答疑业务层体现答疑网格架构的领域特征,其中包含了在网格核心中间件基础上针对答疑领域开发的业务组件,用于支持答疑应用,其中包括问题处理模块、资源节点搜索匹配模块、结果处理模块和领域本体库.问题处理模块主要对用户使用自然语言的提问进行一系列处理后转换成资源请求本体,领域本体库在问题处理过程中起到将关键词转换为资源请求本体的作用,也是与各资源节点联系的纽带;资源节点搜索匹配模块将问题模块处理后得到的资源请求本体作为输入变量在信息服务中心 MDS4 中已注册的本体服务中进行搜索匹配得到包含资源请求本体的资源节点地址列表;结果处理模块将从各个节点搜索返回的结果集进行相关处理生成答案返回给用户.

最上层是答疑网格门户:网格门户为用户提供统一的、整合的界面使用网格,通过网格门户 Portal,用户将不必了解网格中的资源细节,在统一的接口下查询、调用信息资源.作为答疑网格服务的入口,答疑网格门户提供了用户提问以及答案呈现等功能.

该体系结构采用两级分布式本体模型建立各层之间的联系,各个网格节点采用本体技术将问题-答案库本体描述并封装成本体服务;领域本体库用于问题关键词向资源请求本体的转换;资源节点搜索模块所搜匹配出存在资源请求本体的资源地址列表.



(注: 图 1 中只画出资源节点 1 与其它各层关系, 其它资源节点与各层关系同资源节点 1)

图 1 答疑网格系统体系结构

Fig 1 The system architecture of answering grid system

1.2 系统答疑流程

用户只要通过答疑网格门户提问, 就可以使用平台中各个节点的答疑资源. 由答疑网格系统体系结构图可以看出, 系统答疑流程如下:

1) 用户通过网格答疑门户的用户提问模块提出问题后, 系统将问题交给答疑业务层.

2) 处于该层的问题处理功能模块将问题进行分词、关键词提取后形成关键词集合, 利用资源请求本体转换算法与领域本体库交互以进行关键词语义扩展, 生成资源请求本体.

3) 由于 MDS4 具有索引服务和触发服务的功能, 资源节点搜索匹配模块利用核心服务层提供的 MDS4 服务进行资源节点的搜索匹配工作, 获得具有资源请求本体的资源地址列表.

4) 与列表中的资源节点进行交互: 利用 GRAM 服务处理与任务相关的各种工作, 包括负责任务接收、任务监控、任务结果收集及提交等工作; 利用 GASS 服务进行远程文件的访问, 问题答案检索模块采用基于本体的语义相似度匹配进行

问题答案的搜索匹配, 并通过 GridFTP 服务将得到的结果集传输给结果处理模块.

5) 由于各节点内采用基于本体的语义相似度匹配, 所以结果处理模块对各节点内返回的结果集按语义相似度的高低将对应的答案排序输出.

1.3 系统关键实现技术

答疑网格门户提交的问题经过分词、关键词提取、资源请求本体转换、资源节点搜索匹配、节点内问题答案检索及结果处理等步骤地处理. 在这里, 分词和关键词提取技术比较成熟; 节点内问题答案检索采用基于本体的语义相似度匹配^[6]; 结果处理对语义相似度的高低将对应的结果集排序输出; 该系统实现的关键是两级分布式本体的构建及维护、资源请求本体转换和资源节点搜索匹配.

1) 两级分布式本体建立及节点本体服务描述、封装及发布: 首先构建领域本体库, 资源节点的问题本体是领域本体库的子本体, 构建时参考领域本体库, 同时各节点可以根据实际需要不断

的修正和完善领域本体库, 如各节点可以增加、更新本体, 添加、修改本体之间的关系等. 由于在 Gbbus 中 MDS 实现了基于 LDAP 的树状元数据目录服务来支持资源发现, 逻辑上可以基于 LDAP 树状结构来描述节点本体库中本体的元数据信息; 并且节点本体对外接口的描述必须遵照 OGSi 规范, 将接口用 WSDL/gWSDL 实现.

2) 资源请求本体转换利用领域本体库概念间的关系的将关键词转换资源请求本体, 资源请求本体分为两部分: 标准描述和扩展部分. 标准描述是从领域本体的同义关系表求出当前关键词的同义表达的标准描述. 同义关系表中列出该领域中同一个概念的不同的表达方式, 并采用该领域最常用的描述最为该概念的标准描述. 同义关系表在领域本体库建库时完成. 扩展部分利用领域本体库中概念间的关系如父子关系、兄弟关系扩展关键词的语义.

