

中文本体三元组的单字索引与更新方法研究^{*}

■ 许德山 张运良 李芳

[摘要] 以实现中文本体三元组的快速索引和更新为目的,首先分析中文本体的知识结构,并通过 OWL 建模语言将其内部的各种知识单元转换成关系三元组和属性三元组构成的知识网络;然后利用《通用规范汉字表》对体现实体间语义信息的关系三元组进行编码,形成主语前缀和宾语前缀索引表,同时对描述实体特征的属性三元组进行编码,形成主语前缀和主语后缀索引表;接下来详细论述关系和属性三元组的更新方法;最后选取多个本体文件对三元组的汉字索引和更新方法进行测试和分析。结果表明,单字索引能够有效地减少本体三元组的更新和检索时间。

[关键词] 本体索引 三元组索引与更新 知识库构建

[分类号] TP391

DOI: 10.13266/j.issn.0252-3116.2014.22.018

1 引言

近年来,随着社交网络和科技领域的发展,新产生的词汇和领域知识越来越多,领域本体资源也越来越庞大。作为一种形式化的领域知识模型,本体通过多个三元组的组合将特定的概念组织起来,形成知识网络。一个具有服务能力的本体知识库通常存有大量的三元组,如何对这些三元组进行管理、存储和压缩,是每一个服务系统都需要面临和考虑的问题。本文提出利用汉字表对本体概念和三元组进行索引,形成前缀和后缀存储文件,以便于在检索过程中快速地查找和定位所需的三元组信息,为大规模知识服务提供有效保障。

2 本体知识库的组织与表示

本体的形式化描述为计算机处理领域知识带来了便利,为了满足应用需求,国外研究机构开发了多种英文本体管理工具^[1-2],国内的学术界也在积极开展本体建设和管理方面的研究^[3-4]。本体的每一个知识表示元素可以被看作一个知识片断,每一个知识片断包含概念的名称、关系、属性、定义和说明等元素,这些元素通过各种语义关系将不同的概念实体建立联系,

共同描述领域知识结构。虽然不同的本体描述的领域内容各不相同,但它们都有统一的建模元素,这些知识表示元素主要包括:

(1) 概念类:类是本体中最主要的知识单元,也是本体结构的基本组织单元,用于描述具有共性的实例对象。在本体的实现中,概念用类(class)来定义,通常具有一定的层次结构,是构成领域知识的术语系统。

(2) 属性:属性是概念的内部结构特征,描述概念的各种性质,是一个概念区别于其他概念的个性化标识。

(3) 关系:概念之间产生联系的纽带,是描述领域知识的基础框架。本体内的各种概念通过关系建立连接,形成知识网络。

(4) 约束条件:每个属性或关系可以从多个侧面进行描述,从而对概念的属性或关系的取值范围以及使用方式进行规范,这些侧面构成了本体知识结构的约束条件。

(5) 实例:概念描述了具有特定性质的一类实体,是领域知识的抽象描述。而实例则是领域概念在知识库中的具体载体,大量的领域知识需要通过实例属性及其之间的关系进行表示。

在本体知识库中概念类、属性、关系、约束条件和

^{*} 本文系“十二五”国家科技支撑计划项目“科技知识组织体系共享服务平台建设”(项目编号:2011BAH10B03-2)和中国科学技术信息研究所重点工作项目“结构化知识服务平台建设及应用”(项目编号:ZD2014-3-2)研究成果之一。

[作者简介] 许德山,中国科学技术信息研究所信息技术支持中心助理研究员,E-mail: xuds@istic.ac.cn; 张运良,中国科学技术信息研究所信息技术支持中心副研究员,博士; 李芳,昌吉学院计算机工程系副教授,博士。

收稿日期:2014-10-08 修回日期:2014-11-01 本文起止页码:111-116 本文责任编辑:徐健

实例等元素共同描述领域知识,其中概念类、关系、属性、约束条件构成知识框架,概念类填充实例后,就构成了彼此相连的知识网络。虽然这些元素的语义特征并不相同,但它们之间都可以通过“主语-关系-宾语”、“主语-属性-取值”两种形式产生联系,因此整个的本体知识结构都可以转化为关系三元组和属性三元组,这就为本体知识结构的索引和存储带来了便利。本文将探索利用汉字编码表对构成领域知识的关系三元组和属性三元组进行索引的方法,为概念间语义信息的检索和知识服务提供底层支撑。三元组索引和更新的整体思路如图1所示:

