DOI:10.19751/j.cnki.61－1149/p.2021.04.022

非结构化地质数据内容存储方法研究

魏东琦[1,2],江宝得[1,3*],张静雅[2]

(1. 中国地质大学(武汉)国家地理信息系统工程技术研究中心,湖北 武汉 430074;2. 中国地质调查局西安地质调查中心,陕西 西安 710054;3. 地理信息工程国家重点实验室,陕西 西安 710054)

摘 要:地质工作已迈入大数据时代,但地学信息被记录成的报告、图件等非结构化数据,仍按照较为简单的方式组织归类到一起并存储在文件系统中,形成很多个内部构成复杂的数据集。这种方式不能很好的表达非结构化数据承载的丰富地学信息,也不便表达信息之间的复杂关系,更不利于发现跨数据集存在的深层知识。为尝试解决这个问题,笔者提出了多粒度级别内容树模型和支持演化的数据建模方式。这些特性使得通过模型可以对数据内容进行不同尺度的拆分,对信息的精确定位,还可以使模型根据数据主体需要,拓展主体特征描述的维度,逐步发现数据包含的信息和建立信息与信息之间的关系。考虑到地质大数据的特点,设计了以 HBase 为核心的数据模型持久化方式,以达到使用大数据技术体系下技术分析处理数据的目的;最后给出了对成果地质数据进行建模的实例,将文档、图件等非结构化数据以内容实体为最小单元进行拆分和重构,达到了较好的内容组织和信息表达效果。

关键词:地质大数据;非结构化数据;数据建模;内容存储

中图分类号:P628 **文献标志码**:A **文章编号**:1009-6248(2021)04-0266-08

Research on Content Storage Method of Unstructured Geological Data

WEI Dongqi[1,2], JIANG Baode[1,3*], ZHANG Jingya[2]

(1. National Engineering Research Center of Geographic Information System, China University of Geosciences (Wuhan), Wuhan 430074, Hubei, China; 2. Xi'an Center of China Geological Survey, Xi'an 710054, Shaanxi, China; 3. State Key Laboratory of Geo-information Engineering, Xi'an 710054, Shaanxi, China)

Abstract: Geological work has entered the era of big data, yet the unstructured data, such as reports and maps carrying geosciences information, are still classified in simple ways and stored in the file system, forming a lot of data set with complex internal structures. This method cannot well deliver the abundant geosciences information carried by unstructured data or the complex relationships with information, nor can it discover the knowledge deeply existing across data sets. To solve the problem, this paper proposes a multi-granularity level content tree model and a data

收稿日期:2021-04-15;**修回日期**:2021-05-24

基金项目:中国地质调查项目"国家地质大数据汇聚与管理"(202009000000180722),地理信息工程国家重点实验室基金资助项目、实验室开放基金(SKLGIE2019－Z－4－1)。

作者简介:魏东琦(1983－),男,博士研究生,高级工程师,主要研究方向为地质大数据、数据挖掘、自然语言处理。E－mail:wdongqi@mail.cgs.gov.cn。

*** 通讯作者**:江宝得(1982－),男,博士,助理研究员,主要研究方向为空间数据分析、多尺度表达等。E－mail:jiangbaode@cug.edu.cn。

modeling method that supports evolution. The model can split the data content at different scales and accurately locate the information and meanwhile expand the dimension of the subject's feature description according to the need of the data subject. The information contained in the data is finally discovered and the relationship with information is thus established. This paper designs a persistence method of data model with HBase as the core to achieve the purpose of processing data under the big data technology system. A modeling example shows preferable effect in content organization and information conveying, with the unstructured data of documents and maps split and reconstructed as the smallest unit of the content entity.

Keywords: geological big data; unstructured data; data model; content storage

地质工作已迈入大数据时代,这为挖掘数据内在信息,充分发挥数据自身的价值带来了良好契机(赵鹏大,2018)。大数据是一种高级信息生产力,它促进着信息生产方式的改变,并推动数据应用模式的发展。正如麦肯锡所说的:"数据已经渗透到当今每一个行业和业务职能领域,成为重要的生产因素。人们对于海量数据的挖掘和运用,预示着新一波生产率增长浪潮的到来。"大数据技术虽然起源于互联网行业,但随着它的成功应用和展示出的强大活力,也对诸多传统行业的信息产业发展带来了启示和不容错过的机遇。

