《智能信息处理》课程考试

基于本体的信息集成框架的研究

张晓雪

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020年12月06日

基于本体的信息集成框架的研究

张晓雪

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要 现代企业在信息化建设过程中积累了大量的信息资源，但由于这些信息可能来源于不同的信息系统，表示和存储形式各异，存在语法和语义上的异构，难于共享和互操作，为企业的信息集成带来了新的挑战。传统的信息集成技术很难解决信息源之间的结构和语义的异构，因此，本文将本体技术引入到企业信息集成中来，对异构信息进行形式化和规范化表示，以期解决传统信息集成中存在的问题。本文主要分析了当前信息集成中存在的问题，介绍了基于本体信息集成的几种典型方法，提出了基于本体的信息集成框架，并详细描述了框架的各个组成部分。

关键词 信息集成 异构 信息源 本体 本体映射 本体进化

中图法分类号 TP311.20 DOI 号 10.3969/j.issn.1001-3695.2014.01.030

Ontology-Based Information Integration Framework

Zhang Xiaoxue

(Computer science and technology, Dalian maritime university, LiaoningDalian, 116026, China)

Abstract *Abstract The modern enterprises accumulate massive information resources in the informalization-process, but because these information possibly originates from different information system, expressed and memory-form respectively different, there are grammar and semantic isomerisms, it is difficult to share and operate mutually, this bing the new challenge for enterprise's information integration. It is very difficult to solve information source structure and semantic isomerism with traditional information integration method. Therefore, the thesis brings ontology technology into the enterprise information integration, carries on the formalization and the standardized expression to isomerism information, which is in order to solve the problems exist in traditional information integration.The thesis primarily analyzes the problem which exists in current information integration method,introduces several typical information integrations method, proposes information integration framework that based on ontology technology, and describes the components of the framework in detail.

Keywords Information integration;Isomeric information source;Ontology;Ontology Mapping;Ontology Evolution

1 引言

随着信息技术的不断发展，尤其是网络技术的高速发展，信息技术在企业中获得了广泛的应用，企业信息化程度和所积累的信息资源在飞升。但信息形式多样，缺乏统一的描述，给各企业信息资源的集成和管理带来诸多挑战，传统的信息集成方式很难满足企业的要求。基于本体的信息集成是目前企业信息集成研究的热点，能解决当前企业信息集成中出现的问题，实现企业中相关领域信息语义上的共享。

1.1 论文研究背景及问题提出

20 世纪 90 年代以来，经济和社会生活受到信息的驱动。然而，根据中国网络信息中心的调查报告显示，信息的数量在成倍增长，但质量与利用效率却不高。如何更有效地利用信息资源，是计算机科研及工程人员急需解决的问题。

当前企业中存在数量巨大的信息，它们以不同

形式存在于企业的各个部门，除了结构上的异构还存在语义上的异构，难于实现信息的有效共享。解决这个问题，不仅具有理论价值，同时也具有广阔的应用前景。

1.1.1 信息集成方式

现有的信息集成的方式大多是基于数据库的。这种方式使用多个数据库（信息源）的信息，并利用这些信息构成一个新的数据库，这个数据库可能是逻辑上的包含来自所有这些信息源的信息，因此可以作为一个数据单元来访问。信息源可能是传统的数据库，也可能是非传统的数据库，目前，数据库或者分布式信息源的集成方式常用的有三种。

①基于联邦数据库的信息集成。数据源是独立的， 但一个数据源可以访问其他数据源以提供信息。为了到达目的，联邦数据库实现了各个数据库之间的一对一连接。这些连接允许数据库 D1 以另一种数据库D2 理解的方式来查询D2。这种结构的问题是， 如果N 个数据库两两之间都要进行交互，则需要写N×（N－1）个接口的代码来支持相互查询。

②基于数据仓库的信息集成。来自几个信息源的数据副本存储在单一数据库中，称为数据仓库。存储在数据仓库中的数据在存储之前可能要经过一些处理，数据仓库集成结构中，来自各个信息源的数据被提取后组成一个全局模式。然后，数据存储在数据仓库里，在用户看来，与普通数据库无异。其组织方式如图1.1所示：