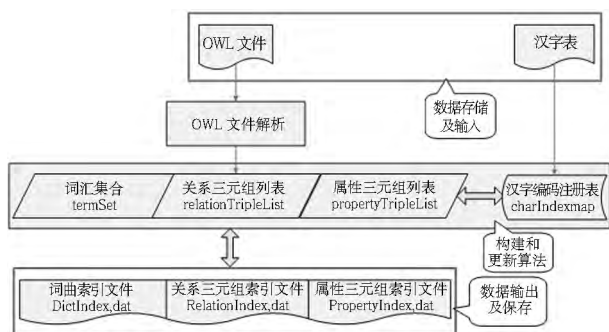


图1 三元组索引和更新的总体策略

本文中构建和使用的本体采用 OWL 网络本体语言进行描述,OWL 是 W3C 推荐的标准本体建模语言,具有丰富的描述元语和推理性能,以保证对概念类、类间关系、属性和个体的语义描述^[5]。单个的 OWL 本体通常使用 XML 文件进行保存,其内部的概念结点和关系构成了领域的知识框架。由于单个的本体文件所提供的信息有限,索引模块应提供数据整合功能,将多个本体中相同或相似的概念建立了语义链接,将新的三元组更新到索引表中,形成多领域的知识库。

3 汉字编码注册表的构建

作为组成词汇的基本单位,中文的每个汉字都有其自身的意义,例如“电动汽车”一词,前缀汉字“电”,后缀汉字“车”,两个汉字都标识了词汇的某一类特征或属性,“电动汽车”与“电”有关,属于一种“车”。如果对词汇的前缀和后缀分别建立索引,能够从不同的侧面高效和快速地定位词汇信息。本文选取《通用规范汉字表》作为构建本体知识三元组的索引基项,该汉字表收录了 8 105 个汉字,这些汉字按照使用频率,分为 3 个等级^[6]。一级字表收录 3 500 个常用汉字,二级字表收录 3 000 个次常用汉字,两个字表合计 6 500 字,可以满足日常阅读、期刊出版、辞书编纂和文字处理等方面的基本需要。三级字表收录了 1 605 个姓氏

人名、地名、科技名词和古文献中的生僻汉字,以满足科研和专门领域的用字需要。

按照汉字的使用频率划分字表可以加快信息的索引和查找,汉字注册模块按一级、二级和三级汉字标识分别进行了注册编码,整个构建过程主要有以下几个步骤:

(1) 首先将《通用规范汉字表》按一级、二级和三级进行整理,并赋以相应的标签,一级字、二级字和三级字的标签缩写分别为 tygf 1、tygf 2 和 tygf 3。

(2) 遍历 tygf 1、tygf 2 和 tygf 3 汉字表,按汉字在字表中的顺序分配编码,形成<汉字,注册码>格式的字码映射表 tygf 1Map、tygf 2Map 和 tygf 3Map。

(3) 将字码映射表以二进制方式保存为 tygf 1.dat、tygf 2.dat 和 tygf 3.dat 文件。

由于《通用规范汉字表》中收录的汉字范围广泛,用其生成的汉字编码具有稳定性,无需经常更新,非常适合作为本体知识结构的索引条目。编码注册表中存有字码键值是汉字在整个知识库中的唯一标识,基于汉字组成的词汇及各种三元组都以该编码作为索引项。

4 三元组索引的构建方法

词汇是本体中组织领域知识的基本元素,更多的复杂概念和语义信息则由多个词汇组成的三元组进行描述。本体中与领域知识有关的三元组主要是关系和属性。关系三元组体现了各种概念间的语义信息,属性三元组用来对概念具有的个性特征进行说明,这些三元组构成本体知识库中的最小知识单元,将各种领域信息组织成相互关联的知识网络,是进行语义推理、关系发现和相似度计算的基本单位。

4.1 关系索引的构建

本体中的关系是组织领域知识的基础框架,其他的语义描述都围绕着概念间的关系网络展开,从而构成具有人机可读性的知识结构。“主语-关系-宾语”构成了关系三元组的基本形式,关系三元组主语和宾语的取值必须是描述领域知识的概念类,以确保关系网络的连通和扩展。从关系三元组的结构来看,主语和宾语由描述领域知识的特定词汇组成,这些词汇在词典索引表中具有唯一的注册编码,可以利用编码作为索引项,将关系三元组分别建立主语前缀和宾语前缀索引。关系索引的构建过程见图2。