在大数据时代到来之前的很长一段时间,地质领域一直采用一些传统的数据(这里指数字化的数据)生产方式,这种记录和存储数据的方式不够先进,但因为简单被大多数工作者所使用。这些传统的生产方式积累了十分丰富的成果数据,其中蕴藏的巨大的价值有待发现(李超岭,2015;陈建平,2017)。在数据的形成过程中,人们往往习惯将数据按照一定的方式组织归类。众所周知,在地学领域一般面对的问题比较复杂,成果结论需要多种方式结合才能很好的表现。这样形式多样的数据被人为的组织到一起,构成一个数据集,用于说明某个问题,表述某种结论等等。这个数据集中数据内容之间、数据之间具有某种人为形成的构成关系。这种数据集合的数量十分巨大,数据集中的数据以非结构化数据为主。每个数据集合中,数据内容组织灵活,信息表现形式多样,数据有其内在的规律但无法进行严格的模式约束。数据中多种价值信息交织在一起,信息总量大但信息点分散,难于梳理提炼和归纳。

笔者研究对象就是具有此类特征的数据集,简称为非结构化地质数据集。它们是行业中的成果数据,而不是原始采集的数据。生产者已对数据进行了一定程度的加工,并且按照一种合理的方式对信息进行了组织和梳理,但由于使用的工具和认识水平的限制,使得数据的存在形式较为单一,信息之间复杂的关联关系不能有效表达,不能方便的发掘数据内在的价值。从另一角度上讲,人们希望借助一定的技术方法,挖掘数据内在的信息,发现和理解信息内容之间的关系,体现数据具有的价值(Ashley, et al., 2014)。从现阶段的技术发展水平来看,大数据技术应该能够较好的满足这方面的需求。

从上面的分析可以看出,这类数据的存在形式与所需的技术产生了代差,为解决这个问题会涉及到数据管理的很多方面,但建立数据合理有效的组织和存储形式是重要环节,也是后续很多工作的基础。

1　相关工作

在对非结构化数据的管理和处理方面,主要运用数据仓库、内容管理系统等手段,传统意义上,这些系统对数据的持久化组织存储主要依托关系数据库这大类技术体系。王珊等(2011)较为全面的分析了使用关系型数据库的优劣,并指出在大数据的时代背景下,随着新技术新方法的产生,带来了非结构化数据管理与处理方面自上而下的变革(覃雄派等,2013)。许多研究表明,NoSQL 技术正逐渐成为非结构化数据管理的优势性技术,相对于传统的数据操作方式(王梅等,2013),大数据技术体系下"靠近数据计算"的设计理念也更适用于大规模非结构化数据的深度分析和知识发现(Cuzzocrea, et al., 2011;吴冲龙等,2016)。

一般情况下一个完整的体系,在数据组织管理

和处理每个层面都有各自的策略和技术手段，在笔者关注的数据的组织层面，这类策略主要是针对数据的建模。在这个问题上，前人已经在理论研究和技术实现上做了很多工作，提出了解决方案，具有代表性的有数据空间（dataspaces）（Franklin et al.，2005）以及 NoSQL 相关的一大类技术（谢华成等，2012；杨鹏等，2018）。笔者将从一个新的角度看待一个地质行业使用传统方式生产并积累下来的大量成果数据，研究大型内容复杂数据集的内容分解提取和组织存储的技术方法，建立起数据与大数据技术之间的桥梁。借助大数据技术的优势，提高数据的信息定位，相关关系建立等方面的效率，让数据的深度分析，知识发现变得更加顺畅，赋予此种地质数据“大数据 computing 的能力”。

2 数据模型

内容复杂、信息存储形式多样的大型非结构化数据集，当中的数据很少能用预先定义的模式进行组织，更为合理的组织方式往往是在收集和处理数据的过程中产生的。再者，由于蕴含在内容复杂数据集中的信息并非以传统的关系数据库方式严格表示，而是大多包含在非结构化的数据中。因此，建立基于内容和特征的数据模型是进行有效描述数据的关键，同时也是进一步知识发现的实现基础。

建立基于内容和特征的数据模型目的在于消除由于内容表现形式多样造成的异构和多源化问题，在不丢失信息量的原则上尽量自然的组织和描述数据的真实内容，使得计算分析能更容易的贴近数据描述的本质，更容易发现数据蕴含的知识。笔者利用模型力求对数据的内在作语义清晰、组织高效的描述性和结构性建模，以便进行有效的知识集成、共享和复用。

