CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 38 卷第 7 期 2016 年 7 月
Vol. 38, No. 7, Jul. 2016

文章编号:1007-130X(2016)07-1322-08

基于 HBase 的领域本体存储方法及其应用研究*

王 红,孙 康
(中国民航大学计算机学院,天津 300300)

摘 要:在分析民航突发事件应急管理领域本体及其存储特点的基础上,提出了一种基于 HBase 的领域本体存储方法,采用将领域本体元数据与 RDF 实例数据分开存储的方式,给出了描述领域本体类及属性信息的元数据和 RDF 实例数据的存储模型,及其基于 MapReduce 的领域本体 RDF 数据并行加载过程。结合应用实现了领域本体基于 HBase API 的基本图模式查询,并在 Hadoop 环境下进行了实验与效果分析,为民航应急管理领域本体的海量数据存储提供了理论与方法支撑。

关键词:民航应急管理;领域本体;RDF;存储模型;HBase
中图分类号:TP311 文献标志码:A
doi:10.3969/j.issn.1007-130X.2016.07.004

A domain ontology storage method and its applications based on HBase

WANG Hong, SUN Kang
(School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Based on the analysis of the civil aviation emergency management domain ontology and its storage features, we propose a domain ontology storage method based on HBase. The storage data of domain ontology metadata and the instance data are stored separately. We describe domain ontology classification, the metadata of attribute information, and the storage model of RDF instance data, as well as the current loading process of RDF data based on MapReduce. Combining with practical applications, a basic graph pattern query algorithm of domain ontology based on HBase API is implemented. We analyze the experimental results in the Hadoop environment, which show that the proposed method can provide theoretical support and a new method for storing the huge data of civil aviation emergency management domain ontology.

Key words: civil aviation emergency management; domain ontology; RDF; storage model; HBase

1 引言

领域本体(Domain Ontology)是指对特定领域内概念及概念间关系的形式化表达。目前有诸多利用分布式系统处理海量 RDF 数据管理问题的相关研究[1~4],相关研究大多为针对在分布式平台下 RDF 实例数据的存储与查询方法本身,鲜有应用案例。文献[5]给出了民航突发事件应急管理领域本体的构建方法,文献[6]讨论了民航应急管理领域本体的关系型数据库存储方法。

由于领域本体与关系型数据模型之间转换与映射开销较大,本体数据的稀疏性特点会导致数据表中出现大量空值,造成存储空间利用率下降,本体的查询与管理效率无法适应大数据环境对民航应急管理的要求。本文在前期民航应急管理的研

* 收稿日期:2015-05-21;修回日期:2015-09-30
基金项目:国家自然科学基金委员会与中国民用航空局联合资助项目(61079007);国家自然科学基金青年基金(61201414);国家青年基金项目(61301245)
通信地址:300300 天津市东丽区中国民航大学计算机学院
Address: School of Computer Science and Technology, Civil Aviation University of China, Dongli District, Tianjin 300300, P. R. China

究基础之上,提出了基于 HBase 的应急管理领域本体存储模型。

2　领域本体及其存储方法分析

2.1　民航突发事件应急管理领域本体

民航突发事件应急管理领域本体的构建主要依据"领域词典"[5],按照民航局以及各大机场等管理部门制定的《民航突发事件应急预案》《应急管理规定》《应急救援计划(或手册)》中的相关术语与业务流程,提取出民航应急预案、应急案例、应急处置过程和应急资源等核心概念、实例及其之间的关系(如图 1 所示)。

其中:

(1)领域本体的概念:由顶层概念(类)与子类概念构成。顶层概念包括应急案例、应急预案、应急处置过程和应急资源四大类;顶层概念可以进一步划分为若干子类。概念之间的关系主要包括子类关系(SubClassOf)及其相关的数据与对象属性。类与属性的信息保存在 OWL 本体描述文件中,其数据格式如下所示:

```
〈owl:Class rdf:about="http://www.cauc-caedo.com/caedo.owl#航空器紧急事件"〉
  〈rdfs:subClassOf rdf:resource="http://www.cauc-caedo.com/caedo.owl#应急案例"/〉
〈/owl:Class〉
```

