基于 Neo4j 对涉藏领域本体的存储方法研究

王　飞[1]，　易绵竹[1]，　谭　新[2]，　陈永升[1]，　向一帆[1]

(1. 信息工程大学洛阳校区 语言工程系　河南 洛阳 471003；2. 91709 部队　吉林 珲春 133300)

摘要：基于本体语义理论构建的涉藏领域本体定义了一种多属性值对的语义表示框架，建立了语义关联数据模型，将文本中的对象和事件通过语义属性联系起来，可以为多种应用提供支持.该语义表示框架可以表示成多维度的资源描述框架(resource description framework，RDF)，采用传统的关系型数据库存储会带来空间浪费和数据管理困难等问题.在大数据存储技术的推动下，产生了非关系型数据库用以存储复杂关系数据.主要研究了将涉藏领域本体数据按照节点和边的形式存储在图数据库 Neo4j 中的方法，使数据库能够更有效地存储语义数据，并提供可视化的查询与处理，比传统的关系型数据库提高了操作效率.

关键词：本体；存储；知识表示；Neo4j；RDF

中图分类号：TP391.9　　**文献标志码**：A　　**文章编号**：1671-6841(2019)02-0060-06

DOI：10.13705/j.issn.1671-6841.2017259

0　引言

随着互联网的发展，一维关系已不足以描述真实数据.本体将领域中的对象和事件通过语义建立起关联关系，用以描述领域知识，体现了本体语义的优势[1].语义数据可表示为 RDF 格式，描述一个节点具有一组属性，或将属性看作节点间的关系[2].数据存储形式影响了知识的应用效率，过去采用关系型数据存储本体知识，随着数据量增加和多关系数据的存储需求，产生了非关系型数据库(NoSQL)[3].关系型数据库将数据存储成一维关系数据，每一维包含一个主键，而非关系型数据库的每个节点和关系均可包含多个属性，更适合存储和处理复杂的多关系型数据.

本文构建的涉藏领域本体以国家安全为目标，以事件为中心将对象建立语义关联，形成多关系数据.涉藏领域对象包含人名、地理实体名和组织机构名.从语言学角度分析，事件通过动词表示，对象通过名词表示，对象作为语义角色参与事件中，形成多属性值对的语义描述框架.关系型数据库在存储和查询含有多属性值对的数据时存在如下问题：查询效率低下；空间浪费；可视化效果不好；不易发现所有关系.涉藏领域本体知识类型较多，且存在属性差异，导致信息存储空间不同，非关系型数据库更适合存储和表示复杂的数据结构，并形成知识图谱，便于处理和挖掘涉藏领域实体的关联关系，帮助发现隐含知识，辅助决策.

本体作为一种形式化知识表示方式，概念之间需要特定的语言对其进行约束和推理[4].本体知识通常采用 OWL(web ontology language)或 RDF 本体描述语言，存储在关系型数据库中[5].RDF 以“实体-属性-值”三元组描述实体与实体之间的关系，对知识共享和知识交换提供支持[6].随着互联网技术和语义 web 的发展，基于 web 的本体描述语言占据了主要地位[7].目前西藏主题的本体主要面向通用领域，以处理藏文信息.Jiang[8]提出基于 HowNet 概念相似度和藏英词典获取藏语本体.Qiu[9]在藏语本体中基于构式获取藏语词汇的概念.Xu[10]提出基于本体计算藏语概念相似度，用于推进藏语信息处理.这类本体不具有国家安全领域关注的对象，并且在知识表示研究的推动下，构建方法也需要进一步更新.本体编辑工具 Protégé[11]因其开源、方便、模块清晰等优点，在本体构建领域受到广泛应用[12].虽然 Protégé 提供了友好界面和一致性检验，但它仍需要人工输入和编辑大量的数据信息，很难构建大规模的本体工程[13]，存在很多限制.Protégé 产生的文件

收稿日期：2017-09-04

基金项目：国家自然科学基金项目(11590771).

作者简介：王飞(1983—)，男，新疆伊犁人，博士研究生，主要从事自然语言处理、数据挖掘研究，E-mail：89738764@qq.com；易绵竹(1964—)，男，四川南充人，教授，主要从事自然语言处理研究，E-mail：mianzhuyi@gmail.com.