2.1 基本定义

定义 1。特征域（值域）D_i 是具有相同类型的值的集合，且这个集合有限集，集合中元素的个数称为域的基数记作 $\|D_i\|=m$。特征域的全集 $D=\bigcup_{i=1}^{N} D_i$，其中任意两个 $D_i \subset D, D_j \subset D, D_i \cap D_j = \varphi$，即 D 是多个不相交的特征域的并集，每个域都存在若干支持的操作符是封闭的。

定义 2。主题特征量 $F=\{f_i \mid \forall f_i = Id(D_i), 0 \leqslant i \leqslant n\}$，其中 f_i 表示一个维度的主题特征分量，它的取值为某一特征域的标识 $Id(D_i)$。F 是一个 n 维特征量，但对于任意的 f_i 其仅可映射到唯一的 D_i，不存在不同 f_i, f_j 取值于同一个特征域 D_i。

由定义 1 和 2 可知，子域的个数等于内容实体中特征量的维数。不同特征量 F_m，F_n 中的值可以映射到于同一个域 D_x，即 $F_m(i)=F_n(j)=Id(D_x)$，称其为享元（flyweight）（这是内容实体聚合、分类的一种方式）。

定义 3。取 D 中的有限子集 $\{D_1, D_2, \cdots, D_n\}$，$D_1, D_2, \cdots, D_n$ 的笛卡尔积为：$D_1 \times D_2 \times \cdots \times D_n = \{< f_1, f_2, \cdots, f_n > \mid f_i \in D_i, i=1,2,\cdots,n\}$。其中，每一个 $(f_1, f_2, \cdots, f_n)$ 叫做一个 n 维特征量，每一个值 f_i 叫做一个分量。这些域中可以存在相同的域。

若域 D_i 的基数 $\|D_i\|=m_i, i=1,2,\cdots,n$，则笛卡尔积的基数为：$M=\prod_{i=1}^{n} m_i$。

定义 4。内容实体（content entity）$C=(K, V, R_c, F, D_c)$ 是一个五元组，其中：键 K 是全局唯一 C 的标识，V 是 C 表示对象化的原始内容（可以是文本对象、空间对象、图像等），K, V 构成一一映射关系；R_c 为 C 与其他项（可以是多个）C' 的联系（relationship），是面向对象的，包括泛化（Generalization）、关联（Association）、聚合（Aggregation）、组合（Composition）、依赖（Dependency）5 种基本关系，联系可以是一对一的和一对多的，但不能自引用；F 为 C 的主题特征量，对不同的内容实体 C_i 和 C_j，特征量的维数可以不同，并且属于 D 的笛卡尔积构成的集合；D_c 是 F 中特征分量（特征值）取值的域，D_c 是 D 的一个子域 $D_c \subseteq D$。

具有相同取值域 D_c 的内容实体的特征量可以组成一个包。集合的元素是无序的，并且元素是不同重复的。包中的元素是无序的，但允许一个元素出现多次。内容实体的组成及构成关系如图 1 所示。

2.2 支持演化的非结构化数据的内容组织

建立数据模型目的是让模型对数据的结构组织和特征描述更加贴近于数据的本质，建模是一个过程，所以数据模型需要通过演化一步步清晰数据之间的关系，更贴切的描述主体的特征。数据模型可分为初始态和演化过程 2 个部分：初始态是模型对数据最初的理解和描述，建立数据内容的基本的定

