

基于本体的石油开发领域知识构建研究

杜睿山 尚福华 吴雅娟

(大庆石油学院计算机与信息技术学院,大庆 163318)

摘要 为了解决石油开发领域多专业之间由于信息术语不统一,而造成的不能信息共享和应用集成的问题,将本体引入石油开发领域的知识构建。通过本体的形式化描述,提出了本体映射的改进方法,设计了石油开发领域本体集成平台,实现石油开发领域知识的表达和共享,对石油开发深层次地综合运用和评估具有重要意义。

关键词 本体 石油开发领域 形式化描述 本体映射

中图法分类号 TP311.12; **文献标志码** A

随着石油开发的深入和科学技术的不断发展,互联网的广泛普及,各种相关信息量极大的丰富,交叉学科知识的增长与更新速度日益加快,为石油开发领域知识构建带来复杂的问题。石油开发领域包含的多个专业普遍存在着系统独立开发、单纯追求功能实现,没有从整个石油行业的高度来规划各种石油开发应用系统的设计和开发,使得从调研确定需求阶段开始到组织管理数据等多方面造成了对各专业知识的不一理解和使用,导致了各系统之间对信息、知识共享的障碍,无法为石油开发领域提供完善的全局解决方案。

石油开发包含石油开发的资源预测、剩余油挖掘、地面系统设计与改造、提高采收率工程等多方面的应用,而且和石油地质有着千丝万缕的联系,地质学中包含多门知识,如岩石、矿物、构造、矿床、地史古生物、地球化学、找矿勘探、物探、化探、遥感等,这些专业的信息管理仍然存在很多问题:

(1) 过去上述专业中相关数据管理和应用是孤立和平行分离的,虽然已经建立和积累了海量石油

开发结构与资源数据,但通常是简单地生成各不相同的数据表,很少做综合应用高层次的再开发,因此这些基础数据库将会成为一系列脆弱的新型“知识孤岛”;

(2) 它们主要以单独数据表为核心,缺乏在石油开发行业知识基础上统一的概念模型、数据模型、元数据标准、相互融合的接口,其数据库的整合性差、利用率低;

(3) 多数石油开发过程中所测绘原始数据的产生、整理和归档是按项目要求进行的,文件方式存储、单机环境进行制图或编写报告,以致这些石油开发信息都以文字、图表形式散落在工程勘察报告中而未被充分利用,难以实现信息复用和综合分析应用,甚至是有些数据库的内容及相应的特征缺失。

按照这些情况建立起来的油田数据库只是数据而没有其内在的规则,无法体现出各个专业词汇之间复杂的联系,使得难以用简单的类别属性来表示各专业词汇在特征和功能上的关系。为解决对专业词汇进行组织和描述的问题,需要引入新的理念,确定合理的数据描述模型以便对知识库中的信息进行适当的表达,从而使计算机真正发挥其为决策人员提供参考的作用。因此从长远看石油开发领域全局知识的构建已成为一种趋势,即从知识的角度对石油开发资源进行管理。

1 本体概念

本体的概念最初起源于哲学领域,可以追溯到公元前古希腊哲学家亚里士多德。它在哲学中的定义为“对世界上客观存在物的系统地描述,即存在论”,是对客观存在的一个系统的解释或说明,关心的是客观现实的抽象本质^[1]。近年来,本体这个概念在各个领域的应用研究得到了较快的发展。

人工智能领域,最早给出本体定义的是 Neches 等人,他们将本体定义为“给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”^[1]。

引用得最为广泛的定义是由 Gruber 提出的:“本体是概念化的明确的规范说明”^[2,3]。W. N. Borst 对该定义进行了引申,认为“本体是共享的概念模型的形式化的规范说明”^[4]。Fensel 又对这个定义进行分析,认为本体的概念包括如下四个主要方面^[5]:

(1) 概念化:通过抽象出客观世界中一些现象的相关概念而得到的模型,其表示的含义独立于具体的环境状态。

(2) 明确:所使用的概念及使用这些概念的约束都有明确的定义。

(3) 形式化:Ontology 是精确的数学描述,而且是计算机可读。

(4) 共享:Ontology 中体现的是共同认可的知识,反映的是相关领域中公认的概念集,它所针对的是团体而不是个体。

本体的目标是研究相关的领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇和词汇之间相互关系的明确定义。本体通过对概念的严格定义和概念之间的关系来确定概念精确含义,表示共同认可的、可共享的知识。