格式为 RDF 或 OWL,表现力还有欠缺[14].NoSQL 数据库可以表示多种关系和属性,并支持知识动态存储和演化,适合大数据实时处理[15].Neo4j 是一种高性能的可视化效果较好的 NoSQL 数据库.

为了将涉藏领域本体中不同类型的知识关联,符合本体语义分布特性,本文提出采用 Protégé 构建领域本体、Neo4j 图数据库进行知识存储的方法,能更加直观表现事件与对象的所有关系,同时可以提高存储效率,减小存储空间,为用户提供更好的可视化查询效果.在涉及国家安全的涉藏领域中可以描绘出各个实体的语义关系网络,便于发现和预测知识.

1　基于 Neo4j 的涉藏领域本体存储模型

基于本体语义构建的涉藏领域本体是事件驱动的,分析文本中动词和专名,就能得到领域中主要概念.事件概念中,语义角色构成其属性,属性值具有语义选择限制;对象概念属性为数据属性,描述对象基本特征;动词的属性是其在句子中的用法,也就是动词的句法结构;专名和实例的属性按照概念定义的语义框架,根据文本内容进行实例化.涉藏领域本体中不同类型知识通过关系相互映射,概念之间具有“is-a”关系,概念和实例之间具有“instance-of”关系,概念与词汇之间通过语种标签指示,共同形成涉藏领域本体知识库.

1.1　基于本体语义的涉藏领域本体表示

本体语义理论将表示文本意义的概念和词汇分开,词汇和短语的意义通过词汇所映射的本体概念表达.基于本体定义的语义框架对文本进行分析,生成文本意义表征(text meaning representation, TMR)[16].TMR 基于输入文本的语义依存关系构建,以动词为中心,文本中的句法模式作为属性被记录在词汇框架中,文本中的语义角色被填充在 TMR 中,表达文本含义.本体定义的语义框架区分粒度较细,包含中心、属性、侧面和值,表示为(head(slot(facet(filler)+)+)+),以相关领域中的句子(1)“Indian PM Narendra Modi has visited Mr. Obama at the White House.”为例,产生的 TMR 为

(1) VISIT-1

agent	value	Narendra Modi
theme	value	Mr. Obama
location	value	the White House
root	value	visit

VISIT 是动词 visit 所映射的概念,它表示从属于社会事件的访问事件,root 属性指出文本中表示访问意义的词汇是动词 visit,句子(1)所表示的访问事件被记为索引 VISIT-1.概念 VISIT 的语义框架表示为

(2) VISIT

agent	value	PERSON
theme	sem	PERSON
	relaxable-to	LOCATION
location	value	LOCATION
time	value	TIME
en-lex	value	visit

本体概念 VISIT 中的第 1 列表示语义属性,第 2 列是属性的不同侧面,sem 表示基本语义限制,relaxable-to 表示语义限制的放宽,第 3 列是属性值的限定条件,在概念和词汇中由概念或变量填充,在事件和对象实例中由文本的具体值填充,否则为空.en-lex 表示动词 visit 是概念 VISIT 的英语表示,描述为

(3) visit

subj	value	$ var1
dobj	sem	$ var2
comp	value	$ var3
map-concept	value	VISIT

词汇 visit 记录了它的基本句法模式,map-concept 表示词汇 visit 映射到概念 VISIT,由 VISIT 的语义描述框架表示其含义.句中的 Obama 是专有名词,它作为对象实例表示为

(4) Barack Obama

INSTANCE- OF	PERSON
ALIAS	Mr. Obama,
	President Obama,
	the president of the United States,
	the US president
SOCIAL-ROLE	President
GENDER	male
NATIONALITY	USA
BORN	August 4, 1961
SPOUSE	Michelle Robinson(m. 1992)

按照本体语义所定义的框架,领域内的事件和对象都通过关系和属性相互联系起来,建立了语义关联数据,如图1所示.