三元组索引模块首先加载汉字编码表形成 Map<String,Integer> 格式的映射表,然后解析 OWL 文件,在

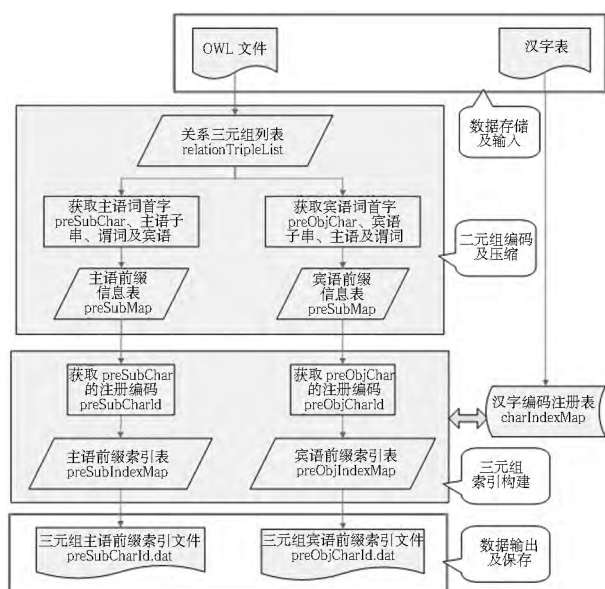


图2 关系索引的构建流程

内存中构建本体模型。领域知识的概念词汇之间通过特定的关系建立语义链接,这些关系在 OWL 本体中具有统一的特征,都采用 owl: ObjectProperty 元语进行标识, sparql 三元组检索语言可以根据 owl: ObjectProperty 获取所有的关系三元组。三元组获取后,索引模块提取主谓前缀和宾谓前缀与汉字编码表进行匹配,最终形成 Map < String, Map < String, Map < String, Set < String > > > 形式的树状映射表,该映射表含有三元组的主谓首字、主谓无首字子串、谓词和宾语 4 个变量,所有关系三元组以主谓或宾谓首字为主键归并到统一的索引文件中,然后对其余变量进行再次索引形成多层次的树状结构。其中主谓前缀索引以谓词为主键对宾谓进行索引,每一首字对应多个谓词,每一谓词对应多个宾谓。宾谓前缀索引则以谓词为主键对主谓进行索引,每一首字对应多个谓词,每一谓词对应多个主谓。关系三元组索引表的详细构建过程可归纳为以下 6 个步骤:

(1) 加载汉字编码注册表 charIndexMap。

(2) 利用 sparql 表达式“select ? subject ? predicate ? object where { ? subject ? predicate ? object. ? predicate rdf: type owl: ObjectProperty }”获取本体中所有的实例关系三元组,存入 relationTripleList 列表。

(3) 遍历 relationTripleList 中的三元组,截取主谓词首汉字 preSubChar,将 preSubChar 后的子串 preSubString、谓词 predicate 和宾语 object 信息存入 preSubMap 映射表中。同时截取宾谓词首汉字 preObjChar,将 preObjChar 后的子串 preObjString、主谓 subject 和谓词 predicate 信息存入 preObjMap 映射表中。

(4) 在汉字索引表 charIndexMap 中查找 preSubChar 对应的注册编码 preSubCharId,并将 preSubMap 存有的信息以 preSubCharId 为主键存入 preSubIndexMap 列表中,同时查找 preObjChar 对应的注册编码 preObjCharId,并将 preObjMap 存有的信息以 preObjCharId 为主键存入 preObjIndexMap 列表中。

(5) 遍历 preSubIndexMap 列表,获取主谓前缀汉字注册编码 preSubCharId 和谓词-宾谓封装列表 preSubMap,以 preSubCharId 为文件名将 preSubMap 存储到 dat 文件中。

(6) 遍历 preObjIndexMap 列表,获取宾谓前缀汉字注册编码 preObjCharId 和主谓-谓词封装列表 preObjMap,以 preObjCharId 为文件名将 preObjMap 存储到 dat 文件中。