由于本体具有概念化,明确化,形式化和共享的特点,并且具有良好的概念层次结构和对逻辑推理的支持,因而将其应用于研究必将能够有效地解

决开发领域知识描述中的语义不确定性和传统模式在知识构建上应用的局限性等问题。

本体(Ontology)及相关技术的引入为石油开发领域知识的全局构建问题的解决提供了全新的视角。从地理信息系统到可缩放向量图形(Scalable Vector Graphics, SVG),再到本体和语义网,人们由概念和逻辑演绎现实世界的一切,又从现实世界中提炼、抽象概念,使之不断进化完美,并致力于用一个完美的框架,来表达、演绎人类所有知识,并以此发现新的知识。从本体到语义网络,人们终于还是在数字世界的演化中,不知不觉地融入了哲学的浩瀚。石油开发知识的构建是数字油田信息化的重要组成部分,石油开发领域的交叉性和综合性决定了它与石油行业内其他领域信息交换和资源共享的频繁性和迫切性。随着石油开发知识涵盖越来越广,新的技术手段不断产生,新的术语及相关概念不断更新,面对浩瀚的数据,传统的领域知识构建和更新知识库的方法比较低效,且覆盖率和实时性都十分有限,难以满足迅速增长的石油开发决策的应用需求。因此快速构建并实时更新石油开发领域知识库的需求更加迫切。

2 本体的形式化描述定义

从目前比较有代表性的结构来看,本体主要有金芝^[6]、崔巍^[7]提出的三元组本体模型;王洪伟等^[8]、Tomai 和 Kavouras^[9]提出的四元组本体模型;Perez 和 Benjamins^[10]、Rodriguez^[11]提出的五元组本体模型和 Myo-Myo Naing^[12]提出的六元组本体模型。本文在六元组的研究基础上,结合石油开发建模的研究背景,建立了石油开发领域本体的形式化定义,各元组具体含义如下:

$$O = \langle C, A^C, R, A^R, H, X \rangle。$$

(1) C 表示概念的集合,每一个概念 C_i 表示同一类型的对象。

(2) A^C 表示概念属性集的集合,每一个属性集都对应一个概念,同一个概念 C_i 可以用同一个属性集中的属性 $A^C(C_i)$ 来表示。

(3) R 表示关系的集合。领域中概念之间的交互作用形式上定义为 n 维笛卡儿积的子集, $R: C_1 \times C_2 \times \cdots \times C_n$ 。在语义上, 关系对应于对象元组的集合, 其中每一个关系 $R_i(C_1, C_2)$ 表示概念 C_1, C_2 之间的二元关系, 其关系的实例就是概念对象的元组 (C_1, C_2) 。

(4) A^R 表示关系属性集的集合, 每一个属性集都对应一个关系。

(5) H 表示概念之间的层次关系, 是 $C \times C$ 的一个子集, 表示概念之间的父子关系, $H(C_1, C_2)$ 表示 C_1 是 C_2 的子概念。

(6) X 表示公理集。公理表示永真断言, 每一个公理都是对概念和关系的属性值进行约束。可使用适当的逻辑语言, 如一阶逻辑来表示。

3 本体映射库的构建

本体映射是指在两个本体的概念对具有基本相同的语义。进行石油开发多学科协作, 需要在集成异构系统的局部本体同领域本体之间构建本体映射。本体映射通常采用某种本体相似度算法, 选取具有最大相似度值的概念构成本体映射。从而形成本体映射库。

本文针对现有相似度计算方法的不足, 进行了一些改进, 主要针对异构本体间的概念映射综合考虑概念名称、属性、实例、层次结构以及概念间关系的信息, 利用同义词集计算概念名称和属性名称的相似度, 通过对概念和属性名称的语义扩展, 提高相似度计算的准确性。

3.1 概念名称的相似度计算

概念中包含多个义原, 每个义原的作用是不一样的, 考虑其不同的影响, 设置不同的权重因子, 计算概念间的语言级相似度可以如下计算:

$$Lsim(c_1, c_2) = \sum_{i=1}^n w_i sim(p_{1i}, p_{2i}) \quad (1)$$

概念间的语言级的相似度计算只是涉及到表示概念名称的标识符, 如果异构本体中, 命名规则不一致, 很有可能同一语义的概念的在不同本体中的名称标识符完全不同, 此时, 各标识符按照公式

(1) 计算的相似度可能为 0。因此, 本文提出在进行概念的语言级相似度计算之前先按照本体定义中的概念的同义词集 (Same As 关系) 及 HowNet 中的中英文注释进行语义扩展, 将计算概念 c_1 和概念 c_2 的相似度转换为计算概念 c_1 和概念 c_2 的同义词集的相似性, 具体算法为对 c_1 的同义词集中的每个元素 $c_{1i} (i=1, 2, \cdots, n)$ 和 c_2 的同义词中的每个元素 $c_{2j} (j=1, 2, \cdots, m)$ 计算相似度, 然后取最大值作为 c_1 和 c_2 的语义相似度, 即

$$Langsim(c_1, c_2) = \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} (Lsim(c_{1i}, c_{2j})) \quad (2)$$