(2)领域本体的实例:实例描述如图中菱形框所示,与上层概念的关系为 rdf:type。采用 RDF 的形式描述,其数据格式如下所示:

```
〈caedo:航空器失事 rdf:about="http://www.cauc-caedo.com/caedo.owl#717MH17"〉
  〈rdf:type rdf:resource="http://www.cauc-caedo.com/caedo.owl#航空器失事"/〉
  ……
〈/caedo:航空器失事〉
  〈caedo:航空器失事 rdf:about="http://www.cauc-caedo.com/caedo.owl#1121MU5210"〉
  〈rdf:type rdf:resource="http://www.cauc-caedo.com/caedo.owl#航空器失事"/〉
  ……
〈/caedo:航空器失事〉
```

其中 717MH17 代表 7 月 17 日马航 MH17 班机坠毁事故,1121MU5210 代表 11 月 21 日东方航空 MU5210 号航班事故等。其事故类型为航空器失事。

2.2　领域本体的存储方法分析

领域本体一般采用三元组[7]或五元组[8]等形式化定义,以 OWL 或 RDF 语言进行描述。其中,资源描述框架 RDF[9]是本体描述语言的基础,用于描述资源的基本数据模型,包括资源(Resource)、属性(Property)和陈述(Statements),一般采用(主语,谓词,宾语)三元组的形式表示,主语为被描述的资源,用 URI 来表示,谓词为属性,宾语表示其它资源或文本,或者表示为主语的属性值。OWL[10]是 W3C 推荐的标准 Web 本体描述

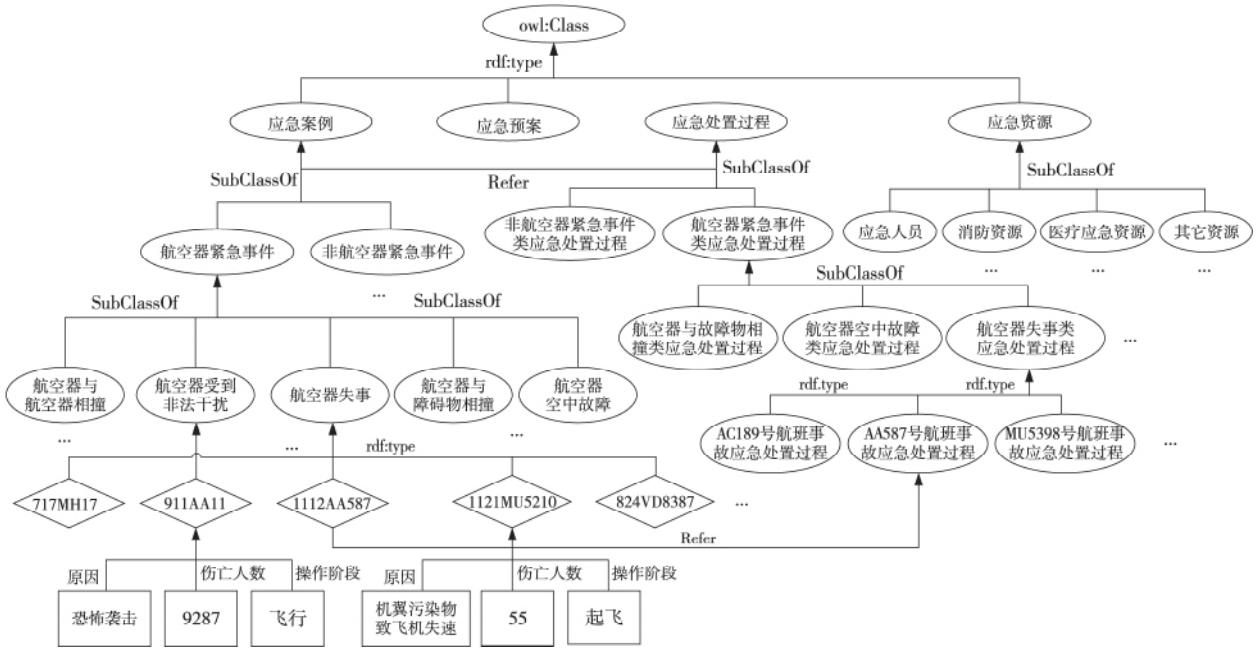

Figure 1　Core concepts and their main relationships of the civil aviation emergency management domain ontology