4.2 属性索引的构建

属性三元组用来描述主谓所具有的个性特征,其宾谓的取值通常由文本、数字和时间等内容组成,这些描述依附于主谓存在,与其他概念不产生关联,而属性三元组的主谓通常为概念类的实例,是构成领域知识网络的核心结点,其稳定性和连通性很强,因此属性三元组的辅助索引选择主谓尾字索引更为合适。属性三元组前缀和后缀索引的构建过程如图 3 所示:

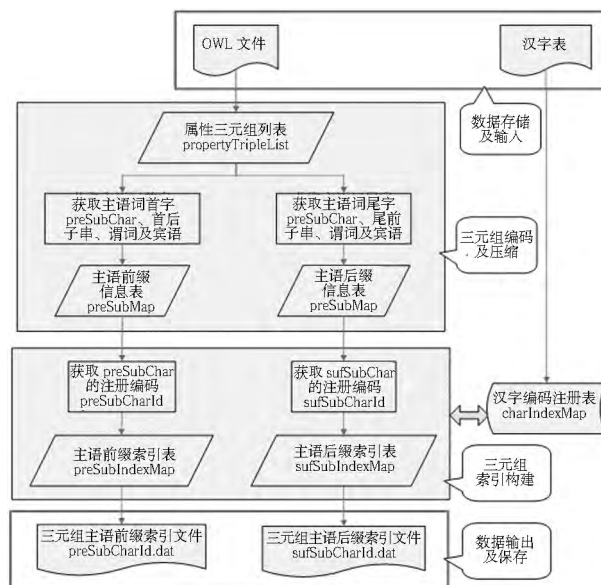


图3 属性索引的构建流程

OWL 本体中的属性采用 owl: DataProperty 元语进行统一标识,OWL 文件解析后便可以根据这一特征获取属性三元组。属性三元组获取后索引模块提取主谓前缀和后缀与汉字编码表进行匹配,然后以主谓首字或尾字为主键将所有属性三元组进行归并处理,并利用谓词对宾谓进行索引形成多层次树状索引表。树状

索引表中每一主键对应多个谓词,每一谓词对应多个宾语。其详细的处理流程由以下 6 个步骤组成:

(1) 加载汉字编码注册表 charIndexMap。

(2) 利用 sparql 表达式 “select ? subject ? predicate ? object where { ? subject ? predicate rdf: type owl: DataProperty }” 获取本体中所有的属性三元组,存入 propertyTripleList 列表。

(3) 遍历 propertyTripleList 中的三元组,截取主语词首汉字 preSubChar,将 preSubChar 后的子串 preSubString、谓词 predicate 和宾语 object 信息存入 preSubMap 映射表中。同时截取主语词尾汉字 sufSubChar,将 sufSubChar 前的子串 sufSubString、谓词 predicate 和宾语 object 信息存入 sufSubMap 映射表中。

(4) 在汉字索引表 charIndexMap 中查找 preSubChar 对应的注册编码 preSubCharId,并将 preSubMap 存有的信息以 preSubCharId 为主键存入 preSubIndexMap 列表中,同时查找 sufSubChar 对应的注册编码 sufSubCharId,并将 sufSubMap 存有的信息以 sufSubCharId 为主键存入 sufSubIndexMap 列表中。

(5) 遍历 preSubIndexMap 列表,获取主语前缀汉字注册编码 preSubCharId 和谓词-宾语封装列表 preSubMap,以 preSubCharId 为文件名将 preSubMap 存储到 dat 文件中。

(6) 遍历 sufSubIndexMap 列表,获取主语后缀汉字注册编码 sufSubCharId 和谓词-宾语封装列表 sufSubMap,以 sufSubCharId 为文件名将 sufSubMap 存储到 dat 文件中。

5 三元组索引文件的更新操作

领域知识的结构框架通常由本体中概念间的关系和属性描述组合而成,而具体的事实则由实例三元组进行描述。实例三元组的规模和数量会伴随着知识库的完善不断地增加,索引文件的规模也会不断增加,如果每次更新都遍历所有的索引文件必然耗时过多,因此在更新前必须挑选出需要更新的索引文件,缩小更新文件的规模和范围。索引更新模块主要负责关系三元组主语前缀、宾语前缀以及属性三元组主语前缀、主语后缀 4 种索引文件的更新操作。