3.2 概念属性的相似度计算

基于属性的相似度计算的基本依据是如果两个概念具有完全相同的属性则认为两个概念可能是相同的, 根据属性计算相似度的算法可以归结为计算两个概念的属性集合的相似程度, 应该不仅和两个概念具有的相似属性有关, 而且应该和属性对概念的影响程度及属性集合中的元素个数有关。

比较两个概念的属性集合的相似度可以使用 Tversky 计算:

$$ppsim(c_1, c_2) = \frac{|A \cap B|}{|A \cap B| + \alpha |A - B| + (1 - \alpha) |B - A|} \quad (3)$$

式(3)中 A, B 分别表示概念 c_1, c_2 的属性集合, $A-B$ 表示属于 A 但不属于 B 的术语集, $B-A$ 表示属于 B 但不属于 A 的术语集, α 根据概念在各自的层次结构中的深度确定。因为同一语义的概念在本体构建中可能按不同的规则命名, 例如, 可能使用汉语、英语或拼音等, 所以单纯从名称上比较属性集合也是不全面的, 这样, 还是从考虑属性名称的同义词集进行算法改进, 将属性集合 A, B 分别按属性的同义词集扩展为 A', B' , 将公式(2)中的 $A-B$ 替换为 $A-B'$, 表示属于 A 但不属于 B' 的术语集, 同理, $B-A$ 替换为 $B-A'$, 即

$$Psim(c_1, c_2) = \frac{|A \cap B|}{|A \cap B| + \alpha |A - B'| + (1 - \alpha) |B - A'|} \quad (4)$$

3.3 概念结构及语义关系的相似度计算

概念间结构的相似度 $Rsim(c_1, c_2)$ 计算主要考

考虑概念的层次关系及非层次关系,层次关系如 Part-of 关系、Kind of 关系等,非层次关系指存在一个映射关系: $F:C_1 \times C_2 \times \cdots C_{n-1} \rightarrow C_n$,即可以由前 $n-1$ 个决定第 n 个。

基本的想法是将两个概念对应的关系集合根据 HowNet 表示成义原的并集,然后根据两个集合的交集的情况计算两概念间的关系层的相似度,集合的比较和义原之间的相似度的计算和前面介绍的算法类似。

3.4 概念实例的相似度计算

在需要映射的两个本体中,可以利用概念的具体实例计算概念相似度。一个概念的实例也是它祖先概念的实例。基于实例计算概念相似度的理论依据是:如果概念所具有的实例全部都相同,那么这两个概念可能是相同的;如果两个概念具有相同实例的比重是相同的,那么这两个概念可能是相似的。用具体实例来计算概念 c_1 和概念 c_2 的相似度,记为

$$Isim(c_1, c_2) = \frac{p(c_1 \cap c_2)}{p(c_1 \cup c_2)} = \frac{p(c_1, c_2)}{p(c_1, c_2) + p(\bar{c}_1, c_2) + p(c_1, \bar{c}_2)} \quad (5)$$

$Isim(c_1, c_2) \in [0, 1]$, $Isim(c_1, c_2) = 0$ 表示两个概念完全无关, $Isim(c_1, c_2) = 1$ 表示两个概念完全相同。

基于实例计算概念相似度涉及到 3 个概率: $p(c_1, c_2)$, $p(c_1, \bar{c}_2)$, $p(\bar{c}_1, c_2)$, 其中 $p(c_1, \bar{c}_2)$ 是从一个本体的实例空间中随机选取的一个实例属于 c_1 但不属于 c_2 的概率,也可以理解为所有属于 c_1 但不属于 c_2 的实例在实例空间中所占的比重。因此,在计算 $p(c_1, c_2)$, $p(c_1, \bar{c}_2)$, $p(\bar{c}_1, c_2)$ 时要用到概念 c_1 和概念 c_2 在各自本体中的实例个数。用 U_i 表示