图 1　民航应急管理领域本体核心概念与主要关系构成

语言,构建于 RDF、RDFS 基础之上,虽然 RDF 与 OWL 在表达能力与语义方面上存在一些差异,但是 OWL 本体可采用 RDF 三元组格式来进行描述。本文重点研究领域本体的 RDF 存储方法。

2.2.1 RDF 的文件与关系型数据库存储

RDF 一般采用文件式存储与集中式关系型数据库存储。

RDF 的文件式存储即基于 XML 文档的 rdf 或 owl 文件。文件式 RDF 存储方法实现简单,但查询时须对整个 OWL 文件进行解析,只适用于数据量较少的情况。RDF 的关系型数据库存储是将 RDF 三元组映射至关系型数据模型。目前基于关系型数据库的存储模型主要有简单的三列表、水平存储、属性表等[11]。RDF 的多值问题难以在关系型数据库中表达。本体数据的稀疏性会导致关系数据表中出现大量空值,造成存储空间利用率下降、查询效率降低,并且关系型数据表结构不适合海量数据模式动态变化的需求[12]。文件式存储和集中式关系型数据库存储都已难以适应大规模 RDF 数据的高效管理需求。

2.2.2 RDF 的分布式数据库存储

基于分布式系统解决本体存储和查询的问题是当前的研究热点,基于 Apache Hadoop 平台[13]与 HBase[1]来构建 RDF 存储系统是广泛关注的方向。其中 Hadoop 是由 Apache 基金会开发的分布式系统基础架构,包含 HDFS、MapReduce 两个核心组件,其中 HDFS 是分布式文件存储系统;MapReduce 是一种并行计算模型,用于大规模海量数据的并行计算。

HBase 是基于 HDFS 的面向列的分布式数据库,其列式存储的特点是解决了数据稀疏性存储问题,最大程度节省了存储开销,提升查询效率的同时还拥有很高的扩展性与可靠性,为本体在 RDBMS 中存储的多值与空值问题提供了很好的解决方案。

文献[2]采用 HBase 与 Hadoop 作为 RDF 存储与查询的平台,给出了以主体谓词为行健、客体为列族的 SP_O 存储结构,创建了 PS_O、PO_S、OP_S、OS_P、SO_P 多张索引表的方式。文献[3]提出了将 RDF 三元组存储于 SP_O、PO_S、OS_P 三张存储表中的方案,三张表都只包含了一个列族,每行数据都存储在一个列族中,由于其对于元数据和实例数据皆采用统一方式划分,数据冗余度较高。文献[4]采用了实例数据按类划分的方法,为本体中的每个类创建了两张实例数据表,分别以主体和客体为行健,谓词为列限定符,客体与主体作为值。其没有给出元数据的查询方法,且在查询实例数据时仅给出谓词且谓词不是 rdf:type 的情况下,查找将退化为全表扫描的方式,查询效率将无法保证。

3 基于 HBase 的分布式领域本体 RDF 存储方法

3.1 HBase 数据模型分析

HBase 数据模型是一个分布式、多维有序的映射。设行健为 R,列族为 C,列限定符为 CQ,时间戳为 T,值为 V,其数据模型的形式化表示如下[10]:

$$R \rightarrow \{C_i \rightarrow \{CQ_i \rightarrow \{T_i \mid V_i\}\}\}, i = 0,1,2,\cdots,n$$

其中,行健直接对应一至多个列族,而一个列族由若干列组成,单元格由行健、列族:列限定符、时间戳唯一确定,具有由行健、列族:列限定符、时间戳到单元格组成的多维有序映射结构。

HBase 表中所有的数据都是采用字节数组的形式存储的原始数据(Raw Data)。在 HBase 中列是最基本的单位,一到多个列组成一行,并由唯一的行健来确定,所有的行按照行健字典序排序存储,并可以由时间戳来存放不同版本的数据。HBase 的表具有稀疏性,其在底层物理存储时按列族存储,只存放有内容的表格单元,逻辑上值为空的列实际并不占用存储空间。所以,表可以设计得非常稀疏,十分适合用来存储稀疏的 RDF 数据。