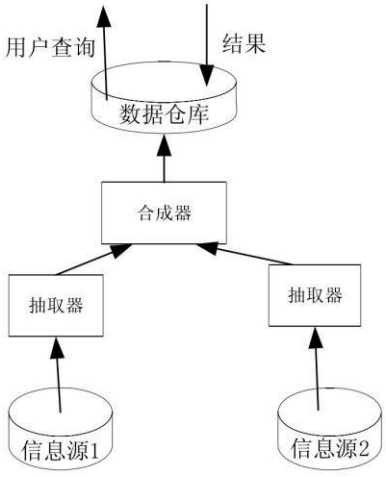

图1.1 数据仓库存取独立数据库的信息集成

一旦数据存储在数据仓库中，用户就可以查询。数据仓库一般采用定期更新方式，通常不允许用户对数据仓库进行更新。数据仓库与基本数据源的数据可能不一致。

③基于Mediator的信息集成。Mediator是一种软件组织，它支持虚拟数据库，用户可以查询这个虚拟数据库。Meditation不存储任何自己的数据，而是将用户的查询翻译成一个或多个数据源的查询。然后，Meditation将那些数据源对用户查询的回答进行综合处理，将结果返回给用户。

Mediator集成几个数据源的方式与物化了的数据仓库的集成方式有些相似，数据源通常多于两个。首先，用户提出的查询交给Mediator，因为没有自己的数据，Mediator必须从它的数据源中得到相应的数据，并使用这些数据形成对用户查询的回答。

Mediator每收到一个查询便将其分解成对应基本数据源的查询，经包装器处理之后传到基本数据源进行实际的查询，各个数据源的返回结果由Mediator组合后呈现给用户。其过程如图1.2所示：

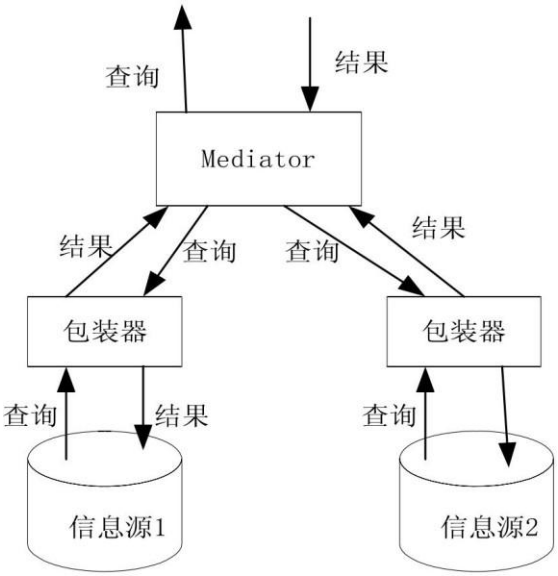

图1.2 Mediator数据集成方式

1.1.2 信息集成的主要问题

不同数据源中的数据一般都独立表示。给不同数据源中的原始数据赋予一些含义，具有相同的语义的数据，可能存在不同的表示。下面给出了在信息集成中需要解决的主要问题：

(1) 数据类型不同：例如，定长字符串的长度可能不同，一些数据源可能使用整数表示序列号。

(2) 值不同：同一概念在不同数据源中可能用不同的数据表示。

(3) 语义不同：许多重要概念在不同数据源中可能有不同的含义。例如：两个代理商可能在关系 CARS 中包含有车的信息的不同。

(4) 数据丢失：一个数据源可能不记录其它数据源提供的类型信息。例如，使用 NULL 或者缺省值。

现有的集成技术基本都是形式上的集成，本体技术的出现，使得概念在某个领域内有了明确的定义与描述。一定程度上为解决数据语义问题提供了可供参考的方法。

1.2 基于本体的信息集成方法适合应用于企业

基于本体的信息集成方法总体目标是获取促进信息资源的重用、信息资源共享和应用技术协作的技术与方法，改善目前信息资源管理中利用传统的技术难以检索和融合相关资源的状况。

南京长江油运公司所作为大型油运企业，承担着重要的运输任务。在长期的生产实践过程成中，积累了大量的信息资源。充分利用这些资源，可以提高企业运作效率，缩短运作时间。但这些信息资源以不同的形式存在于企业的各部门，由于它们在结构上存在很大差异，尤其是语义上的不明确与异构，使得相互之间不能互操作，很难查找和抽取有用的信息，给信息资源的有效利用造成了极大的障碍。基于本体的信息集成方法应用于该公司信息管理有效的解决了上述问题。

2 基于本体的信息集成框架

2.1 本体与信息集成

信息集成的主要内容是：集成不同硬件、不同操作系统、不同数据库管理系统和不同应用软件组成的异构数据处理环境下的数据。信息集成的目的是屏蔽底层数据源的异构性，提供给用户一个可以理解的、简明的视图。