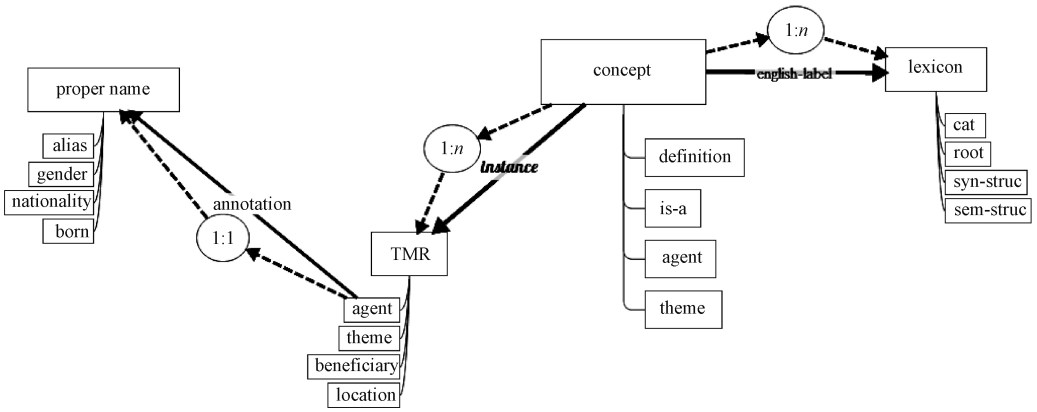

图1 本体语义关联数据

Fig.1 Ontology semantic linked data

1.2 涉藏领域本体知识的构建与RDF表示

本体构建工具使开发者可以只关注本体的结构组织,而不必关心细节,提高了本体构建的效率.本体编辑器Protégé的基本操作都使用图形用户接口,操作时编辑器会检测内部逻辑一致性.Protégé编辑的本体存储为RDF/XML数据格式.XML是可扩展的,允许用户自定义模式或标签,RDF基于XML语法,但是与规范XML模式不同,RDF模式对数据的描述更丰富[17].RDF用三元组来描述语义关系,提供了一个设计好的知识表示环境.每个RDF三元组包括两个实体和一个关系,这类似于句子的主语、谓语和宾语.RDF初始化了本体描述,进行语义声明,开发了复杂的形式表述,使得文本和数据信息能够适应知识表示环境,而这些对使用者来说都是透明的,使用者只需要将知识保存为相应的数据格式.

本体语义的知识框架可以看作是复合型三元组,每个三元组可以看作一维的属性值对,每个概念至少具有一个属性,适合采用RDF来描述,容易被计算机处理.基于本体语义理论构建的涉藏领域本体中每个知识条目都可以表示成多维RDF数据,这好比将一个复杂的高维结构在形式上降维处理,使得后期的存储应用将会更加高效.因此对于涉藏领域语义数据的存储就是对多维RDF数据的存储.例如对(1)事件实例的RDF表示为

<VISIT, rdf: type, concept>

<VISIT, agent, Narendra Modi>

<VISIT, theme, Mr. Obama>

<VISIT, location, the White House>

<VISIT, root, visit>

专有名词(4)的RDF可以简化表示为

<Barack_Obama, rdf: type, PERSON>

<Barack_Obama, ALIAS, Mr. Obama>

<Barack_Obama, SOCIAL-ROLE, President>

<Barack_Obama, NATIONALITY, USA>

用 Protégé 构建初始本体并保存成 RDF 格式数据可以减少一定的工作量,它能够建立起本体概念的知识表示,以及概念与词汇和实例的映射关系,词汇的属性需要单独记录为 XML 格式文件.多维 RDF 数据本身构成了语义网络,可以在格式转换后用图数据库存储和表示.