本体 O_i 中的实例集, $N(U_i)$ 表示实例集中的实例个数。用 $N(U_i^{c_1, c_2})$ 表示在 U_i 中既属于 c_1 又属于 c_2 的实例个数。在存在足够样本的情况下,可以取一些样本作为正反实例,通过机器学习的方法训练学习器,然后使用学习器对实例集 U_i 中的实例进行分类,分成属于 c_i 和不属于 c_i 两类。这样通过机器学习的方法就可以获得 $N(U_1^{c_1, c_2})$ 、 $N(U_2^{c_1, c_2})$ 、 $N(U_1)$ 和 $N(U_2)$, 则可用式(6)计算概率。

$$p(c_1, c_2) = \frac{N(U_1^{c_1, c_2}) + N(U_2^{c_1, c_2})}{N(U_1) + N(U_2)} \quad (6)$$

类似计算 $p(c_1, \bar{c}_2)$ 和 $p(\bar{c}_1, c_2)$, 计算式如下:

$$p(c_1, \bar{c}_2) = \frac{N(U_1^{c_1, \bar{c}_2}) + N(U_2^{c_1, \bar{c}_2})}{N(U_1) + N(U_2)} \quad (7)$$

$$p(\bar{c}_1, c_2) = \frac{N(U_1^{\bar{c}_1, c_2}) + N(U_2^{\bar{c}_1, c_2})}{N(U_1) + N(U_2)} \quad (8)$$

把式(6)、式(7)和式(8)代入式(5)即可求得基于实例的概念的相似度。

3.5 概念的综合相似度的计算方法

将概念名称相似度、属性相似度、结构相似度和实例相似度的计算结果进行加权求和,即可求到概念的综合相似度:

$$sim(c_1, c_2) = w_1 Langsim(c_1, c_2) + w_2 Psim(c_1, c_2) + w_3 Rsim(c_1, c_2) + w_4 Isim(c_1, c_2)。$$

其中的 w_i 为权重系数,满足 $w_1 + w_2 + w_3 + w_4 = 1$, 各权重系数需由领域专家根据不同的环境适当选择,依赖于专家的经验。

3.6 映射库的建立流程

通过相似度算法可以得到绝大多数映射数据,以本文本体相似度计算为主和适当程度的用户参与的方法,可以建立起一个可靠而完备的本体映射库。本体映射库的建立过程如图 1 所示。

解和交流成为可能,并保持语义上的一致性从而达到知识的共享,这一统一的框架或规范模型即本体。

它可以有效地进行知识表达、知识查询,或不同领域知识的语义消解。此外,这些共享的知识,其最终目的还是应用于解决问题,即使在没有领域专家的情况下,计算机也能够自动地反复运用相关的领域知识解决各种实际的问题,即通过知识重用提高问题求解的质量和效率。实现这类知识重用的系统即基于知识的系统,它通过知识表示和基于知识的推理技术自动化地运用知识辅助人们进行问题求解。

按照前述改进的混合本体集成方法,开发了一个石油开发领域知识集成平台,结构如图 3 所示。

石油开发领域知识集成平台的结构由七部分组成:本体建立/编辑模块、本体库、相似度计算模块、本体映射库、本体映射库管理模块、石油开发领域知识集成模块、系统接口、以及参与集成的信息系统。石油开发领域知识集成模块是平台的核心模块,该模块能够根据石油开发领域多专业的要求,发布各系统开发数据到集成平台和从集成平台查询、转换、更新,删除石油开发数据。