2.1.1 本体研究现状

起源于哲学的本体论，近年来受到信息科学领域的广泛关注，在信息科学中的应用始于人工智能领域。作为知识库和知识系统构建技术的学科“知识工程”发展起来。通过复用，描述性的知识、问题解决方法以及推理服务都可以在系统间实现共享，从而可以构建更大更完善的知识库。同时，数据库管理系统领域（DBMS）的研究也逐步发现，但早期数据库的概念模型以专门化和不一致为特征，导致了数据集成的许多实际问题[3]。另外，伴随着面向对象技术的兴起，软件工程领域也开始意识到领域建模的重要性。上述的 3 个领域的发展面向了同样的一类问题，即对某一个领域进行通用概念的形式化描述。因为本体具有以下突出特点：

(1) 从功能上来讲，本体和数据库有些相似，但是本体表达的知识更丰富。

(2) 本体是领域内重要实体、属性、过程及其相互关系形式化描述的基础。这种形式化的描述可成为软件系统中可重用和共享的组件。

(3) 本体可以为知识库的构建提供一个基本的结构。以描述对象的类型而言：有简单事实及抽象概念，这些可以描述成一个本体的静态实体部分，它们主要描述的是事物或概念的各个组成部分以及这些组成部分之间的静态联系；本体也可以描述事物或概念的运动和变化。[1]

(4) 本体适合表示抽象的描述。企业模型是人们对企业或者企业的某些模型的抽象描述，在企业逻辑建模中，本体的使用可以帮助我们清楚地理解企业特定领域的相关元素、关系和概念，让知识表达更加准确便捷，帮助人们进行更好的企业决策[2]。

2.2 基于本体信息集成方法

本体技术应用于信息集成领域，产生了几种基于本体的信息集成方法。目前应用本体技术进行信息集成主要有三种方法：单本体方法、多本体方法和混合方法。

2.2.1 单本体方法

单本体方法使用一个全局本体提供的共享词汇表来表示信息的语义（图2.1 a）。所有信息源均与这个全局本体有关。所有信息源必须通过某种方式（如映射）与全局本体发生联系，全局本体的词汇是所有信息源词汇的综合。

2.2.2 多本体方法

在多本体方法中，每个信息源由自己的本体描述（图2.1 b），即可以为每个信息源抽象出一个单独本体。原则上，源本体可以是其他本体的综合，但不能是其他共享同一词汇表的不同源本体集合。多本体方法的优势首先在于不需要全局本体。各个源本体可以独立发展不需要考虑其他的源本体。本体结构很容易变换。

2.2.3 混合方法

类似多本体方法，在混合方法中每个信息源用自己的本体描述。但是为了确保源本体的相互兼容，它们建立在一个共享的词汇表上。这个共享的词汇表包含了某个域上的基本术语。通过将基本术语用某些操作综合起来，可以构建源本体中的复杂术语。源本体中的术语全部基于基本术语，因此术语比多本体方法更易兼容。某些时候共享词汇表也是一个本体（图2.1 c）。

混合方法的优势在于新的信息源可以很容易

的加入源本体，而不需要修改映射或者共享词汇表。共享词汇表的使用使得源本体兼容并且避免了多本体方法的弊端。混合方法的不足在于，已有的本体不易重用，因为所有源本体必须与共享词汇表相关。

表2.1 概括了各种本体集成方法的优缺点：

	单本体方法	多本体方法	混合方法
执行效果	直接	代价高	可接受
语义异构	视图相似	支持异构视图	支持异构视图
信息源增加/删除	需要调整全局本体	提供新的与其他本体相关的本体	提供新的源本体
与多本体方法的比较	—	缺乏共享本体，困难	使用共享词汇表，简单

