

基于语义的领域虚拟数据中心研究与实现

李华昱^{1,2} 胡长军² 欧阳纯萍² 叶银珠³

(中国石油大学(华东)计算机与通信工程学院 东营 257061)¹

(北京科技大学信息工程学院 北京 100083)² (中国石油勘探开发研究院 北京 100083)³

摘要 油井工程领域涉及的数据具有分布、异构和自治性强等特点,并且存在复杂的语义关联,无法为全局决策提供有效的数据支持。利用本体和虚拟视图集成技术,提出了油井工程虚拟数据中心解决方案,通过构造领域全局本体并利用本体抽取、本体映射和查询转换实现领域数据语义集成,提供了统一和基于语义的数据查询与共享服务。实际应用表明,虚拟数据中心能够为油井生产决策提供全面和实时的数据支持。

关键词 虚拟数据中心,本体,语义映射,油井工程

中图分类号 TP391 文献标识码 A

Research and Implementation of Domain-specific Virtual Data Center Based on Semantics

LI Hua-yu^{1,2} HU Chang-jun² OUYANG Chun-ping² YE Yin-zhu³

(College of Computer and Communication Engineering, China University of Petroleum, Dongying 257061, China)¹

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)²

(Research Institute of Petroleum Exploration & Development, PetroChina, Beijing 100083, China)³

Abstract The data involved in the field of oil-well-engineering are distributed, heterogeneous and autonomous, and there is also a complex semantic association among them. Under this circumstance, effective global support for decision-making is difficult to ensure. Using integration technology of ontology and virtual view, this paper presented a virtual data center solution for oil-well data integration. Through global-ontology construction, ontology extraction, ontology mapping and query conversion, the data in oil-well engineering field can be semantically integrated. Consequently, a unified and semantic-based data query and shared services were achieved. By practical application, this virtual data center is able to supply production decision with comprehensive and real-time data services.

Keywords Virtual data center, Ontology, Semantic mapping, Oil-well engineering

随着石油工业信息化的快速发展,油井工程领域积累了庞大的数据资源,如生产、措施、作业、井下和设备数据等。这些数据呈现以下特点:面向不同的应用系统,在模式、语义等层面存在差异;分布存储,数据之间存在复杂的语义关联;数据量大、动态变化、实时性要求高等。这些特点导致数据交互变得复杂、困难,迫切需要进行语义数据集成,提供统一和基于语义的数据访问接口。

针对油井工程领域数据特点,采用虚拟视图集成技术和本体技术,本文建立了基于语义的油井工程虚拟数据中心 We-VDC (Oil Well Engineering Virtual Data Center)。WeVDC 以油井工程本体 WeOnto (Oil Well Engineering Ontology) 作为全局数据访问视图(即虚拟视图),提供基于领域概念的语义查询和共享服务,为油井生产决策提供全面和实时的数据支持。

本文第1节给出了虚拟数据中心概述;第2节描述了

WeVDC 架构层次及实现流程;第3节描述 WeVDC 中的语义抽取、语义映射和语义查询实现方法;第4节简要介绍了 WeVDC 的原型实现;最后得出结论。

1 虚拟数据中心概述

虚拟数据中心 VDC (Virtual Data Center) 是采用虚拟视图集成技术,提供统一、透明的数据访问视图与共享服务的数据集成框架。VDC 利用全局模式与数据源局部模式之间的映射将用户提交的全局查询请求转化为对底层数据源的查询访问并返回查询结果。VDC 只保存部分数据源元数据及模式映射信息,数据仍然分布在数据源中。由于结果直接来自数据源,因此 VDC 能够满足数据实时性要求,从而避免数据频繁加载引起的资源消耗,提高了系统运行效率。随着计算机处理能力和网络带宽的提升,虚拟视图法正成为数据集成的主流方法。

收稿日期:2010-03-03 返修日期:2010-05-18 本文受国家 863 高技术研究发展计划基金项目(2008AA01Z109),国家“十一五”科技支撑计划(2006BAK11B03)资助。

李华昱(1977—),男,博士生,讲师,主要研究方向为领域数据工程,E-mail: lhy_21th@sina.com;胡长军(1963—),男,教授,博士生导师,主要研究方向为数据工程、高性能计算等;欧阳纯萍(1979—),女,讲师,主要研究方向为领域数据工程;叶银珠(1983—),女,博士生,主要研究方向为油气井工程。