3.2 基于 HBase 的领域本体存储模型

3.2.1 元数据存储

民航突发事件应急管理领域本体由元数据和 RDF 实例数据组成。其中元数据为描述实例的数据,即类与属性的信息,其保存在 OWL 本体描述文件中。根据民航应急管理领域本体的特点,采用将元数据与实例数据分开存储的方式,将民航突发事件应急管理领域本体的元数据信息存放在 CAEDOClass 表中,再将实例数据保存在每个类的实例数据表中。将数据按类划分,一是针对数据源无需过多额外的预处理工作;二是可以有效地缩小查询范围,从而提高查询效率。本文的存储模型建立在文献[13]提出的存储模型基础之上,针对在仅给出谓词且谓词不是 rdf:type 的情况下查询会退化为全表扫描的情况,为每个类添加一个 Class-

Name_PSO 实例数据表,从而具备了更加高效的索引结构,使得除了(?S ?P ?O)之外任何形式的三元组模式查询都有较好的查询效率。

领域本体元数据的存储模型为:

$$ClassName \rightarrow \{(SubClass, Property) \rightarrow \{Value_i \rightarrow \{\text{null}\}\}\}, i = 0,1,2,\cdots,n$$

即以 *ClassName* 为行健,子类和属性两个列族表示类的子类关系信息与属性信息,用列限定符存储具体值,将单元格设计为空值,根据 HBase 对单元格为空值不占用存储空间的特点,可以减少不必要的存储开销。由于列式存储的灵活性和表结构的稀疏性,可以动态地增加列来存储多值。其逻辑存储结构如表 1 所示(表中省略了属性列与时间戳)。

Table 1　Logic storage structure of CAEDOClass

表 1　CAEDOClass 逻辑存储结构

行健	子类(SubClass)			
应急案例	子类:航空器紧急事件	子类:非航空器紧急事件		
应急预案	子类:国家总体应急预案	子类民航局应对突发事件总体预案	子类地方专项应急预案	子类民航机场/航空公司应急预案
应急处置过程	子类:航空器紧急事件类应急处置过程	子类:非航空器紧急事件类应急处置过程		
应急资源	子类:医疗急救资源	子类:消防资源	子类:应急人员	子类:其它资源

根据输入三元组模式语句快速定位到该实例所属类的实例数据表,这样可以达到有效缩小查询范围的目标。创建 Class_type 表存储实例所属类及元数据所属类型 owl:Class,以实例名为行健,列限定符存储实例所属的类。创建 CAEDOProperty 表存储属性信息,以属性名为行健,定义域、值域两个列族,列限定符存储具体值。

3.2.2　实例数据存储

为每个类创建三张实例数据表:ClassName_SPO、ClassName_OPS 和 ClassName_PSO,以满足所有组合形式的三元组模式查询匹配条件需求。ClassName_SPO 表以该类实例数据三元组的 *Subject* 作为行健,列限定符存储 *Predicate*、*Object* 为单元值,表中数据均是以此类的实例作主语的三元组,表中只有一个列族 I,列限定符由此类具有的所有属性组成。

ClassName_OPS 表以该类实例数据三元组的 *Object* 作为行健,列限定符存储 *Predicate*、*Subject* 为单元值。ClassName_PSO 表以该类实例数据三元组的 *Predicate* 作为行健,列限定符存储 *Subject*、*Object* 为单元值。

作为示例,本体描述文件中定义了事件类航空器失事:

〈owl:Class rdf:about =" http://www.cauc-caedo.com/domain.owl#航空器失事" 〉〈/owl:Class〉

并且定义域为航空器失事的属性,有操作阶段、原因、遇难人数等。航空器失事_SPO 实例数据如表 2 所示。

Table 2　Logic storage structure of Aircraft Crash_SPO

表 2　航空器失事_SPO 逻辑存储结构

行健	I			
	I:rdf:type	I:操作阶段	I:原因	I:遇难人数
1121MU5210	航空器失事	起飞	机翼污染物致飞机失速	55
717MH17	航空器失事	起飞	飞机被导弹击中	298
824VD8387	航空器失事	着陆	飞行机组违反规定	96

4　应用与效果分析

4.1　基于 MapReduce 的数据加载

串行的加载方式在数据量较大的时候往往会耗费很长的时间。本文采用通过 MapReduce 并行加载的方式进行数据加载。图 2 描述了 MapReduce 并行加载实例数据时的数据流。