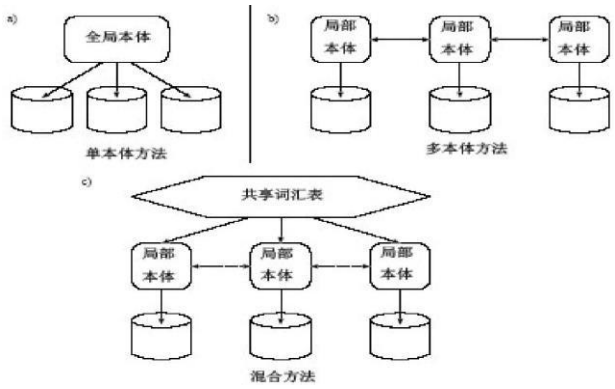

图 2.1 三种基于本体的信息集成方法

表2.1显示了三种信息集成方法中，混合本体是最实用，最有优势的一种集成方法。

2.3 信息集成中语义与本体的关系

传统的信息集成中，信息的表示方法大多是基于语法的，它能不做到的用户查询时得到条件语义相关的信息。在基于本体的信息集成中，本体技术是其应用的重要技术。

概念、符号、和现实对象之间的关系通常可这样描述：

1)概念和对象有着直接的联系，概念反映了对象的客观存在。

2)符号是概念的形式化表示，符号是人们对概念的具体表达，符号和概念也是直接联系的。

3)符号与对象之间不具有直接的联系，而是具有一定的任意性，即通常所谓的约定俗称。

人类对现实世界的认识是通过概念映射到事物上去的。各种概念与现实世界具体事物的映射是人类长期经验积累以及约定俗成产生的结果。

机器通过将符号向语义形式编码映射达到对符号的语义理解。在现实应用中，往往需要制定一些规范来定义语义概念的含义，以使得不同系统之间能够理解交互信息的含义。

语义和本体之间的关系可以用以下几点说明：

(1)概念是人类长期以来在实践的基础上，在某行业内约定俗成的符号与现实世界之间的一种映射关系。因此常常出现一词多义或者多词一义现象。

(2)现实中的概念不作改变，很难被机器或者应用程序所理解。也就难以将符号向现实世界进行映射。

(3)要实现对语义的理解，需要有一个完备的符号系统来实现概念和现实世界的对应或者映射。

(4)在机器表示中，语义是通过概念向现实世界进行映射的，本体是机器语义理解的最小单位。

(5)人们希望能够对资源进行语义上的查询，即需要系统能根据信息资源所具有的资源语义进行融合后提交给用户，查询机制是基于本体的。

2.4 基于本体信息集成框架

基于本体的信息集成框架如图2.2所示，基于本体的信息集成框架主要包括四层：信息源层、信息描述层、中间层和应用层。

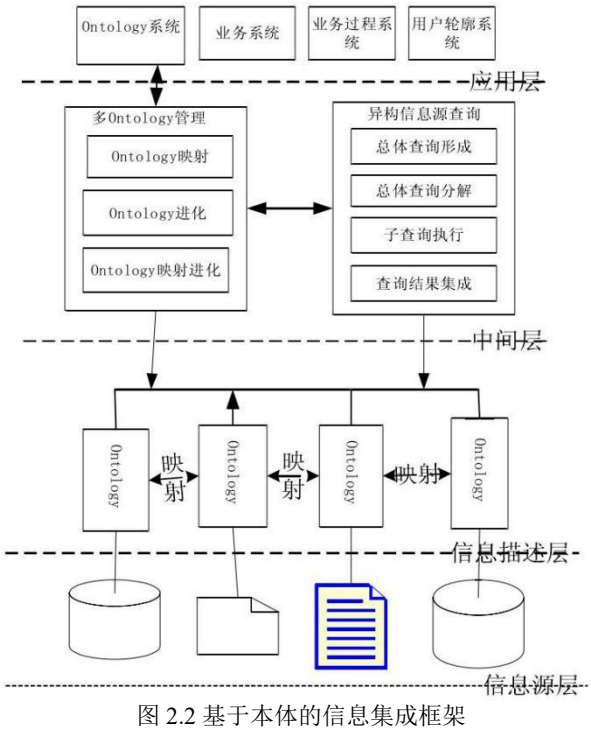

图 2.2 基于本体的信息集成框架

2.4.1 信息源层

信息源层包括企业中的不同信息源。企业中大量不同的信息源分布于各个部门之中。各个部门的应用不同，因此对信息源的表述以及存储采取了不同的格式，这导致了信息源的语法异构和语义异构问题。

信息源层的主要目的就是发现各个独立的信息源，并找出其语法以及语义上的异构，为上面的描述层提供相关的信息。

2.4.2 信息描述层

信息源层的各个信息源之间存在语法异构和语义异构，为解决语法和语义异构问题，每个信息源采用一个单独的Ontology进行描述，对每个信息源的概念以及关系清晰表述，来解决了异构问题。但是，多个Ontology之间仍然存在异构现象，这是因为：建立每个信息源对应本体时所采用的模式可能不同，或者不同的人对同一信息源采用不同的方法描述。

