基于语料库和规则库的石油本体自动构建研究

文必龙,段 炼,汪志群,李云静,王琪超
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:石油领域文本所蕴含的信息丰富但其数目繁多复杂,现有大多数本体都是通过手工构建的,这种方法难以方便快捷地抽取文本信息,难以构建一个较完善的石油领域本体。为提高本体构建的效率,文中综述了本体的主要概念,分析了本体构建的一般原则和方法。利用文本处理软件对文本进行分词处理,生成特征词集并对其进行缩减,利用Petro-Onto方法实现语料库的构建,提出了基于语料库和规则库区分概念、属性并抽取它们之间关系的方法。该方法能大大提高本体的构建效率,并在一定程度上保证结果本体的质量,达到了本体自动建立的目的。

关键词:语料库;规则库;领域本体;本体自动构建
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2015)09-0209-04
doi:10.3969/j.issn.1673-629X.2015.09.044

Research on Automatic Construction of Petroleum Domain Ontology Based on Corpus and Rule Base

WEN Bi-long, DUAN Lian, WANG Zhi-qun, LI Yun-jing, WANG Qi-chao
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Texts of petroleum domain contain rich but numerous and complex information, and most existing ontology are built by manual. But this method is difficult to extract information in text conveniently and rapidly and build a complete text of petroleum domain. In order to improve the efficiency of building ontology, sum up the main concept of ontology and analyze the general principles and methods. Use text processing software to segment words and generate feature word set, then shrinking them. Through building corpus through Petro-Onto, propose a method based on distinguishing concepts and attributes of corpus and rule base and extract the relationships between them. This method can greatly improve the efficiency of building ontology and can guarantee the quality of the result of ontology, and eventually achieve the purpose of building text automatically.

Key words: corpus; rule base; domain ontology; ontology automatic construction

0 引 言

本体是在语义和知识层次上描述信息系统的概念模型,它自从被提出以后就受到了广泛关注,不但被广泛应用在计算机领域,而且近些年来随着油田数字化的采集技术、存储技术的进步,其在石油领域的应用研究也正成为热点。本体应用价值在石油领域凸显的同时,一个非常现实的问题就是:如何方便快捷地对数以万计的石油文本构建领域特定的本体,成为研究的难点。文必龙等[1]提出运用业务模型、数据模型为参照的领域本体构建方法,设计了石油领域本体的顶层本体框架及规范,建立了一个石油领域本体(Petro-Onto)。Petro-Onto可以明确专业术语及其之间的关系、领域公理,并使其形式化、规范化;在人和人之间、人和机器之间共享;达到石油勘探开发领域知识复用的目的。杜睿山等[2]在本体六元组表示的基础上,建立了本体在石油开发领域的形式化定义,提出了本体映射的优化改进方法,设计了石油开发领域本体集成平台。但是他们提出的都是建立石油领域的本体框架、规范,没有更具体的进一步说明,也没有利用语料库和规则库去构建石油本体。

文中对主流的一些本体构建原则、方法进行介绍,

收稿日期:2014-07-28 **修回日期**:2014-11-04 **网络出版时间**:2015-08-26
基金项目:国家科技重大专项(2011ZX05023-005-012)
作者简介:文必龙(1967-),男,博士后,教授,硕士生导师,研究方向为软件工程与集成技术、数据库应用;段 炼(1988-),男,硕士研究生,研究方向为软件设计开发与集成。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1535.006.html

在此基础上,提出一种新的基于语料库和规则库处理技术的本体构建方法。

1 领域本体及构建方法

1.1 本体的定义

本体在汉语中最早出现在公元 300 年左右西晋司马彪所著的《庄子注》中,“性,人之本性也”。在当今汉语中,“本体”指的是事物的主体或自身,事物的来源或根源。在哲学中,本体指的是对客观存在的一个系统解释或说明,所以,它描述的是客观现实的抽象本质[3-4]。在计算机领域,本体由 Gruber 提出,指的是概念化明确的规范说明;在人工智能领域最早定义本体的是 Neches 等,他们将本体描述为“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”。总之,本体的概念最主要包括四个方面[5-6]:

(1)概念化:描述客观世界现象的抽象模型;

(2)形式化:使用数学工具对其进行精确的描述;

(3)明确:精确地定义概念及其联系;

(4)共享:所反映的知识是使用者公认的。

综上所述,本体可以被认为是用来描述概念以及概念之间关系的模型。

1.2 领域本体构建的原则及方法

本体是对相关领域中的概念、概念之间关系的显示描述。目前领域本体的构建方法还在摸索性研究阶段,很多研究人员提出了一些有益于构造本体的标准,以此对构建本体进行指导。通过总结分析,本体的设计原则包括如下几个方面[7-8]:

(1)明确性和客观性:本体的定义应该明确、客观。

(2)完全性:给出的定义能完整表述术语的含义。

(3)一致性:由术语推断出的结论与术语本身含义是基本一致的。

(4)最大单调可扩展性:往本体中添加相关的术语时,其内容不用修改。

(5)最小承诺:建模对象所给的约束应该尽可能少。

(6)兄弟概念语义的差异越小越好。

(7)最小编码偏差:本体的构建和具体的编码语言尽可能相独立。

(8)术语名称使用的越标准越规范越好。

(9)在实现多继承机制时应使用尽可能丰富的概念层次结构。

基于以上原则,目前出现了很多本体的构建工程思想,其中比较有名的有 IDEF-5 方法、Skeletal Methodolody 骨架法、TOVE 企业建模法、Methonotology 方法、循环获取法、七步法等[9]。

文中结合以上本体的建立原则和方法,归纳出对石油领域文本进行本体构建的步骤:首先确定处理的对象是石油领域的文本文件;其次抽取出文本的特征值并对其进行缩减[10];然后根据语料库区分出概念和属性;最后根据语料库和规则库建立起概念和属性之间的关系,并用 OWL 语言对其进行描述。

2 石油领域语料库的构建

根据 Petro-Onto 构建领域本体的方法,对语料库进行构建。语料库中包含石油领域相关的概念及与概念相对应的属性,每个概念有指向其父节点、子节点的指针,大致步骤如下:

(1)确定语料库的需求:石油勘探与开发是油田企业的核心业务,之所以建立语料库就是为了捕获石油勘探开发的领域知识,为自动建立本体服务,因此构建语料库需要满足两方面的需求:信息需求(围绕油田企业信息共享、应用集成的需要)和业务需求(围绕勘探、钻井、采油等核心业务)。

(2)建立语料库的框架:一般来说,可以将语料分为三层:顶级语料层、领域语料层、应用语料层。其中通用的概念用顶级语料表示,比如空间、时间、地点、事件、行为等,概念和特定的领域是相互独立的,可使用在不相同的领域。处于第二层的领域语料指的是特定领域下对领域进行描述的词汇和术语。顶级语料定义的词汇可以被领域语料用来引用,以此来描述领域语料的词汇。具体的应用使用处于第三层的应用语料来描述,它不但可以引用特定的领域语料,而且也可以用来引用任务中的概念。

在顶级语料库中包括 4 个子概念:活动(勘探、钻井、测井、录井、采油等都是它的子类)、对象(井、区块、层位、组织机构、油田各种资源等都是它的子类)、特性(孔隙度、饱和度、渗透率、产量等都是它的子类)、元(描述标准化的个体概念;描述概念之间的关联;描述别名信息)。

(3)确定语料库的参照体系:石油领域经过长期的信息化建设,已经建立了大量的业务模型、数据模型、应用软件,好多主要的领域知识都蕴涵其中,这些信息系统可以作为参照源模型,通过分析和处理,构建语料库。

(4)对语料库进行规范、标准化定义:从参照体系中获得的领域知识是一个没有经过细化的比较粗糙的语料库模型,因此还需要进一步对其进行更加精确的描述和规范化的定义,通过对获得的概念进行一致性分析、冗余分析、完整性分析、抽象、分类,最后形成语料库。

3　语料库与规则库相结合的自动本体构建方法

在对石油领域文本构建本体之前,需要使用文本处理工具对文本进行预处理,抽取出文本的特征值,并使用评估函数对特征值进行选择,形成缩减后的特征值集合[11]。预处理后文本特征值集合的每个特征都是孤立的,而孤立的特征很难对研究者有所帮助,所以要区分出概念和属性,建立它们之间的关系。下面针对这两个问题进行讨论。

使用缩减后的文本特征分别与语料库中的概念语料和属性语料进行扫描匹配,查找出概念和属性,如果未匹配成功,则手工区分概念和属性,然后再把这些概念和属性添加到语料库中。

概念和概念之间的关系主要包括部分与整体(part-of)、包含(kind-of)、概念与属性(attribute-of)这三种关系[12],下面分别进行讨论:

(1)部分与整体(part-of)和包含(kind-of)。

在图 1 中,概念 A 首先在语料库中查找父概念、同级概念和子概念,记为概念集 C,使用特征值和 C 进行匹配,若匹配成功,则概念 A 和概念 B 关系建立成功;若未找到,从规则库[13]里查找和概念 A 模式相匹配的概念。这种规则库匹配方法是一个循环过程,开始可以人工总结出一些模式,如 N_0<是> N_1[<是> N_2, <是> N_3,…,<是> N_i]或者 N_0<组成> N_1[<组成> N_2, <组成> N_3,…,<组成> N_i]。利用这些模式得到一些概念间的关系,然后从中取出一对有关系的概念对文本进行搜索,又可以发现此关系新的模式,再利用这些新的模式又可以从特征值中建立新的概念关系。这种句法匹配模式方法对新模式的发现和定义是一个循环的过程,这种方法对于人工发现和定义模式有了一定的改进。

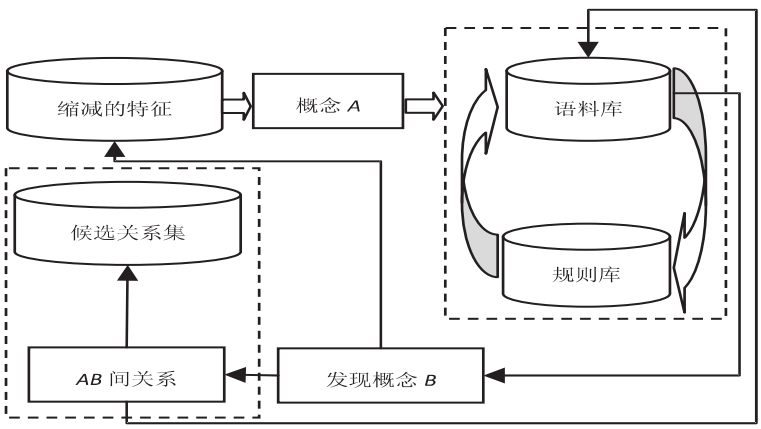

图 1　关系的抽取

(2)概念与属性(attribute-of)。

把概念和语料库中概念相匹配,匹配完成后从语料库中查找该概念的属性,并把查找出来的属性和特征词里面的属性相匹配。若匹配完成,则概念和属性之间的关系建立成功,反之通过手工的方法建立关系,然后把建立起的关系对语料库进行完善。

通过以上讨论,已经建立起了概念与概念、概念与属性之间的关系,使用编程技术把上述生成的概念之间的关系读取出来,然后使用 OWL 语言对抽取的概念和建立的关系进行描述,生成本体库[14-15]。

4　实　验

4.1　实验方法

1)石油领域文本的选择。

分别在石油的勘探、钻井、测井、录井、采油领域选择若干个文本,按照以上步骤进行本体构建工作。

2)文本预处理。

(1)根据第 3 节提到的方法对文档进行预处理,获得缩减后的特征值集合 S。

(2)使用第 2 节构建的语料库和集合 S 中的元素进行扫描匹配,找出概念 S_0、属性 S_1。

(3)使用第 3 节的方法建立 S_0 和 S_1 之间的关系,使用 OWL 语言对关系进行描述,生成本体文件。

4.2　结果及分析

随机选取 10 篇油田勘探、钻井、测井、录井、采油领域方面的文档,分别使用以上方法和传统的手工方法对文档进行本体构建实验。实验结果如表 1 所示。

表 1　实验结果

实验方法	D	F	R	P_F	P_R	T
手工方法	10	128	236	92	189	243
文中方法	10	102	206			67

其中,D 为实验样本数;F 为通过实验方法提取到的缩减后的特征值;R 为通过实验方法提取到的概念和属性之间的关系结果数;P_F 为 $F_{手工方法} \cap F_{文中方法}$ 的元素个数;P_R 为 $P_{手工方法} \cap P_{文中方法}$ 的元素个数;T 为耗用的时间数(单位:min)。

通过实验结果可知,使用文中提出的方法虽然达不到手工方法的准确率,但其节省人力物力,节省时间,并能大致地反映文本内容。

5 结束语

所提出的基于语料库和规则库的石油领域本体自动构建的方法,利用预处理软件对文本特征进行抽取,利用语料库区分概念和属性,利用规则库建立概念属性之间的关系,较传统的手工方法有明显的优势。

当然,文中的研究内容还存在需要提高的方面。比如,语料库中存放的语料不全面,规则库中存放的规则不完全,这些因素影响概念、属性抽取的准确率,直接导致本体自动构建时间的增加。这些问题,都将作为今后的研究内容。

参考文献:

[1] 文必龙,张 莉. 石油勘探开发领域本体的构建方法研究[J]. 计算机工程与应用,2009,45(34):1-3.

[2] 杜睿山,尚福华,吴雅娟. 基于本体的石油开发领域知识构建研究[J]. 科学技术与工程,2010,10(19):4656-4662.