位方式和相关关系,是模型进化的基石;演化是数据模型的一个重要特性,指的是模型可以随着对所表示的主体内容的理解,随着时间和应用的变化不断完善自身。

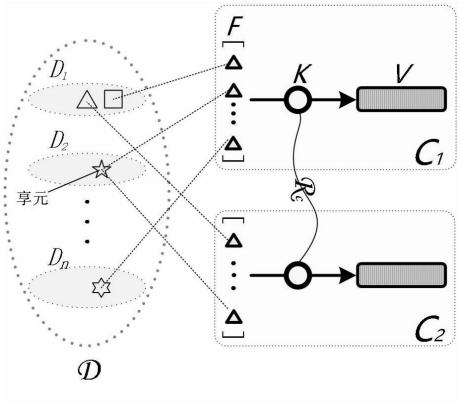

图 1　内容实体的组成要素及它们之间的关系

Fig. 1　Components of the content entity and their relationship

笔者使用树形结构表示模型初始态的骨架,这样可以较好的描述复杂内容数据集中数据诸多要素之间的关系,建立其最基本的坐标定位。为此,使用内容实体 C 来表示内容树的节点,同时将 C 退化出一种特殊的 $C'=(k,R_c,F)$ 用来表示一种纯结构性的节点(后面提到的数据集和可再分解的数据实体都属于此种节点)。于此同时,使用项间关系 R_c 表示树枝,则 R_c 被具体化为 2 种关系{R-parent,[R-children]},即一个节点的父亲节点和多个孩子节点。在此需要引入另一个概念。定义 5。路径 $Ph=\{c_i,c_{i+1},\cdots,c_{i+k}\}$ 是一个非空的有序集合,c_i 是树 T 的任意一个节点。对任意 c_i,c_j 存在且仅存在一条关联路径 Ph。特别的,定义从根结点到任意内容节点的路径称为 c_i 的根路径,记为:$Ph_{root}(c_i)$。

路径是一种可行的内容树节点间原始的关联关系表示和位置坐标的定义,其将原本独立存在的信息联系起来,形成了一个可供操作的整体。

由此,内容组织模型是具有层次和特征继承关系的树 T(图 2)。T 存在 root-set、data-set、data-entity 和 content-entity 三个层次的划分,节点的深度越深描述粒度越细。

root-set 是逻辑上的根节点,他的孩子由大量的类型多样、内容结构复杂的 data-set 组成,记为如下。

$Rs=\{Ds_0,Ds_1,\cdots Ds_n\}$。

由于 root-set 具有超大数量的子节点,因此在实现上需要分布式的非关系型的数据库技术作为支持(Chang, et al.,2006)。

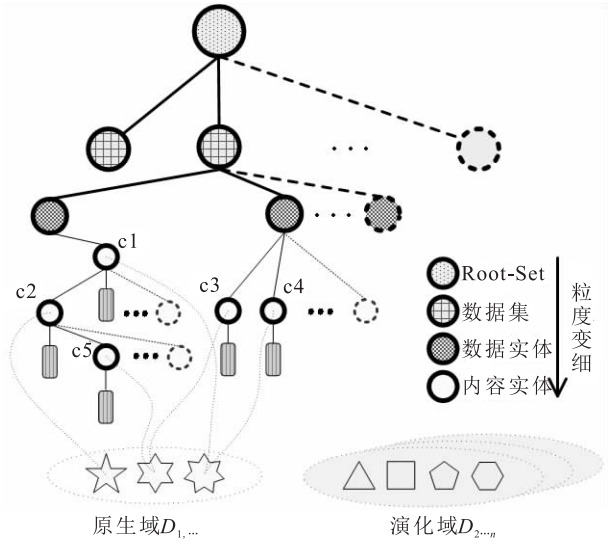

图 2　内容树模型

Fig. 2　Content tree model

数据集 Ds 由有限类型的数据实体 De 组成,记为如下。

$Ds=\{\{De_i^{\alpha}\mid 0\leqslant i\leqslant m\},\{De_j^{\beta}\mid 0\leqslant j\leqslant n\},\cdots,\{De_k^{\gamma}\mid 0\leqslant j\leqslant l\}\}$

De 用于表示某特定数据集中包含数据实体的种类(上标)和数量(下标),例如,a 数据集包含 1 篇正文报告、4 张附图、1 篇审批报告、3 张简单的统计表等等。每个种类的数据会有专门的策略去解析数据中包含的内容。

数据实体是内容实体组合成的复合体。具体来讲,De 是可用面向对象思想建模的由内容实体构成的复杂对象。

内容实体好比是原子,内容实体好比是分子,数据集合就好比是由他们构成的物质。数据实体之间、数据内部的内容之间存在某些联系,两两之间的原有结构性关系可以使用路径描述。

上述提到,数据模型基本的结构性特征使用树状骨架建模,并使用路径的概念定义了节点在整体中的位置坐标,组织起内容节点之间的关联关系。然而仅是数据建模的开端,更深层次的特征刻画需要使用特征量和特征域完成,特征量的概念见定义 2,其取值于相应的特征域,特征域在此分为原生域