为了解决不同Ontology之间的通信和互操作，采用Ontology映射建立异构Ontology之间的语义互操作，从而实现异构信息源的互操作。描述层的主要功能就是试图寻找能解决上述问题的合理方案。

2.4.3 中间层

中间层是系统的关键，是实现系统信息服务的核心要素。包括用户信息需求、异构信息源查询、多Ontology管理等。

（1）用户信息需求

用户信息需求是系统的驱动因素，系统根据用户信息需求进行异构信息源查询，将查询结果主动提供给用户。用户信息需求分为信息需求生成和信息需求学习两部分。

用户信息需求生成是指根据用户自身特点和业务过程中用户所承担的任务和充当的角色等因素，确定用户的信息需求，用户个性化信息需求采用用户轮廓表示。但是由于用户对其兴趣不能确认或者不希望在创建轮廓时投入太多精力，所以，获取正确的用户轮廓是非常困难的。大多数研究采用术语及其权重表示用户轮廓[3][4][5]。在企业应用中，我们用用户知识熟悉程度及知识术语间的关联程度来表示用户个性化需求。采用以下形式对用户知识需求、用户轮廓、业务过程进行描述。

用户知识需求形式化描述为：

KN=(BKN,UP)，KN表示用户知识需求；$BKN = \{t_1,t_2,...,t_q\}$，表示业务需求的术语集合，描述业务知识需求；UP 表示用户轮廓，描述用户个性化知识需求。用户轮廓采用图描述为：UP=(V,E)，V = { $(t_1,f_1), (t_2,f_2), ..., (t_m,f_m)$ }为图的顶点，表示术语集合及其熟悉程度；$E= \{(v_i,v_j,rw_{ij})|v_i,v_j \in V\}$ 为图的边，表示术语之间的关联程度集合；

业务过程描述[6][7] 为： P=(T/(A,R),S)，$T=\{T_1,T_2,...,T_n\}$，为业务过程中的任务；$A=\{A_1,A_2,...,A_P\}$，为业务过程中的业务活动；$R=\{R_1,R_2,\ldots R_k\}$，为业务过程中的角色；$S=\{S_1,S_2,...,S_l\}$，为组织业务活动的策略。

令 $F=A\times A\times C$ 或 $F=R\times R\times C$ 描述活动或者角色的改变。C 中的每个元素 c 为布尔表达式，表示单一活动或角色中是否有某些因素发生了变化。则 F 描述了当条件 c 的计算结果为真时，活动或角色即发生了变化。

信息需求学习分析用户对提供信息的反馈，根据用户所从事的工作经历和对系统推荐的信息源的反馈状况，对用户的个性化信息需求即用户轮廓进行修改，以反映用户信息的动态变化。

将查询获得的信息推荐给用户，供用户选择使用。用户通过阅读信息，完成相应的业务活动，其知识水平会随之而变化。因此，为反映用户个人信息需求变化，必须对用户个性化信息需求进行学习。用户信息需求进化算法学习原则为：

用户对于主动提供的信息，有两种处理方式。一种处理方式是阅读，另一种处理方式是不阅读。由于为用户提供的信息源是完成业务活动所必备的，即缺少任一信息源都不能够完成业务活动。若用户阅读信息源，则完成业务活动后相应信息的熟悉程度增加。若不阅读，则说明用户对于该信息源已足够熟悉，则将相关信息术语熟悉程度设置为熟悉程度极限值。用户完成一业务活动后，所完成活动的信息术语和阅读的信息源的信息术语，将会增加到用户轮廓之中，对于新增加的信息术语，其熟悉程度设置为熟悉程度初始值。

用户完成业务活动后，在业务活动信息需求和阅读的信息源中，同时出现的信息术语之间的关联程度增加。信息需求学习的有关算法详见文献[8]。

多 Ontology 管理主要包括 Ontology 映射、Ontology进化、Ontology映射进化。多Ontology 映射的建立非常耗费时间，如采用非自动方法，时间上的延迟难以让人接受，因此，必须采用半自动或全自动方法建立Ontology映射。由于企业中的信息

源不断变化，例如：有新的业务加入，就需要增加新的术语以及新的关系；相应的Ontology必须做出相应的变化，Ontology进化需要处理变化表示、变更传播、一致性检测等任务，Ontology进化有很多问题现在还是难于解决，这也是当前本体技术研究的热点之一。Ontology的变化会引起相应Ontology映射的变化，Ontology映射方法本身也要随之进行进化。