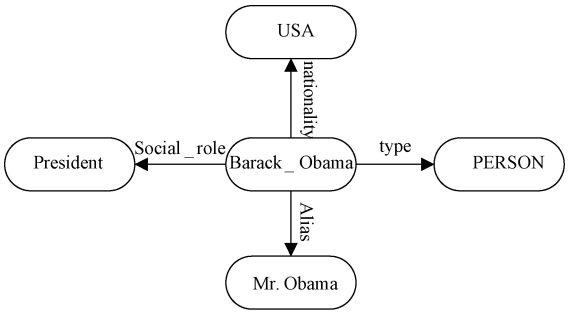

图 2　RDF 数据结构

Fig.2　RDF data structure

1.3　基于 Neo4j 的涉藏领域本体存储模型设计

Neo4j 图数据库模型的基本组成是节点(nodes)、关系(relationships)和属性(properties),它们都是独立存储的.节点和关系都可以创建任意多的属性,通过 key-value 对表示,类似于一个 hashMap 数据结构.Neo4j 能够有效解决多维 RDF 数据中属性个数多少不一带来的内存浪费问题,通过深度遍历接口在多数据连接查询时保持较快的查询速度,与存储的数据量无关,在大规模数据集中体现出了良好的性能.

由于 Neo4j 存储的知识表示方式与本体语义的知识表示框架有一定的区别,并非简单地将基于 Protégé 构建的涉藏领域本体转储进 Neo4j 中,为了达到较好的可视化存储和查询效果,需要重新设计两种知识表示方法的转换.按照图 1 的示例,本文设计基于涉藏领域本体的 RDF 数据与 Neo4j 模型中的元素对应关系如下:

1) 节点.(a) Neo4j 模型中的节点表示为领域本体中不同类型的知识,如概念、TMR 和专名的名称,也就是将涉藏领域事件和对象的概念名和实例名作为实体节点.由于动词词汇仅包含抽象的句法属性,在 Neo4j 中存储时会产生大量相同的存储结构,造成空间冗余,因此将动词词汇作为事件实例的一个词根属性表示,不再将其具体的属性存储在 Neo4j 中.(b) 除事件类型的节点外,其他节点按照本体语义定义的语义框架创建多个属性-值对.

2) 关系. (a) 模型中的关系连接领域本体中不同类型的知识节点,如概念之间的“is-a”关系,概念与 TMR 或专名之间的“instance-of”关系,词汇与概念之间的“map-concept”映射关系,TMR 与词汇之间的“root”关系,即通常所说的对象属性.(b) 涉藏领域本体的语义框架在存储到 Neo4j 模型中时,一个重要的变化就是需要将事件属性(agent、theme、location 和 time)转换为 Neo4j 模型中的关系.做这个转换的原因有两点:事件关联了很多对象,如果某一个对象作为属性值出现,就不再具有实体节点的特点,Neo4j 模型中关系和属性是有区别的,不便于展示一个对象同时关联到多事件的情况;在 Neo4j 模型中,一个事件或者对象作为实体节点是可以共享关系的,但是不能共享属性.如果将对象作为事件属性进行查询,则只能遍历所有的数据,分别查到包含该属性的所有事件,而将事件属性作为关系则可以多关系共享同一个对象节点,查询效率高,且符合查询需要,是多关系数据价值的体现.

3) 属性. 经过如上的调整,则属性就仅作为对象概念和实例性质的基本描述,即所谓的数据属性,例如 PERSON 对象概念的属性主要包括 ALIAS、SOCIAL-ROLE、GENDER 和 NATIONALITY 等.

4) 索引. 涉藏领域本体数据本身是分布式的,存储在 Neo4j 模型中也是分布式的,一个概念节点可能对应着多个词汇节点或者实例节点,也就存在着多个关系.Neo4j 模型对节点、关系和属性都是分别存储的,因此以概念为中心,将概念所映射的同类型知识单独索引,例如概念“INFORM”所映射的英语动词“say”、“tell”和“urge”可分别索引为“INFORM-1”、“INFORM-2”和“INFORM-3”.增加索引的好处在于封装,使得对数据库操作变得简化.

2　实验

2.1　实验数据规模及存储效果

为了验证存储效果,本文从公开网站按照领域专家给出的相关主题词,基于网页标签信息采集西藏相关英语新闻,数据采集的时间跨度从 2010 年 1 月至 2017 年 4 月,获得了 30 410 个句子,经处理后得到领域相

关的概念和实例,筛选了一些具有典型意义的知识,共建立了771个节点、3 833条属性、1 033个关系.