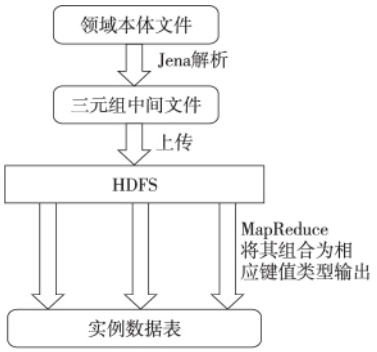

Figure 2　Domain ontology loading process

图 2　领域本体加载过程

首先将本体数据通过 Jena 解析为(*Subject*, *Predicate*, *Object*)格式的三元组中间文件,上传至 HDFS,而后通过 MapReduce 作业将其加载至 HBase 中。

加载元数据时,map 任务输入的数据 value 为(*Subject*, *Predicate*, *Object*),因为元数据表的列族为 SubClass,所以这里将原 *Subject* 作为 *Ob-*

ject,原 *Object* 作为 *Subject*,只将 *Predicate* 为“rdfs:SubClassOf”与“rdfs:domain”的三元组组合成(*Subject*,(*Predicate*,*Object*))的形式输出。reduce 任务获取 map 任务的输出,即(*Subject*,ListOf(*Predicate*,*Object*))键值形式的数据,并只将 *Predicate* 为“rdfs:SubClassOf”与“rdfs:domain”的三元组封装成 Put 对象输出至 HBase。其中 *Subject* 作为 Put 对象的行健,列族为“SubClass”或”Property”,*Object* 作为列限定符,单元值为空值。

加载实例数据时,map 任务将以(*Subject*,*Predicate*,*Object*)格式输入的数据组合成(*Subject*,(*Predicate*,*Object*))的形式输出,reduce 任务获取 map 任务的输出,并将(*Subject*,ListOf(*Predicate*,*Object*))键值形式的数据封装成一个 Put 对象输出至 HBase。其中 *Subject* 作为 Put 对象的行健,列族为”I”,*Predicate* 作为列限定符,Object 作为单元值。

以加载实例数据为例,map 任务将以(*Subject*,*Predicate*,*Object*)格式输入的数据组合成(*Subject*,(*Predicate*,*Object*))的形式输出,reduce 任务获取 map 任务的输出,并将(*Subject*,ListOf(*Predicate*,*Object*))键值形式的数据封装成一个 Put 对象输出至 HBase。其中 *Subject* 作为 Put 对象的行健,*Predicate* 作为列限定符,*Object* 作为单元值。

有如下 RDF 实例:

```
〈caedo:航空器失事 rdf:about=" http://www.cauc-caedo.com/caedo.owl#1121MU5210"〉
  〈rdf:type rdf:resource=" http://www.cauc-caedo.com/caedo.owl#航空器失事" /〉
  〈caedo:原因〉机翼污染物致飞机失速〈/caedo:原因〉
  〈caedo:操作阶段〈起飞〈/caedo:操作阶段〉
  〈caedo:遇难人数〉55〈/caedo:遇难人数〉
〈/caedo:航空器失事〉
```

通过 Jena 解析为三元组文件,格式如下:

Subject:1121MU5210 *Predicate*:http://www.w3.org/1999/02/22-rdf-syntax-ns#type Object:航空器失事

Subject:1121MU5210 *Predicate*:http://www.cauc-caedo.com/caedo.owl#原因 Object:机翼污染物致飞机失速

Subject:1121MU5210 *Predicate*:http://www.cauc-caedo.com/caedo.owl#操作阶段 Object:起飞

Subject:1121MU5210 *Predicate*:http://www.cauc-caedo.com/caedo.owl#遇难人数 Object:55

上传至 HDFS 后,通过 MapReduce 作业将其组合为相应键值类型输出至 HBase,输出结构如表 2 所示。

4.2 基于 HBase API 的 SPARQL 查询算法

在基于 RDBMS 的 RDF 管理系统中,一般直接将 SPARQL 转换成 SQL 语句交给数据库来执行,其底层查询过程对于用户来说完全透明。HBase 本身并不支持 SQL 和 SPARQL 的执行,实现 RDF 数据在 HBase 中的查询就需要完成从 SPARQL 查询语句到 HBase 的查询流程与接口的转换。