集成平台以服务器/客户端的方式运行。服务器端可以提供基于本体的语义转换、实例发布、数据查询、数据更新、删除以及本体映射相似度值的维护等服务。

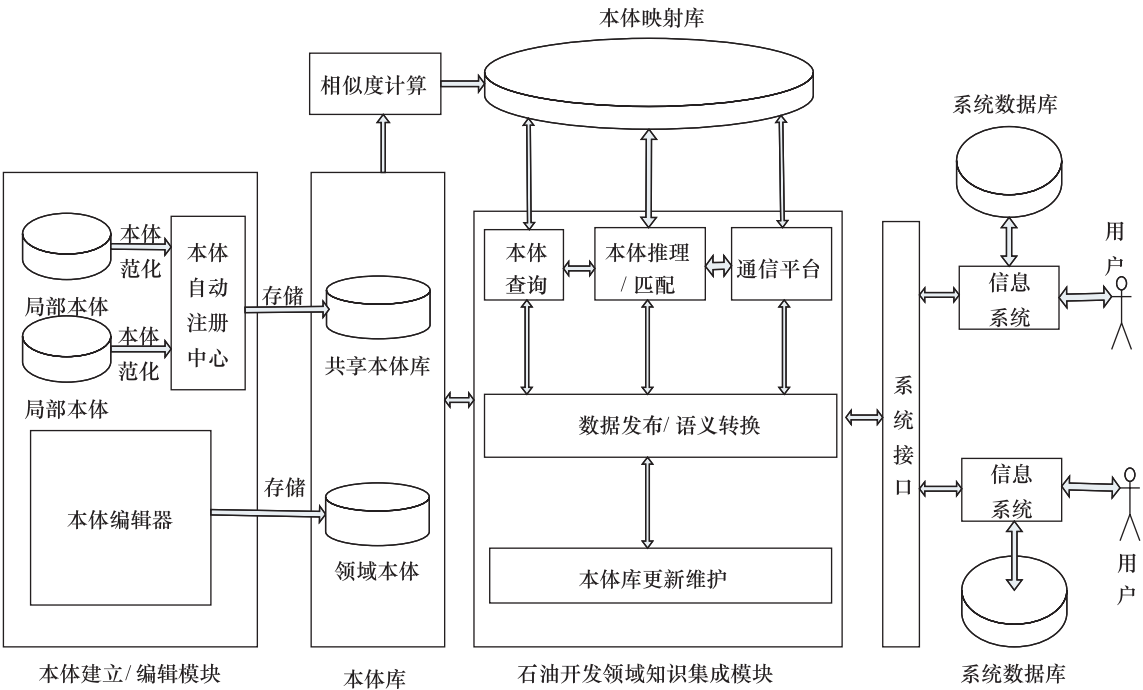


图 3 石油开发领域知识集成平台结构

6 结束语

本文研究的基于本体的石油开发领域知识构建是由石油开发专家系统、知识密集型的信息系统等发展而成,辅助石油开发人员进行问题求解。该系统中的学科知识是用本体描述的;即利用本体对

知识进行表达、组织和再组织,并深入地运用知识进行石油开发问题求解。

本文中的基于本体的石油开发领域知识构建的方法是:首先根据本体的形式语言描述,建立开发领域本体;其次进行相似度计算,建立本体概念之间的映射;通过石油开发领域集成平台为中介,提供各种石油开发领域服务,这些服务具有基于语义的数据查询、数据发布、数据转换等功能,异构系

统通过客户端可以调用集成平台的服务从而实现基于语义的石油开发领域知识构建。

参 考 文 献

- 1 邓志鸿,唐世渭,张 铭,等. Ontology 研究综述. 北京大学学报(自然科学版),2002;38(5):730—738
- 2 Gruber C T R. A translation approach to portable ontologies. Knowledge Acquisition, 1993;5(2): 199—220
- 3 Gruber T R. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human and Computer Studies,1995;43(5/6):907—928
- 4 Borst W N. Construction of engineering ontologies of knowledge sharing and reuse. PhD thesis, Enschede;University of Twente, 1997
- 5 Ontologies D. Silver bullet for knowledge management and electronic commerce. Springer, 2001
- 6 金 芝. 基于本体的需求自动获取. 计算机学报,2000;23(5):486—492
- 7 崔 巍. 用本体实现地理信息系统语义集成和互操作. 武汉:武汉大学博士论文,2004
- 8 王洪伟,吴家春,蒋 馥. 基于描述逻辑的本体模型研究. 系统工程,2003;21(3):101—106
- 9 Tomai E,Kavouras M. From “onio – geoNoesis” to “onto – genesis” the design of geographic ontologies. Geoinformatica, 2004;8(3):285—301
- 10 Perez A, Benjaminx V. Overview of knowledge sharing and reuse components: ontologies and problem-solving methods. In: Proceedings of the IJCAI’99 Workshop on Ontology and Problem-Solving Methods: Lesson learned and Future Trends, Amsterdam: CEUR Publications, 1999;(18): 1—15
- 11 Rodriguez A. Assessing semantic similarity among spatial entity classes. PhD thesis, University of Maine, 2000
- 12 Naing M M, Lim E P, Hoel D G. Ontology-based Web annotation framework for hyperlink structures. Singapore:Proceedings of the International Workshop on Data Semantics in Web Information Systems (DASWIS’02),2002;183—194

Study on Ontology-based Knowledge Construction of Petroleum Exploitation Domain

DU Rui-shan,SHANG Fu-hua,WU Ya-juan

(School of Computer and Information Technology, Daqing Petroleum Institute, Daqing 163318, P. R. China)

[Abstract] Knowledge construction of petroleum exploitation domain is introduced, in order to solve problems of information sharing and application integrating which are caused by not unified of information terms between multi – disciplinary of oil development domain. Through the formal description of ontology, an improved method of ontology mapping is put forward, designs an ontology integration platform of petroleum exploitation domain, and realizes knowledge representing and knowledge sharing of oil development domain. The significance for deeply comprehensive application and evaluation of petroleum exploitation domain is important.

[Key word] ontology petroleum exploitation domain formal description ontology mapping