使用Neo4j提供的接口将Protégé保存的本体数据和扩展的XML属性描述数据导入数据库,分别使用Node、Relationship和Property数据类型建立节点、关系和属性.在数据库中,每一个节点会根据它的类型定义不同的特征,节点关系通过边表示,对象属性则单独展示.如事件meet-with1在Neo4j中可视化为以事件为中心的轮子图,事件中所有参与对象都以语义角色作为关系与meet-with1关联,如图3所示.

分别查看每个节点,又会显示其数据属性.整个库就是大量的事件和对象关联形成的知识图谱,通过全部连接可以发现隐含知识,局部可以发现细节.以对象为中心可以查看关联的事件,如the White House作为两个meet-with事件中的location而将两个事件关联起来,由此可以发现两个事件的隐含关系,继续查看事件可挖掘更多相关属性,如图4所示.

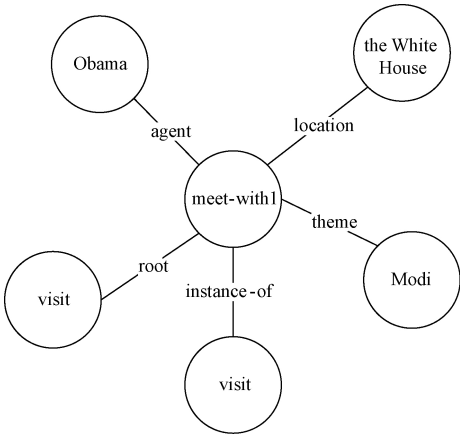

图3 涉藏领域事件存储效果

Fig.3 Event storage effect in Tibet domain

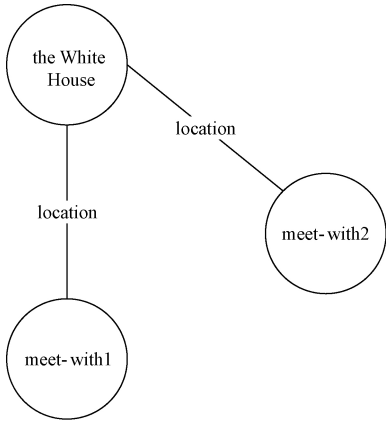

图4 关联同一地点的事件

Fig.4 The events linked with the same location

2.2 存储空间及查询响应时间结果

实验对RDF三元组以OWL文件式存储和在Neo4j中存储所占的存储空间以及查询响应时间进行了对比.在Protégé构建过程中,按阶段分别测试了3组规模不同的RDF文件和Neo4j模型数据,结果如表1所示.

表1 涉藏领域数据存储结果

Tab.1 Data storage result of Tibet domain

三元组	存储空间/KB		平均响应时间/ms	
	RDF	Neo4j	RDF	Neo4j
376	67	58.7	805.3	26
4 936	1 914.88	1 280	1647.54	38
32 618	5 335.04	3 809.28	2 103.7	45

在相同数据量的情况下,Neo4j存储空间更小,通过多次查询同一类型知识系统反馈的响应时间计算出查询的平均响应时间.由于响应时间受到计算机系统性能的影响很大,表1作为一个参考值,数据量较小时对比效果不明显,但相同条件下Neo4j的响应时间更短.

3 结论

本文分析了基于本体语义理论构建的涉藏领域本体的语义框架,发现其数据结构不适合用传统的关系型数据库进行存储,提出了非关系型数据存储的方法.将涉藏领域的语义框架转换成RDF数据格式,按照语义逻辑将不同类型的知识与图数据库Neo4j中的基本存储元素一一对应.该方法的查询效果较好,能够按照查询者的意图只显示相关实例及关系,同时得出以下结论:1)无论知识表示方式如何,对于Neo4j来说,都将分解成节点和边的关系进行存储;2)Neo4j的可视化效果较好,能够显示出不同类型知识节点之间的所有关系和属性;3)Neo4j消除了很多冗余信息,存储效率高,存储空间较小,查询响应速度较快.基于Neo4j图数据库存储涉藏领域本体的方法借助了目前较为流行的NoSQL存储技术,为本体语义的知识存储和表示提供了更适合的解决方案,有利于涉藏领域的知识发现和查询.