[3] 朱恒民,姬小利,黄卫东,等. 电信领域本体构建方法研究[J]. 现代情报,2008,28(1):184-186.

[4] 王 进. 基于本体的语义信息检索研究[D]. 合肥:中国科学技术大学,2006.

[5] Zhong J, Aydina A, McGuinness D L. Ontology of fractures [J]. Journal of Structural Geology,2009,31(3):251-259.

[6] Raskin R. Guide to SWEET ontologies[R]. Pasadena:NASA/Jet Propulsion Lab,2013.

[7] 徐力斌,刘宗田,周 文,等. 基于 WordNet 和自然语言处理技术的半自动领域本体构建[J]. 计算机科学,2007,34(6):219-222.

[8] López M F, Gómez-Pérez A, Sierra J P, et al. Building a chemical ontology using methontology and the ontology design environment[J]. IEEE Intelligent Systems & Their Applications,1999,14(1):37-46.

[9] Uschold M, King M. Towards a methodology for building ontology[C]//Proc of international joint conference on artificial intelligence. [s. l.]:[s. n.],1995:373-380.

[10] 王晓盈,王晓璇,刘 鹏. 中文本体构建及可视化研究[J]. 计算机技术与发展,2010,20(2):121-124.

[11] 陈晓云. 文本挖掘若干关键技术研究[D]. 上海:复旦大学,2005.

[12] 徐 健,张智雄,吴振新. 实体关系抽取的技术方法综述[J]. 现代图书情报技术,2008(8):18-23.

[13] 刘 威. 基于中文文本的本体构建方法研究[D]. 哈尔滨:哈尔滨工程大学,2008.

[14] 杜小勇,李 曼,王 珊. 本体学习研究综述[J]. 软件学报,2006,17(9):1837-1847.

[15] 鲍 文,李冠宇. 本体存储技术研究[J]. 计算机技术与发展,2008,18(1):146-150.

(上接第 208 页)

[3] 郄秀书,吕达仁,卞建春,等. 中高层大气瞬态发光事件(TLEs)及可能的影响[J]. 地球科学进展,2009,24(3):286-296.

[4] Funaki K, Fukunishi H, Tsuji Y, et al. Giant cystic leiomyoma of the uterus occupying the retroperitoneal space[J]. Journal of Radiology Case Reports,2013,7(12):35-40.

[5] Pasko V P, Yair Y, Kuo C L. Lightning related transient luminous events at high altitude in the earth's atmosphere: phenomenology, mechanisms and effects[J]. Space Science Reviews,2012,168(1-4):475-516.

[6] Siefring C L, Morrill J S, Sentman D D, et al. Simultaneous near-infrared and visible observations of sprites and acoustic-gravity waves during the EXI98 campaign[J]. Journal of Geophysical Research,2010,115:A00E57.

[7] 黄文耿,古士芬. 雷暴云准静电场和夜间低电离层的电离[J]. 空间科学学报,2002,22(3):227-233.

[8] 黄文耿,古士芬. 雷暴云准静电场对夜间电离层 D 区的影响[J]. 地球物理学报,2003,46(2):162-166.

[9] 吴明亮,徐寄遥. 时变的准三维"红闪"电场模式研究[J]. 地球物理学报,2005,48(3):480-486.

[10] Yang J, Qie X S, Feng G L. Characteristics of one sprite-producing summer thunderstorm [J]. Atmospheric Research, 2013,127:90-115.

[11] Rockwood S D, Greene A E. Numerical solutions of the Boltzmann transport equation[J]. Computer Physics Communications,1980,19:377-393.

[12] Elliott C J, Greene A E. Electron energy distributions in e-beam generated Xe and Ar plasmas[J]. Journal of Applied Physics,1976,47(7):2946-2953.

[13] Morgan W L, Penetrante B M. ELENDIF: a time-dependent Boltzmann solver for partially ionized plasmas[J]. Computer Physics Communications,1990,58,127-152.

[14] Razier Y P. Gas discharge physics[M]. New York:Springer-Verlag,1991.

[15] 郭冠军,李树楷. 对流层内激光垂直精密测距研究[J]. 光电子·激光,2001,12(4):400-402.

[16] 肖存英. 临近空间大气动力学特性研究[D]. 北京:中国科学院研究生院(空间科学与应用研究中心),2009.

基于语料库和规则库的石油本体自动构建研究

作者： 文必龙， 段炼， 汪志群， 李云静， 王琪超， WEN Bi-long， DUAN Lian， WANG Zhi-qun， LI Yun-jing， WANG Qi-chao

作者单位： 东北石油大学 计算机与信息技术学院,黑龙江 大庆,163318

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2015(9)