3) 答疑网格的资源分散分布在各个节点中, 通过资源节点搜索匹配算法得到能够回答某一问题的资源所在的节点, 算法简单描述如下:

Step1: 设资源请求本体为 O_{req} 即 $O_{req} = \{O_{standard}, O_{extension}\}$, 其中 $O_{standard}$ 表示资源请求本体标准描述, $O_{extension}$ 表示资源请求本体扩展部分;

Step2 设 $R = \{R_1, R_2, \dots, R_m\}$ ($m \geq 1$) 为 MDS4 中符合资源请求的资源候选集, R_k ($1 \leq k \leq m$) 为网格节点发布的本体服务中包含的本体集, 匹配资源列表 R 中的任一资源 R_k ($k = 1, 2, \dots, m$) 与资源请求本体 O_{req} 有以下几种情况:

①若 $O_{req} \cap R_k \equiv R_{req}$, 即在资源节点 R_k 中找到与 $O_{standard}$ 相同的本体且扩展关系与 $O_{extension}$ 一致, 则该节点完全匹配.

②若 $O_{req} \cap R_k \subset O_{req}$, 即在资源节点 R_k 中找到部分 O_{req} 本体, 则该资源节点可作为候选资源节点;

③若 $O_{req} \cap R_k$ 为空, 即在资源节点 R_k 中未找到 O_{req} 中任何本体, 则该资源节点不能作为匹配的节点;

Step3 返回资源节点地址列表. 若有完全匹配的节点, 则返回该资源节点; 若没有完全匹配的节点, 返回候选资源节点; 否则为空.

1.4 系统特点和优势

从答疑网格系统结构中可以看出, 基于领域本体的答疑网格系统具有如下特点和优势:

1) 将资源封装成服务. 通过将问题 - 答案库资源描述为本体, 并将本体封装成服务, 使得资源、信息统一起来, 使得分布式系统管理有标准的接口和行为. 用户可以通过网格操作系统透明地

使用整个网格答疑资源, 有利于实现资源共享.

2) 系统采用两级分布式本体模型, 使本体成为联系各层关系的纽带, 不仅扩展了语义, 在一定程度上解决目前答疑系统中语义的理解不足的问题, 使得其答疑准确率更高; 又为基于本体服务的资源节点搜索提供依据.

3) 实现了答疑资源分布式建设, 打破了答疑资源建设局限于本地库的限制. 在答疑网格中, 各资源节点独立建设答疑资源、发布资源信息, 各节点间相互独立且共享资源, 同时, 又通过 GT4 协调控制各资源节点行为.

4) 具有良好的开放性和扩展性. 由于采用标准的 OGSA 体系结构, 采用 GT4 为统一的中间件开发平台, 使答疑网格的开放性和扩展性体现在两个方面: 一是对答疑资源是开放的, 只要符合接口规范, 各节点都可以将资源封装为本体服务后加入到系统中; 二是 OGSA 中间件是开放的.

2 结论

基于领域本体的答疑网格系统把网格技术与智能答疑系统完全结合起来构建了一个充分利用高速网格计算的智能答疑系统. 很好地实现了答疑资源的分布建设与共享以及各节点资源的统一协调管理, 并且本体的引入提高了系统答疑准确率. 随着人工智能技术和远程教育技术的迅猛发展, 这种基于网格的开放式独立于教育平台的答疑系统将会逐渐走向成熟和完善.

参考文献:

- [1] Borís Katz, Gregory Marton, Gary Borchardt et al. External knowledge sources for question answering[C] // Proceedings of the 14th Annual Text REtrieval Conference (TREC2005), Gaithersburg US MD, 2005: 45-53
- [2] 赵金海. 基于 Web 的自动答疑技术研究及系统实现[D]. 重庆: 重庆大学, 2007.
- [3] 李爽, 陈丽. 国内外网上智能答疑系统比较研究[J]. 中国电化教育, 2003(1): 80-83.
- [4] 韩华, 戴亚非, 李晓明. 远程教学系统: 用户期望与技术挑战[J]. 计算机科学, 2001, 28(12): 1-4.
- [5] 赵赟. 基于本体的问答查询系统关键技术的研究与实现[D]. 南京: 东南大学, 2004.
- [6] Foster I, Kesselman C. 网格计算[M]. 金海, 译. 北京: 电子工业出版社, 2004.
- [7] 都志辉, 陈渝, 刘鹏. 网格计算[M]. 北京: 清华大学出版社, 2002.
- [8] Zhuge H. China's E-Science Knowledge Grid Environment[J]. IEEE Intelligent System, 2004, 19(1): 13-17.