5.1 关系索引的更新

关系索引更新操作也需要汉字编码注册表的辅助,索引的更新操作与构建过程的不同之处在于:索引更新之前需要将更新内容与索引表中存储的内容进行对比,避免三元组信息的重复存储。关系索引的更新

过程如图 4 所示:

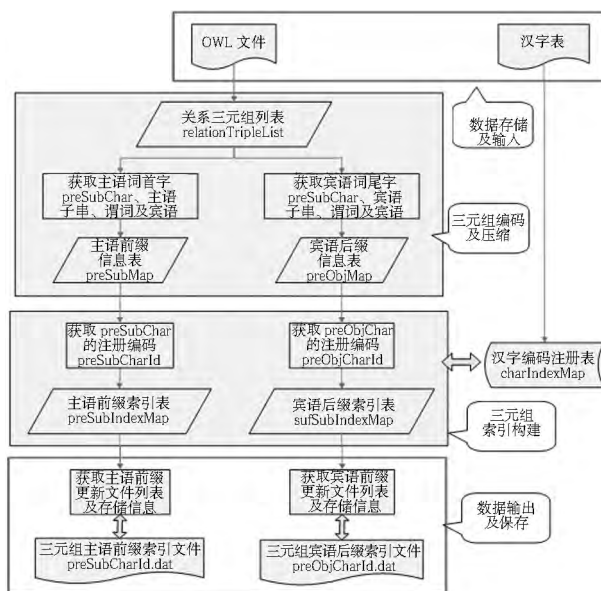


图 4 关系索引的更新流程

由于三元组索引以主语词汇和宾语词汇作为索引项,关系三元组的更新操作要同时完成主语前缀索引文件和宾语前缀索引文件的更新。在进行索引更新时首先根据主语的词首汉字和宾语的词首汉字获取需要更新的索引文件,然后检测索引文件中是否含有待更新的三元组信息,只有在确保新增的三元组信息未在关系索引文件中记录时再完成更新操作,以避免频繁、大量的字符串匹配,其详细流程由以下 6 个步骤组成:

(1) 加载汉字编码注册表 charIndexMap。

(2) 获取本体中所有实例的关系三元组,存入 relationTripleList 列表。

(3) 遍历 relationTripleList 中的三元组,截取主语词首汉字 preSubChar,将 preSubChar 后的子串 preSubString、谓词 predicate 和宾语 object 信息存入 preSubMap 映射表中。同时截取宾语词首汉字 preObjChar,将 preObjChar 后的子串 preObjString、主语 subject 和谓词 predicate 信息存入 preObjMap 映射表中。

(4) 在汉字索引表 charIndexMap 中查找 preSubChar 对应的注册编码 preSubCharId,并将 preSubMap 存有的信息以 preSubCharId 为主键存入 preSubIndexMap 列表中,同时查找 preObjChar 对应的注册编码 preObjCharId,并将 preObjMap 存有的信息以 preObjCharId 为主键存入 preObjIndexMap 列表中。

(5) 遍历 preSubIndexMap 列表,获取主语前缀汉字注册编码 preSubCharId 和谓词-宾语封装列表 preSubMap,读取以 preSubCharId 为文件名的索引文件,判断该文件中是否含有 preSubMap 存储的信息,如果不

含相应内容,则将 preSubMap 中的信息添加到 preSubCharId 对应的 dat 索引文件中。

(6) 遍历 preObjIndexMap 列表,获取宾语前缀汉字注册编码 preObjCharId 和主语-谓词封装列表 preObjMap,读取以 preObjCharId 为文件名的索引文件,判断该文件中是否含有 preObjMap 存储的信息,如果不含相应内容,则将 preObjMap 中的信息添加到 preObjCharId 对应的 dat 索引文件中。

5.2 属性索引的更新

新增的属性三元组与特定的实例描述有关,属性索引的更新操作与关系索引的更新方法相似,但索引项不同。属性三元组的更新与主语首字和尾字有关,需要获取对应的前缀和后缀索引文件,其更新方法如图5所示:

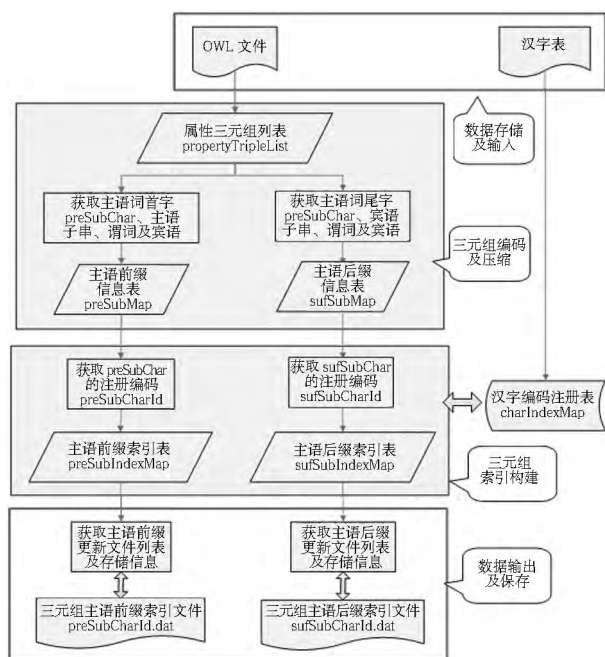


图5 属性索引的更新流程

由于属性三元组宾语的取值不适合做索引项,属性索引的更新仅涉及主语信息匹配和计算,索引更新时需要获取属性三元组的主语词汇,利用主语的首字和尾字确定需要更新的属性索引文件,然后再次判断哪些数据需要更新,完成数据的添加操作。其详细的处理流程由以下6个步骤组成:

(1) 加载汉字编码注册表 charIndexMap。

(2) 获取本体中所有的属性三元组,存入 propertyTripleList 列表。

(3) 遍历 propertyTripleList 中的三元组,截取主语谓词首汉字 preSubChar,将 preSubChar 后的子串 preSubString、谓词 predicate 和宾语 object 信息存入 preSub-

Map 映射表中。同时截取主语词尾汉字 sufSubChar,将 sufSubChar 前的子串 sufSubString、谓词 predicate 和宾语 object 信息存入 sufSubMap 映射表中。

(4) 在汉字索引表 charIndexMap 中查找 preSubChar 对应的注册编码 preSubCharId,并将 preSubMap 存有的信息以 preSubCharId 为主键存入 preSubIndexMap 列表中,同时查找 sufSubChar 对应的注册编码 sufSubCharId,并将 sufSubMap 存有的信息以 sufSubCharId 为主键存入 sufSubIndexMap 列表中。

(5) 遍历 preSubIndexMap 列表,获取主语前缀汉字注册编码 preSubCharId 和谓词-宾语封装列表 preSubMap,读取以 preSubCharId 为文件名的索引文件,判断该文件中是否含有 preSubMap 存储的信息,如果不含相应内容,则将 preSubMap 中的信息添加到 preSubCharId 对应的 dat 索引文件中。

(6) 遍历 sufSubIndexMap 列表,获取主语后缀汉字注册编码 sufSubCharId 和谓词-宾语封装列表 sufSubMap,读取以 sufSubCharId 为文件名的索引文件,判断该文件中是否含有 sufSubMap 存储的信息,如果不含相应内容,则将 sufSubMap 中的信息添加到 sufSubCharId 对应的 dat 索引文件中。

6 研究结论

本体三元组索引的构建和更新操作是领域知识库系统的核心功能,对知识库中的大量三元组进行管理、存储和压缩,是提高系统服务能力的有效手段。本文根据关系三元组和属性三元组的各自特征设计了不同的索引方法,以便于在检索过程中快速地查找和定位所需的三元组信息,支持大规模的知识服务应用。本文的研究工作以科技本体发布与服务平台为基础,是其中本体知识链接模块的重要组成部分^[7]。该平台旨在为多个本体资源提供规范化的管理和链接方式,并通过统一的服务接口供外部用户使用。本文研究选取本体发布与服务平台中的新能源本体作为索引构建数据用例,新能源本体采用 owl 语言构建,拥有“政策”、“资源”、“产品”和“技术”4个大类,广泛覆盖了新能源领域的相关概念,其整体结构见图6。

新能源本体中上下位关系的最大深度超过4层,概念众多,关系复杂,其内部含有526个概念类、12种类间关系、23种属性描述以及2762个实例,这些实例通过各种关系链接组成一个庞大的知识网络。经过解析处理后,获取关系三元组46484个,属性三元组43342个。此外,本文从新能源汽车、新一代生物技术