参考文献:

[1] NIRENBURG S, RASKIN V. Ontological semantics (language, speech, and communication)[M]. Cambridge: the MIT Press, 2004.

[2] 陈红红, 李辉, 李新春. 基于领域本体的概念格语义匹配[J]. 郑州大学学报(理学版), 2010, 42(2):70-73.

[3] GUPTA A, SCHACHNE A, CONDIT C, et al. GeoSciGraph: an ontological framework for earthcube semantic infrastructure[C]//AGU Fall Meeting Abstracts. San Francisco, 2015.

[4] 陈晓美, 毕强. 面向文本的领域本体学习方法与应用研究综述[J]. 图书情报工作, 2011, 55(23):27-31.

[5] 刘言, 林民. 基于 OWL 的双语领域本体构建方法研究[J]. 计算机技术与发展, 2014(8):84-88.

[6] 李楠. 基于关联数据的知识发现研究[D]. 北京:中国农业科学院, 2012.

[7] NOORDIN M F, SEMBOK T M T, OTHMAN R, et al. Constructing an ontology-based and graph-based knowledge representation of English quran[J]. Jurnal teknologi (sciences and engineering), 2016, 9(34):465-469.

[8] JIANG X, QIU L. A Tibetan ontology concept acquisition method based on HowNet and Chinese-Tibetan dictionary[C]//International Conference on Asian Language Processing. Urumqi, 2013:189-192.

[9] QIU L, WENG Y, ZHAO X. Acquisition method of hyponymy concepts based on patterns in Tibetan semantic ontology[J]. Journal of Chinese information processing, 2011, 25(4):45-49.

[10] XU G X, HE Q, ZHAO X, et al. Tibetan concept similarity computation based on ontology[C]//6th International Conference on Intelligent Networks and Intelligent Systems. Shenyang, 2013:292-295.

[11] 洪娜, 张智雄. Protege 在科研本体构建与推理中的实践研究[J]. 现代图书情报技术, 2009, 25(z1):1-5.

[12] 章勇, 吕俊白. 基于 Protégé 的本体建模研究综述[J]. 福建电脑, 2011, 27(1):43-45.

[13] MARI C S F, ASUNCION G P, ENRICO M, et al. Ontology engineering in a networked world[M]. Berlin: Springer Publishing Company, 2012.

[14] MUSEN M A. The protégé project: a look back and a look forward[J]. AI matters, 2015, 1(4):4.

[15] BANERJEE S, SARKAR A. Ontology driven meta-modeling for NoSQL databases: a conceptual perspective[J]. International journal of software engineering & its applications, 2016, 10(12):41-64.

[16] NIRENBURG S, MCSHANE M, BEALE S. Operative strategies in ontological semantics[C]//Hlt-Naacl 2003 Workshop on Text Meaning, Association for Computational Linguistics. Stroudsburg, 2003:22-29.

[17] 陈哲, 魏衍君. 基于本体的 XML 数据源语义集成研究[J]. 郑州大学学报(理学版), 2006, 38(2):36-39.

Storage Method for Tibet-related Domain Ontology Based on Neo4j

WANG Fei[1], YI Mianzhu[1], TAN Xin[2], CHEN Yongsheng[1], XIANG Yifan[1]

(1. *Department of Language Engineering, Luoyang Campus, PLA Information Engineering University, Luoyang* 471003, *China*; 2. 91709 *Troops, Hunchun* 133300, *China*)

Abstract: Based on the theory of ontology semantics, the Tibet domain ontology defined a semantics representation framework of multi key-value pairs to establish semantics linked data model. The objects and events in the text were linked to support a variety of applications by semantic attributes. The semantic representation framework could be represented as multi-dimensional RDF data, and the traditional SQL database had some problems, such waste of storage space and difficuty in managing data. Big data storage technology produced NoSQL database to store complex relational data. The method storing the Tibet domain ontology was studied by translating entities to the nodes and the edges in the Neo4j graph database. So that the database could store the semantics data more effectively and provided visual query and processing, it also had improved operational efficiency than the traditional relational database.

Key words: ontology; storage; knowledge representation; Neo4j; RDF

(责任编辑:方惠敏)