目前,VDC的研究与应用主要面向特定领域,为领域生产与科研提供数据支持。如美国政府的COSMOS VDC^[1](强震观测虚拟数据中心)和我国的CEDC(国家地震科学数据共享中心)通过集成地震观测点数据为地震预测与研究提供数据检索和共享;哈佛大学的Dataverse VDC^[2]面向社会科学领域实现研究成果的上传和共享。

随着语义Web^[3]的广泛应用,利用本体技术实现领域虚拟数据中心成为当前的研究热点。如INDUS^[4]对生物科学异构数据源进行集成,通过建立用户本体与数据源本体之间的映射实现数据查询与共享;针对中医药遗留关系数据库,浙江大学Chen Huajun等开发的DartGrid^[5]通过将数据表映射为全局本体的一个虚拟数据视图,实现异构数据语义集成,为中医药研究提供数据支持。

2 WeVDC实现架构

2.1 体系结构

WeVDC采用4层架构实现:数据源层、包装器层、虚拟数据视图层和数据服务层,如图1所示。

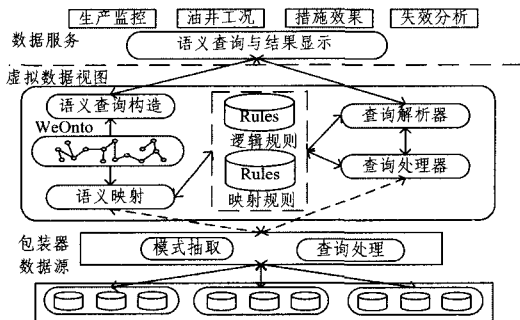


图1 WeVDC实现架构

1)数据源层:包括油井日常生产、措施作业、工况诊断、井下设备等专业数据库,这些数据主要以关系数据库形式存在。

2)包装器层:负责抽取关系数据源的模式信息,执行分解后的全局查询语句。

3)虚拟视图层:该层是WeVDC框架的核心,负责数据源语义集成和查询转换,为数据服务层提供虚拟数据访问视图,主要包括WeOnto、模式映射规则库、逻辑规则库、查询构造器和查询处理器等组件。WeOnto形式化描述了油井工程领域概念以及概念之间的联系,为数据服务层提供统一和基于语义的全局数据访问视图;模式映射规则存放数据源局部本体以及局部本体与WeOnto之间的映射,为查询转换提供参照信息;逻辑规则库保存定义的逻辑规则,这些规则用以增强WeOnto中类之间和属性之间的关联关系;查询解析器对提交的SPARQL查询进行解析,以获取查询需求;查询处理器利用全局本体与局部本体、局部本体与数据源之间的映射,将解析的查询语句转换为底层数据源的数据查询。

4)数据服务层:为油井生产决策提供语义查询和结果显示界面。语义查询界面允许用户采用领域术语定义查询需求并将其表示为SPARQL查询语句,提交给虚拟视图层;结果显示界面对返回查询结果进行组合,以表格或图形方式显示。

2.2 框架实现流程

首先给出WeVDC中的虚拟数据视图SvDV(Semantic Virtual Data View)的定义。SvDC为5元组: $SvDC = \langle G, L, D, M_{GL}, M_{LD} \rangle$ 。其中,G代表领域全局本体,表示全局数据访

问视图, Σ_G 表示G上的公理集合;L是局部本体,是对数据源的语义描述,L上的公理集合记为 Σ_L ;D表示数据源,数据模式记为 Σ_D ;M_{GL}描述全局本体与局部本体之间的映射关系;M_{LD}描述局部本体与数据源模式之间的映射关系。

基于以上定义,WeVDC框架实现流程描述如下:

- 1)建立领域全局本体G,对G添加公理约束。
- 2)抽取每个数据源D_i的模式信息 Σ_D ,利用参照完整性和实体完整性约束构造数据源局部本体L_i。
- 3)建立映射M_{LD}。基于 $\Sigma_D \cup \Sigma_L$ 上的公理集,在每个D_i与其局部本体L_i之间建立映射,映射关系通过构造专门的数据源描述本体实现。
- 4)建立映射M_{GL}。基于 $\Sigma_G \cup \Sigma_L$ 上的公理集,在全局本体G与每个局部本体L_i之间建立映射,映射关系通过构造专门的映射本体实现。
- 5)对全局本体G进行解析,构造语义查询界面。
- 6)实现语义查询。利用查询解析器与查询处理器将提交的全局查询分解为底层数据源的查询请求,完成数据查询。

3 虚拟视图实现

SvDC的实现过程主要包括WeOnto本体构造、数据源局部本体抽取、M_{LD}和M_{GL}语义映射等步骤。

3.1 WeOnto本体构造

参考Noy等^[6]、Villanueva-Rosales等^[7]和Kashyap等^[8]提出的本体构造方法,如图2所示,首先明确WeOnto的范围和需求,识别核心领域术语;确定核心概念并建立核心概念之间的层次与关联;进一步细化模型中的概念层次和属性;为领域概念增加约束定理。

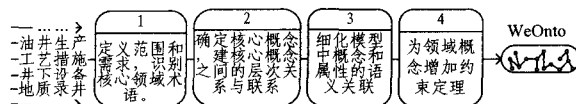


图2 WeVDC构造过程

利用Protégé本体编辑工具完成WeOnto构造,WeOnto类及类之间的关系(部分)如图3所示。

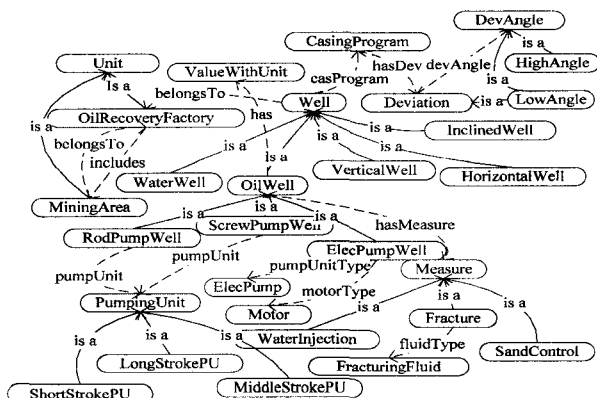


图3 WeOnto类以及类之间的关系(部分)

以下主要介绍增加WeOnto公理约束的方法及过程。增加公理约束主要包括增加必要条件和充分必要条件。

1)增加必要条件。对WeOnto中的类增加语义信息。例如,对斜井类InclinedWell设定限制性公理,即斜井的井身结构具有中等的井斜角,其公理描述如下:

$$\text{InclinedWell} \sqsubseteq \text{Well} \sqcap \exists \text{casProgram.} (\text{Deviation} \sqcap \forall \text{de-}$$

系;*dsdonto:references* 属性描述了参照表的主码;*dsdonto:isNullable* 和 *dsdonto:datatype* 属性描述了列的空值约束和数据类型。此外,设计 *dsdonto:relatedClass* 和 *dsdonto:relatedProperty* 属性描述表、列与本体中的类、属性之间的关联。

在建立映射关系时,需要抽取数据源模式信息即元数据作为实例数据填充到 DSDOnto 中,然后进行映射关联。以下代码给出了利用 DSDOnto 本体描述 M_{LD} 映射的一个实例。该实例表示数据源 D_1 中的工艺措施信息表 *JobInfo* 与其局部本体 $Onto_{D1}$ 中的 *Measure* 类关联。

```
<dsdonto:Table rdf:about="&.dsdonto;JobInfo">
  <dsdonto:relatedClass rdf:resource="&.ontod1;Measure">
  <dsdonto:belongsToDatabase rdf:resource="&.dsdonto;D1">
  <dsdonto:hasColumn rdf:resource="&.dsdonto;Well">
  <dsdonto:hasColumn rdf:resource="&.dsdonto;WellType">
  .....
```

</dsdonto:Table>

2) M_{GL} 的实现

设计映射本体 MapOnto 描述全局本体 WeOnto 与各个局部本体之间的映射关系。

针对类之间的等价和父子关系,利用 OWL 中提供的描述符(*equivalentClass*, *subClassOf* 等)来表示^[11]。以下 MapOnto 中的 OWL 代码表示 WeOnto 中的 *OilWell* 类与局部本体 $Onto_{D1}$ 中的 *WellInfo* 类等价:

```
<rdf:Description rdf:about="&.ontod1;WellInfo">
  <owl:equivalentClass rdf:resource="&.weonto;OilWell"/>
<rdf:Description>
```