三元组模式[14]是一个三元组,表达形式为 $tp=(sp,pp,op)$,采用 V 代表变量集合,则有 $sp\in V\cup URI\cup B$,$pp\in V\cup URI$,$op\in V\cup URI\cup B\cup L$,其中,$sp$、$pp$、$op$ 即为三元组的主语、谓语、宾语,B 表示空节点集合,L 为文本描述集合,URI 即表示统一资源标识之和。基本图模式[14]由一组三元组模式组成,是 SPARQL 查询的基本组成部分。下面分别给出三元组模式与基本图模式查询算法。

4.2.1 三元组模式查询算法

算法 MatchCAEDO_TP 接受输入的资源与变量组成的三元组,查询请求中指定的元组为常量,未指定的元组为变量。算法将三元组模式查询语句转换为 HBase API 进行查询,返回相应的三元组。其算法如下:

算法 1 MatchCAEDO_TP

输入:三元组模式 $tp=(sp,pp,op)$;

输出:相匹配的三元组集合 *Result*。

开始

Result 初始化为空集

if ($tp.sp$ 为常量)

以 $tp.sp$ 为 row-key 在 Class_type 表中查询列族 *is* 中的列限定符 $c1$,得到实例所属的类;

if($c1$ 为 owl:Class)

以 $tp.sp$ 为 row-key 查询 CAEDOClass 表,将结果加入 *Result*;

else

Result = *Query_By_SPO* (tp,$c1$);

else if ($tp.pp$ 为常量)

以 $tp.tp$ 为 row-key 在 CAEDOProperty 表中查询列族 domain 得到定义域类 $c1$;

Result = *Query_By_PSO* (tp,$c1$);

else if ($tp.op$ 为常量)

以 $tp.op$ 为 row-key 在 Class_type 表中查询列族 *is* 中的列限定符 $c1$,得到实例所属的类;

if(*c*1 为 owl:Class)
以 *tp*.*sp* 为 row-key 查询 CAEDOClass 表将结果加入 *Result*;
else
Result = *Query_By_OPS*(*tp*,*c*1);
else
Result = *Query_All*(*tp*);
return *Result*;
结束

其中,若进行的是元数据的查询(即通过 Class_type 表中查询所得行健数据类型为 owl:Class),则扫描 CAEDOClass 表并返回结果。若为实例数据查询,则分为四种情况:(1) 如果 *Subject* 为已知常量,则调用 Query_By_SPO 查询匹配的三元组,该函数以已知 *Subject* 和类名通过查询相应 *ClassName_SPO* 表来返回匹配的三元组。由于指定了行健,可以快速响应查询。(2) 如果 Predicate 为已知常量,则调用 Query_By_PSO 查询匹配的三元组,该函数以已知 Predicate 常量通过查询 ClassName_PSO 表来返回匹配的三元组。(3) 如果 *Object* 为已知常量,则调用 Query_By_OPS 查询匹配的三元组。(4)其他情况则调用 Query_All 进行全表扫描。这四个底层查询接口直接调用相应 HBase API 查询并返回结果。

4.2.2　基本图模式查询算法

算法 MatchCAEDO_BGP 做总体的调度工作,对每一个三元组模式调用三元组模式查询算法,来查询与 SparQL BGP 查询匹配的三元组集合。算法如下:

算法 2　MatchCAEDO_BGP
输入:基本图模式中所有三元组模式语句;
输出:相匹配的三元组集合 *B*。
开始
B 初始化为空集;
IF (*tp*.*sp* and *tp*.*pp* and *tp*.*op* is not a variable)
将 *tp* 加入 *B* 中;
设 BGP 三元组模式集合为(tp_1,tp_2,…,tp_n);
FOR_EACH tp_i in (tp_1,tp_2,…,tp_n)
if (tp_i与 tp_{i-1},…,tp_1 中的语句有共享变量)
用 B 中该共享变量的值替换 tp_i 中的值形成新的集合 *TP*;
FOR_EACH tp_j in *TP*
B= $B \cup MatchCAEDO\_TP(tp_j)$;
return *B*
结束

算法开始先将输入的 Triple Pattern 对每一条语句先进行共享变量替换,构成新的三元组集合对其调用 MatchCAEDO_TP 算法,并将返回的结果加入结果集中。