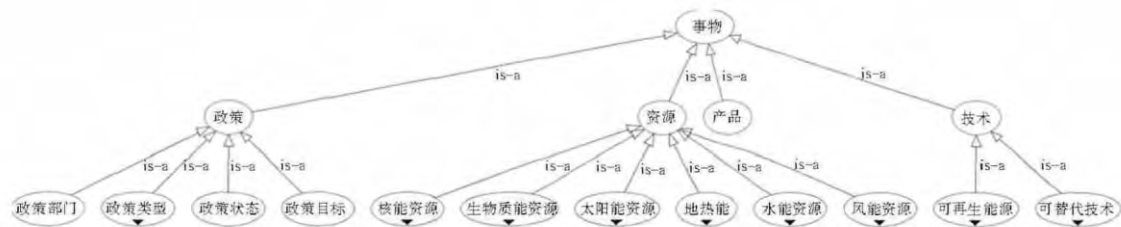


图6 新能源本体的总体结构

和新材料3个本体中整理了与新能源相关的83 123个关系三元组和54 081个属性三元组作为更新数据用例。上述数据在实验机器上进行了测试,实验机器的基本配置为:AMD Athlon II×2 B26 Processor 3.20GHz的CPU,内存4G,32位Windows7系统。索引构建和更新模块的运行情况如表1所示:

表1 三元组索引构建和更新效率

操作项目	数据类型	数量(条)	用时(秒)	效率(条/秒)
构建操作	关系三元组	46 484	165	281
	属性三元组	43 342	144	300
更新操作	关系三元组	83 123	55	1 511
	属性三元组	54 081	40	1 352

由于索引的构建和更新以《通用规范汉字表》表中汉字为基准,三元组前缀和后缀索引文件分别有8 105个文件,每个文件中存有以汉字注册编码为前缀或后缀的三元组信息。由于汉字数量较为固定,采用汉字注册编码作为索引项对三元组结构进行索引能够有效地减少索引文件的数量和规模。为了测试三元组的检索效率,本文随机选取部分实例进行测试,程序运行显示其检索时间小于0.001秒,概念词汇输入后能

够快速的匹配到相应的知识结构。从运行结果来看,索引构建和更新阶段由于涉及大量的字符串匹配以及本地文件的读取和存取操作,消耗时间较长。但因作为后台程序运行,不需频繁操作,对系统整体性能的影响较小,后期可以采用分布式存储和并行方式对索引构建和更新算法进行完善,提升其运行效率。

参考文献:

- [1] Auduna. Sesame [EB/OL]. [2014 - 07 - 23]. <http://www.openrdf.org/about.jsp>.
- [2] TopQuadrant. TopBrand Enterprise Vocabulary Net [EB/OL]. [2014 - 07 - 22]. <http://download.topquadrant.com/evn/45doc/evnIntro.html>.
- [3] 徐立恒,刘洋,来斯惟,等. 基于多特征表示的本体概念挂载[J]. 中文信息学报, 2012, 26(3): 122 - 128.
- [4] 唐晓波,胡华. 中文UGC信息源的本体概念抽取研究[J]. 现代图书情报技术, 2014(5): 41 - 49.
- [5] Manola F, Miller E. RDF primer [EB/OL]. [2014 - 08 - 23]. <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.
- [6] 通用规范汉字表[M]. 北京: 语文出版社, 2013.
- [7] 许德山,张运良. 集成化本体管理平台的设计与实现[J]. 数字图书馆论坛, 2013(11): 15 - 20.

The Building and Updating Methods of Chinese Ontology Triples Index Based on Chinese Characters

Xu Deshan¹ Zhang Yunliang¹ Li Fang²

¹Institute of Science and Technical Information of China, Beijing 100038

²Changji College, Xinjiang 831100

[Abstract] This paper focuses on the building and updating methods of Chinese ontology triples index based on Chinese characters. The architecture of Chinese ontology is described firstly, as well as knowledge network consisting of relation triples and property triples through OWL. And then, the building and updating methods about relation and property triples indexing by Chinese characters are introduced in details. The authors make an experiment with a large number of instances in order to test performance of index and update algorithms. The result indicates indexing Chinese triples with Chinese characters can reduce the retrieval and update time. Finally, the article discusses the experimental result and proposes the research emphasis in the future.

[Keywords] ontology index triple index and update repository building