针对属性之间的等价和父子关系,利用 OWL 中所提供的描述符(*subPropertyOf*, *equivalentProperty* 等)来表示^[11]。此外,对于某些复杂映射关系,则通过逻辑规则进行描述。例如,在数据源 D_1 中措施效果类 *JobEffect* 的月增油 *oilIncrMon* 属性属于效果指标 *effectIndex* 中的月增液 *liquidIncrMon*,而 WeOnto 中的月增油 *oilIncrMonth* 直接属于生产指标 *outPutQuota*,这种映射可表示为

```
ontod1:JobEffect(? x) ^ ontod1:effectIndex(? x, ? m) ^ ontod1:liquidIncrMon(? m, ? y) ^ ontod1:oilIncrMon
(? y, ? z) ->
(weonto:MeasureEffect(? x) ^ weonto:outPutQuota
(? x, ? y) ^ weonto:oilIncrMonth(? y, ? z))
```

3.4 语义查询

语义查询是利用 M_{LD} 和 M_{GL} 语义映射信息将提交的 SPARQL 查询语句转换为底层数据源的 SQL 访问,并将包装器执行 SQL 后的查询结果返回给数据服务层。实现步骤如下:

1) 查询构造器获取用户设置的查询需求,生成全局查询语句 Q_G , Q_G 采用 SPARQL 描述。

2) 根据 MapOnto 映射信息,利用逻辑推理机进行 TBox 推理,将查询 Q_G 中的全局本体中的概念/属性及其子概念/属性转换为相应局部本体中的概念/属性。

3) 将查询 Q_G 分解为针对每个局部本体的子查询 Q_{L1} , $Q_{L2}, \dots, Q_{Ln}$ 。

4) 根据 DSDOnto 映射信息,将 Q_{L1} , $Q_{L2}, \dots, Q_{Ln}$ 重写为针对每个数据源的本地子查询 Q_{D1} , $Q_{D2}, \dots, Q_{Dn}$ 。

5) 包装器执行本地子查询,按照统一的语法格式返回查询结果 R_{D1} , $R_{D2}, \dots, R_{Dn}$ 。

6) 将 R_{D1} , $R_{D2}, \dots, R_{Dn}$ 组合在一起形成最终查询结果 R_G 。

4 WeVDC 实现原型

根据虚拟数据中心构造方法,基于 Java 技术建立了基于 Web 的油井工程虚拟数据中心原型系统 WeVDC。WeVDC 采用 JavaApplet 实现用户查询界面,利用 Jena 作为访问 OWL 本体的 API。Jena 是面向语义 Web 应用的一个 Java 框架,包含本体解析和推理子系统。推理机采用 Pellet,包装器采用 JDBC 访问关系数据库。

WeVDC 主要由 3 个 Java 包实现,即 user, vdc 和 wrapper。user 包中的类负责实现交互界面,主要包括查询构造器和查询结果显示界面的实现;vdc 包中的类负责对全局本体、各个局部本体和实现映射关系的本体进行解析,提供类和属性的查找及匹配,实现虚拟数据访问视图;wrapper 包中的类负责局部数据源概念的查找和查询结果的执行与封装。系统实现主界面如图 5 所示。

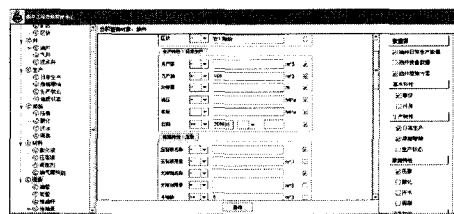


图 5 WeVDC 原型系统主界面

用户从界面左侧选择领域分类信息,确定要查询的领域对象;根据选取的对象列出该对象具备的属性信息,即在界面右侧显示相应的“基本特性”、“生产特性”、“措施特性”和设备特性等,这些特性会根据用户选取概念的不同而动态变化;通过点击特性标识按钮,可以进一步选择要查询的具体性能,界面中间部分会根据用户选择的特性及具体性能条目进行动态更新,每个条目右侧的复选框标识该条目是否作为查询选项显示在返回结果中;界面右侧上部列出了已经注册到 WeVDC 的异构数据源,用户可以选择要查询的数据源;用户输入相关的信息后,点击查询按钮,WeVDC 会生成相应的 SPARQL 查询,进行数据查询。