4.3　实验与分析

实验基于 5 个节点的 Hadoop 环境。其中主节点 NameNode、JobTracker 以及从节点的主要硬件配置为 Intel Core i7-3770,4 GB 内存,1 TB 机械硬盘。操作系统为 CentOS 7.0.1406,开发环境为 JDK 1.7.0_72,Hadoop2.4.1,HBase0.98.23。

采用中国民航大学构建的民航突发事件应急管理领域本体作为实验数据。数据主要构成如图 1 所示,包含 289 个类,171 个对象属性,1 203 个数据属性及 6 138 个实例。

首先通过解析本体元数据文件,创建 CAEDOClass 表存放类的元数据信息,创建 CAEDOProperty 表存放属性信息,创建 Class_type 表存储实例数据所属类。为本体中每个类创建名为 ClassName_SPO、ClassName_OPS 和 ClassName_PSO 的三张实例数据表。将数据集解析为三元组中间文件上传至 HDFS,再通过 MapReduce 任务分别将数据加载至相应表中。

Sesame 是一个典型的基于 RDBMS 的开源 RDF 管理系统框架,底层使用 MySQL 存储架构,为 RDF 建立了多表索引机制,拥有较好的查询性能。测试导入性能时,将实验数据采用类似 LUBM 数据集的组织方式生成三组不同大小的数据,记为 *D*1、*D*2、*D*3,大小分别为 10 MB、225 MB、1.0 GB。测试 Sesame 与 HBase 方案的数据加载时间,结果四舍五入,如表 3 所示。

Table 3　Response time for data loading
表 3　数据加载时间　s

数据加载	Sesame 方案	HBase 方案
*D*1	26	19
*D*2	415	181
*D*3	6 066	1 557

可以看出,相对于 Sesame,基于 MySQL 的方案,HBase 方案在数据加载方面有较大的性能优势,且在数据量增大时优势愈加明显。此外,还可以采取预先创建空 Region,多线程并发写入、批量写入与关闭自动 flush 等优化措施来提升数据的加载效率。

对本文的存储模式、OWL 文件存储模式与 Sesame 方案设计两组不同的三元组查询,对查询时间进行统计,比较它们的平均响应时间。对民航突发事件应急管理领域本体进行以下两种三元组

查询:

(1)元数据的查询,查询某个类拥有的子类即SubClassOf关系:

```
SELECT ?x WHERE{
  ?x 〈http://www.w3.org/2000/01/rdf-schema#subClassOf〉
  〈http://www.owl-ontologies.com/caedo.owl#航空器紧急事件〉
}
```

即查询航空器紧急事件类的子类,该查询有一条三元组模式语句,其中主语为变量,谓词、宾语为已知常量。

(2)实例查询,查询某个事件所属的事件类型:

```
SELECT ?x WHERE{
  〈http://www.owl-ontologies.com/caedo.owl#1121MU5210〉
  〈http://www.w3.org/1999/02/22-rdf-syntax-ns#type〉 ?x
}
```

即东方航空MU5210号航班空难所属事件类型,该查询有一条三元组模式语句,其中主语、谓词为已知常量,宾语为变量。

每组查询运行5次,取平均查询响应时间,其对比如表4所示。

Table 4 Average response time for queries

表4 查询的平均响应时间 ms

查询	OWL方案	Sesame方案	HBase方案
元数据查询(*D*1)	1 819	236	265
实例查询(*D*1)	1 754	238	261
元数据查询(*D*2)	19 339	596	423
实例查询(*D*2)	17 842	604	438
元数据查询(*D*3)	-	7 971	965
实例查询(*D*3)	-	8 142	1 011

两组查询语句响应时间较快,本文的存储模式查询效率比OWL文件式存储要高很多。其主要原因是,OWL文件存储模式查询时必须对整个OWL文件进行解析,其解析时间较长。在数据量较大时文件式查询很快变得极其缓慢,可能受制于实验环境影响,甚至于有时无法测试出实验结果。在数据量较小时,Sesame方案查询时间略胜于HBase方案,在数据量增大时,Sesame方案的查询响应时间快速上升,与本文方案的差距越来越大。以上对比可以看出,本文基于HBase的方案能够支撑大数据量的本体存储需求。

5 结束语

本文针对民航应急管理领域本体设计了一种基于HBase的领域本体存储模型,实现了领域本体类及属性信息的元数据和RDF实例数据的存储,并实现了在此模型上基于HBase API的基本图模式查询算法,通过实验分析了存储与查询方案的优势,为大数据环境下的民航应急管理提供了方法支持,为基于HBase的领域本体推理等应用奠定了良好的数据基础。

参考文献:

[1] George L. HBase: The definitive guide[M]. Sebastopol: O'Reilly Media, Incorporated, 2011.