和演化域 2 种。在此规定,原生特征项取值于原生域,其存储的是域值的原值;演化特征项取值于演化域,其存储的是域值的引用(reference)。

原生域:原生域是一种 schema-first 的逻辑结构,其有预先定义的数据模式,并且这种模式在模型的生命周期中保持不变。域值的产生严格依赖数据模式,构建方式遵循 pay-before-you-go(Franklin, et al.,2005)的原则。原生域的元素数据类型简单,取值严格准确,域值之间无关联性。

演化域:其随着对内容的理解逐步加入模型,具有先有数据后有模式的 pay-as-you-go(Franklin, et al.,2005)的演化特性。域中元素是基于对象的,元素之间可以构建复杂的关系结构。演化域也是对结构性特征和原生描述性特征表现力有限的一种补充。也就是说,对于任意一个演化域,其中的元素可以是结构性或者是描述性的。

取自原生域的特征值在内容树结构的构建过程中被初始化。演化是内容树模型的重要特点,取自演化域的特征值不做任何的假设和限定,其数据模式可以是松散的滞后的,数据模式是在数据的基础上根据内容主体的需求逐渐演化而来的。演化域元素是基于对象的,因此元素之间的关联关系可以是复杂的,并且也是根据主体需要动态建立的,在其上定义的数据操作也是 best-effort 的,即允许次优的结果产生(李玉坤等,2008)。

2.3 内容存储的持久化方式

采用多尺度、细粒度的内容拆分有利于数据的深度分析和信息的精确定位,但也因此带来了内容片段数量的几何级数式增长,数据的存储和管理面临新的挑战,NoSQL 技术的兴起为解决此类问题提供了良好的方法。更重要的是,数据模型的特征量和特征域的模式滞后的特性导致其更倾向使用 NoSQL 技术来实现。因此,使用面向列簇的存储技术——HBase 是一个较为理想的方案(王梅等,2013; Chang, et al., 2006; 谢华成等, 2012)。HBase 表设计中,行键的设计至关重要,由定义 3 可知,每一个内容节点都有一个唯一根路径可作为行键(rowkey),行键的构成采用如下策略。

(1)递归的将同一父节点下的子孙节点按深度为权值升序存放在连续的行中,这一点利用 $Ph_{root}(c_i)$ 的排序很容易做到。

(2)如果数据项存在或继承有原生域的值,则将其编码并附加到行健最后,否则占位补齐。

(3)保证每个 rowkey 的长度是相等的,并且需要进行散列处理。由于面对的问题是大型的数据集,所以选用低碰撞概率的散列算法(Biham, et al.,2006)。

(4)为读操作优化行键,以便他们可以被快速的读取。表结构如图 3 所示,行键是根路径哈希和原生特征编码的组合,表中包含 2 个列簇,proto 用于存储内容实体的 value、关系和来自原生域特征值;Evolvement 存储取值自演化域的特征项的值的引用。时间戳简单起到版本号的作用。演化域内元素之间关系是域内部表示范畴,属于另外的研究领域。

rowkey	Proto					Evolvement			
	Value:content	Relation	D0	D1	Dk	D_{k+1}	D_{k+2}	...	
Ph_{root}:hash(ci)+pf	V(ci):byte[]	$\{R_p, [R_{chd}]\}$:ci	e0	e2	ei	ref:ex	Φ		
Ph_{root}:hash(cj)+pf	V(cj):byte[]	$\{R_p, [R_{chd}]\}$:cj	e3	e2	Φ	Φ	ref:ey		

图 3 内容树模型在 Hbase 中的存储结构

Fig. 3 Storage structure of content tree model in HBase

2.4 数据建模与模型演化的主要思想

数据建模是将原始状态的数据集中的内容解析重构成内容树模型的过程,建模分为 2 个独立的步骤。步骤一,建立模型的初始态,包括内容树的骨架结构和原生特征项的赋值,由于内容树的继承特性,下一级别的节点可以选择性的继承其祖先节点的特征;步骤二,为模型的演化过程,这是通过对节点内值的分析修正原生特征值和增加演化特征值,从而

丰富节点的描述性特征和附加结构性特征。

演化操作是由已知信息得到潜在信息的过程,对于内容实体个体,这个操作的结果将以特性值的形式存在;而对于总体的内容实体集合,一种演化操作的所有特征值构成一个演化域。前述提到演化域随着对内容的理解逐步加入模型,即演化操作事先知道演化域的取值范围(包括域值间的关系),演化操作则按照一定的模式进行内容实体与域值之间的匹配。