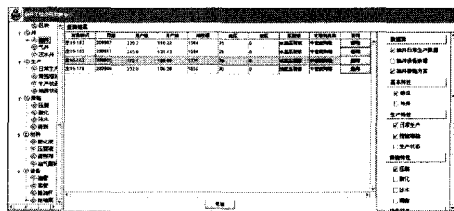


图 6 查询结果

图 5 显示了用户设置的查询需求:“在 2009 年 5 月以后月产量大于 100m^3 、位于岔 15 断块、实施了压裂措施并且月增油大于等于 5m^3 的油井,同时返回这些井的月产液、月产油、动液面、流压、套压和使用的压裂液及支撑剂名称等信息。”通过数据查询处理,返回结果被合并后将显示在一个 Web 页面上,如图 6 所示。查询结果页面提供了更为详尽的查询关联,用户可以通过点击每一条查询结果记录中的“查询”按钮查看针对该记录的更多相关信息。

结束语 利用本体和虚拟视图集成技术对领域分布异构数据源进行语义集成,构建虚拟数据中心,是目前一个研究与

应用热点。领域虚拟数据中心能够为全局应用提供有效的数据查询与数据共享服务。针对油井工程领域,本文提出了基于语义的油井工程虚拟数据中心的实现框架。其中,对于框架实现过程中的两个关键步骤,即全局本体构造和全局本体与数据源模式之间的语义映射进行了详细描述。利用全局本体实现虚拟数据访问视图,并借助全局本体与局部本体、局部本体与数据源模式之间建立两级映射关系,能够将基于领域概念的查询请求转换为底层数据源的 SQL 访问,以实现统一、透明的数据查询与共享服务。

通过在华北油田进行实际应用,油井工程虚拟数据中心能够为油井生产决策提供全面和实时的数据支持。

参 考 文 献

- [1] Archuleta R J, Steidl J, et al. The COSMOS Virtual Data Center: A Web Portal for Strong Motion Data Dissemination [J]. *Seismological Research Letters*, 2006, 77(6): 651-658
- [2] King G. An introduction to the Dataverse Network as an infrastructure for data sharing [J]. *Sociological Methods & Research*, 2007, 36(2): 173-199
- [3] Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. *Scientific American Magazine*, 2001, 284(5): 34-43
- [4] Caragea D, Pathak J, Bao J, et al. Information Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data Sources [C]// San Diego, CA, USA. *Proceedings of the Data Integration in Life Sciences (DILS 2005)*. Berlin: Springer-Verlag, 2005: 175-190

(上接第 184 页)

局、色彩搭配等。其中包含的若干 QA 服务均能随着时间的推移,动态改变其答案内容。由于本文的 IQAS 组成单元是 Web 服务,因此用户不仅可以通过提问的方式来获取或构建 QA 服务,还可直接要求系统提供某些服务,例如用户要求提供一个能够发短信的服务,也集成到该界面中。

结束语 本文提出以流程服务组合为核心的 IQAS 构造方法,即利用 Web 服务自身提供的操作及各种信息,通过交互协同,回答用户的提问。根据 RGPS 需求元描述框架,在领域模型的指导下,用户提问被转换为 QA 需求,通过按需服务发现以及基于 Web 服务关系网络的服务自动组合机制,最终生成能够回答此类问题的 QA 服务。我们开发了一个面向交通领域的智能按需问答原型系统 Uniponse,同时搭建了 Web 服务管理库和 QA 服务管理库,该系统验证了本文提出方法的可行性和有效性。