异构信息源查询根据用户信息需求，系统自动查询企业所有的信息源。异构信息源查询根据用户信息需求中的术语要求形成总体查询，通过Ontology映射将总体查询分解为针对各特定信息源的子查询，子查询在各特定信息源上进行执行，将执行后的查询结果集成后反馈给用户，供用户参考。反馈给用户的结果集大多以视图的方式呈现。异构信息源查询主要需解决查询分解的问题。

2.4.4 应用层

应用层主要包括企业业务系统、业务过程系统、Ontology 系统、用户轮廓系统等。

企业业务系统，主要包括用来完成企业业务活动的软件系统，如CAD、CAE、ERP 等。

业务过程系统，主要分为过程建模系统和过程执行机两部分。过程建模系统对企业业务过程进行建模，过程执行机执行业务过程实Ontology 系统，用来建立描述各信息源的Ontology。

用户轮廓系统，提供用户轮廓的建模、修改功能。应用层需要根据用户实际的需求采用适当的系统。

2.4.5 信息集成框架所需的有关技术

基于本体信息集成框架的实现需要用到若干关键技术。

(1) 为对信息源中概念以及概念间的关系有一个清晰的表述，必须考虑到采用何种描述逻辑的问题，需要用到描述逻辑的若干技术。

(2) 信息源本体建立后，为了屏蔽各个信息源本体的异构，实现各个知识源本体之间的通信和互操作，必须进行本体间的映射。

(3) 信息源总是处在不断的变化当中，因此必须考虑本体本身以及本体构建规则的变化，也就是本体进化问题。

(4) 为了实现全局查询能够分离为各个信息源的查询，对于用户提出的查询如何进行分解，即所谓的分解策略问题也是必须考虑的。

结束语

在网络技术的高速发展的时代，信息技术不断应用于企业，但信息存在的形式是各式各样的，它们的不统一性给各企业信息资源的集成和管理带来了各种问题。主要表现在以下几个方面：(1)如何使被管理的信息资源具有应用程序能够理解的含义，实现信息资源处理过程自动化、智能化；(2)如何对特定领域中积累的大量信息资源进行有效管理，使用户可以找到与需求相关的信息资源；(3)如何根据信息资源所具有的领域知识含义，将分散在各种异构系统中的相关信息方便、快速地融合后呈现给用户。基于本体的信息集成，重点在于解决当前企业信息集成中出现的大量的问题。基于本体信息集成方法在企业应用中同传统的信息集成方法相比，优势在于：①更有效地利用信息资源；②解决信息在结构上的异构和语义上的异构。本文分析当前企业信息集成的主要技术及存在的问题，阐述了本体与信息集成的关系、基于本体信息集成的方式以及阐明了基于本体的信息集成框架等。

参 考 文 献

[1] Lu Ruqian,Zhang Songmao. PANGU-An agent-oriented knowledge base[J]. In Processing of Conference on Intelligent Information Processing (16th WCC 2000): 486-493.

[2] 石杰,宿彦,史晓峰. 知识工程中的本体论研究[J]. 西安电子科技大学学报(社会科学版), 2004(02):1-3.

[3] Kyung-Yong Jung,Kee-Wook Rim,Jung-Hyun Lee. Automatic Preference Mining through Learning User Profile with Extracted Information[C]. SSPR&SPR2004,LNCS3138,2004:815–823.

[4] Yue feng Li,Y.Y. Yao. User Profile Model:A View from Artificial Intelligence[C]. RSCTC 2002,LNAI 2475,200:493–49.

[5] Ugene Santos,Hien Nguyen. Empirical Evaluation of Adaptive UserModelingina Medical Information Retrieval Application[C]. UM 2003,LNAI 2702:292–296.

[6] 徐焕良. 企业知识资源计划及其关键技术研究[D]. 南京航空航天大学博士论文. 2003. 10.

[7] 徐焕良,李绪荣,丁秋林.基于角色模型的业务过程再工程（BPR）的研究[J]. 计算机科学. 2003.Vol.30 No.1:154-157.

[8] 张磊,谢强,王金栋,丁秋林. 基于业务过程的知识需求研究[J]. 吉林大学学报（信息科学版）. 2005.Vol 23 No.5.

[9] Haghgoo Maliheh, Sychev Ilya, Monti Antonello, et al. SARGON – Smart energy domain ontology. 2020, 2(4):191-198.

[10] 岳丽欣,刘文云. 国内外领域本体构建方法的比较研究[J]. 情报理论与实践, 2016, 39(08):119-125.