[2] Papailiou N, Tsoumakos D, Konstantinou I, et al. H2RDF+: An efficient data management system for big rdf graph[C]// Proc of Semantic Web Information Management, SWIM 2014.

[3] Sun J, Jin Q. Scalable rdf store based on hbase and mapreduce [C]// Proc of 2010 the 3rd International Conference on Advanced Computer Theory and Engineering, 2010: 633-636.

[4] Zhu Min, Cheng Jia, Bai Wen-yang. A storage model For RDF data based on HBase[J]. Computer Research and Development, 2013, 50(Suppl.): 23-31. (in Chinese)

[5] Wang Hong, Yang Xuan, Wang Jing, et al. Research on ontology-based knowledge presentation and reasoning in civil aviation emergency decision[J]. Computer Engineering & Science, 2011, 33(4): 129-133. (in Chinese)

[6] Wang Hong, Zhu Yue-li, Wang Jing, et al. The applied research of the method in ontology mapping based on the relational mode [J]. Journal of Convergence Information Technology, 2013, 8(11): 292-302.

[7] Li Shan-pin, Yin Qi, Hu Yu-jie, et al. Overview of researches on ontology[J]. Computer Research and Development, 2004, 41(7): 1041-1052. (in Chinese)

[8] Li Man, Wang Da-zhi, Du Xiao-yong, et al. Dynamic composition of web services based on domain ontology[J]. Chinese Journal of Computers, 2005, 28(4): 644-650. (in Chinese)

[9] Resource description framework (RDF)[EB/OL]. [2014-02-25]. http://www.w3.org/RDF/.

[10] Web ontology language (OWL)[EB/OL]. [2012-12-11]. http://www.w3.org/2001/sw/wiki/OWL.

[11] Zou Lei, Chen Yue-guo. Massive RDF data management [J]. Communications of the CCF, 2012, 11(8): 32-43. (in Chinese)

[12] Baeza-Yates R, Castillo C, Junqueira F, et al. Challenge on distributed web retrieval[C]// Proc of IEEE the 23rd International Conference on Data Engineering, 2007: 6-20.

[13] Hadoop: The definitive guide [M]. Sebastopol: O'Reilly Media, Incorporated, 2012.

[14] Du Fang, Chen Yue-guo, Du Xiao-yong. Survey of RDF Query Processing Techniques [J], Journal of Software, 2013, 24(6): 1222-1242. (in Chinese)

附中文参考文献:

[4] 朱敏,程佳,柏文阳. 一种基于 Hbase 的 RDF 数据存储模型[J]. 计算机研究与发展,2013,50(Suppl.):23-31.

[5] 王红,杨璇,王静,等. 基于本体的民航应急决策知识表达与推理方法研究[J]. 计算机工程与科学,2011,33(4):129-133.

[7] 李善平,尹奇,胡玉杰,等. 本体论研究综述[J]. 计算机研究与发展,2004,41(7):1041-1052.

[8] 李曼,王大治,杜小勇. 基于领域本体的 Web 服务动态组合[J]. 计算机学报,2005,28(4):644-650.

[11] 邹磊,陈跃国. 海量 RDF 数据管理[J]. 中国计算机学会通讯,2012,11(8):32-43.

[14] 杜方,陈跃国,杜小勇. RDF 数据查询技术综述[J]. 软件学报,2013,24(6):1222-1242.

作者简介:

王红(1963-),女,重庆人,教授,CCF 会员(E200014403M),研究方向为本体技术、数据挖掘与智能信息处理。**E-mail**: hwang@cauc.edu.cn

WANG Hong, born in 1963, professor, CCF member (E200014403M), her research interests include ontology, data mining, and intelligent information processing.

孙康(1991-),男,安徽淮南人,硕士生,研究方向为语义网络与本体。**E-mail**: skctvc15@163.com

SUN Kang, born in 1991, MS candidate, his research interests include semantic web and ontology.