在未来的工作中,我们还将进一步研究 QA 需求的验证方法,即系统构造的 QA 服务是否能够较好地回答用户提问以及 QA 服务在用户问题发生变化时的动态演化算法。

参 考 文 献

- [1] Eliza[EB/OL]. <http://www-ai.ijs.si/eliza/eliza.html>, 1966
- [2] 余正涛,樊孝忠,郭剑毅,等. 基于潜在语义分析的汉语问答系统答案提取[J]. *计算机学报*, 2006, 29(10): 1889-1893
- [3] 孙昂,江铭虎,贺一帆,等. 基于句法分析和答案分类的中文问答系统[J]. *电子学报*, 2008, 36(5): 833-839
- [4] 李良富,樊孝忠,李宏乔. 语义相似计算驱动领域自动问答[J]. *北京理工大学学报*, 2005, 25(11): 958-962
- [5] Lee Cheng-wei, Lee Yi-hsun, Hsu Wen-lian. Exploring Shallow

- [5] Chen H, Wang Y, Wang H, et al. Towards a Semantic Web of Relational Databases: a Practical Semantic Toolkit and an In-Use Case from Traditional Chinese Medicine [C]// *Proceedings of the 5th International Semantic Web Conference*. Berlin: Springer, 2006: 750-763
- [6] Noy N F. Semantic integration: a survey of ontology-based approaches [J]. *ACM SIGMOD Record*, 2004, 33(4): 65-70
- [7] Villanueva-Rosales N, Dumontier M. yOWL: An ontology-driven knowledge base for yeast biologists [J]. *Journal of Biomedical Informatics*, 2008, 41(5): 779-789
- [8] Kashyap V, Cheung K H, Doherty D, et al. Ontology-based Data Integration for Biomedical Research [M]// Jain R, Sheth A. eds. *The Semantic Web Real-world Applications from Industry*. New York: Springer US, 2008: 97-122
- [9] Ghawi R, Cullot N. Database-to-Ontology Mapping Generation for Semantic Interoperability [C]// *Proceedings of the 3rd International Workshop on Database Interoperability*. New York: ACM, 2007: 43-50
- [10] De Laborda C P, Conrad S. Relational. OWL: a data and schema representation format based on OWL [C]// *Proceedings of the 2nd Asia-Pacific conference on Conceptual Modeling*. Australia: Australian Computer Society, 2005: 89-96
- [11] Seksun S, Djamal B, Philippe T. OWL-based Approach for Semantic Interoperability [C]// *Proceedings of the 19th International Conference on Advanced Information Networking and Applications*. Washington, DC: IEEE Computer Society, 2005: 145-150

Answer Ranking Features in Cross-Lingual and Monolingual Factoid Question Answering [J]. *Computational Linguistics and Chinese Language Processing*, 2008, 13(1): 1-26

- [6] Kikuchi Y, Hidaka K, Nakajima S, et al. Chat-Robot-based Web Content Presentation Interface and Its Evaluation [J]. *Human-Computer Interaction*, 2007, 4553: 934-943
- [7] McGuinness D L. Question answering on the semantic Web [J]. *IEEE Intelligent Systems*, 2004, 19: 82-85
- [8] Zhang Pengfei, Chunhua Wu, et al. Personalized Question Answering System Based on Ontology and Semantic Web [C]// *IEEE International Conference on Industrial Informatics*, 2006: 1046-1051
- [9] Chen Zhe, Wen Dunwei. A new Web-service-based architecture for question answering [C]// *IEEE International Conference on Natural Language Processing and Knowledge Engineering*. 2005: 179-183
- [10] Jang Minsu, Sohn Joo-chan, Cho H K. Automated Question Answering Using Semantic Web Services [C]// *The 2nd IEEE Asia-Pacific Service Computing Conference*, 2007: 344-348
- [11] Braga D, Ceri S, Daniel F, et al. Optimization of MultiDomain Queries on the Web [C]// *Vldb'08*. Auckland, New Zealand, August 2008: 562-573
- [12] Wang Jian, He Keqing, et al. RGPS: A Unified Requirements Meta-Modeling Frame for Networked Software [C]// *Proceedings of the 3rd International ACM Workshop on Applications and Advances of Problem Frames (IWAAPF'08)*. Leipzig, Germany, 2008: 29-35
- [13] Liu Wei, He Keqing, et al. Requirements Dynamic Elicitation and Analysis: Heavyweight Semantic Inducement Solution [J]. *Special Issue on Software Engineering and Complex Networks, DCDIS Series B*, 2007, 14(S6): 238-245