具体来说,针对内容树模型,演化操作的对象是内容树的节点,已知信息是模型初始过程后形成的一些,可以表示成($Ph_{root}(c_i)$, $V(c_i)$, $\{R_{parent}(c_i), [R_{child}(c_i)]\}$, F'),其中 F'是原生域,演化操作 f 的结果是一个新加入的演化域 Dx 和一个新维度的特征值 e_x, $F'=F'+\{e_x\}$。根据前述数据结构,演化操作可以使用 map-reduce 计算框架实现(Dean, et al.,2004)。这种考虑主要是因为作为操作对象的内容节点的数量级十分巨大,而彼此之间的数据结构相对独立。

3　成果地质数据建模实例

对成果地质数据而言,内容模型是能够突破档案的界限约束,突破文件边界的限制,将原先孤立于每个档案中,每份数据中的信息统一整合在一个模型中,增强信息之间的联系,简化获取信息的步骤,丰富获取知识的途径。从图 4 可以看到,模型对数据的描述粒度逐渐细分,首先是档案级目录元数据,包括档案标题、档号、项目编号等元信息;再者是文件级描述,包括空间要素,文档目录,附图、附表、附图等索引表;最后是将数据实体分解后的章节段落,插图、表格、图件要素等内容实体。这种金字塔形状的内容粒度划分,前面几级是已经事实存在的数据和数据固有的组织形式,最后的内容级是此模型的主要内容,也是最小粒度级的操作对象。

参照前述数据模型的有关概念和具体化成果地质数据,表 1 将模型中的概念与现在需要建模的归档数据层级关系和数据构成做了对应。

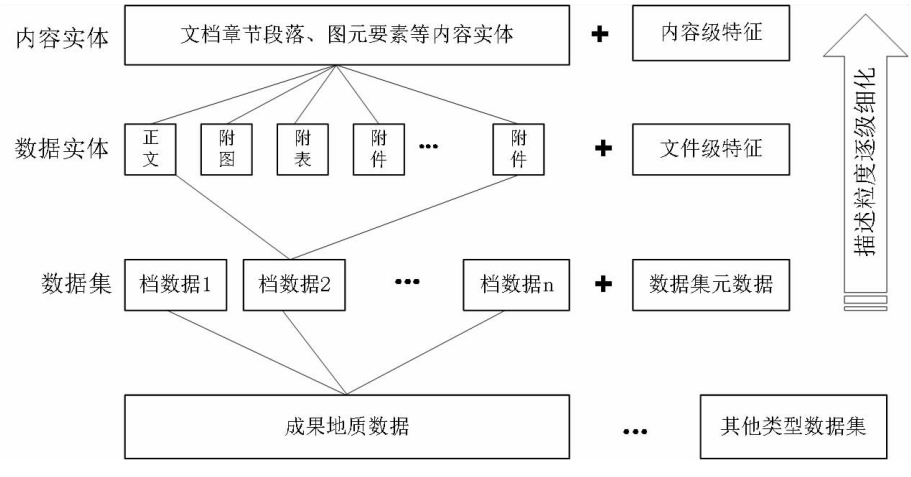

图 4　地质成果数据的分级组织

Fig. 4　Hierarchical organization of geological achievement data

表 1　成果地质数据和数据模型概念间的对应表

Tab. 1　Correspondence table between achievement geological data and data model concepts

数据模型中的概念	成果地质数据	示例
root-set	成果数据的集合	存放成果数据的根目录
data-set	档数据(目录级)	某一档成果数据
数据实体(data-entity)	每档数据中的文件(文件级)	正文、附图、附表等文件
内容实体(content-entity)	文件的内容及关系(章节结构,段落、图元要素等)	文档的章节段落及文档结构,图件的图元要素

由于笔者目前只对成果地质数据的建模，所以根 Root－Set 中只包含一种类型的数据集，也就是成果地质数据集。每一个数据集 Data－Set 表示一档地质成果数据，例如，如果有 5 000 档地质成果归档数据，那么 Root－Set 就包含 5 000 个 Data－Set，可表示成如下形式。

$$Rs = \{Ds_0, Ds_1, \cdots Ds_{5000}\}。$$

如果按照树状结构表示 Root－set 和 Data－set 的关系，Root－set 可能会有大量的子节点，这在实际的数据结构实现中会做优化处理的，不可能会让一个节点存在太多的子节点，但仅从逻辑视角观察，暂且可以认为所有的 Data－set 全部是 Root－set 的直接孩子节点。

段落、表格、图片、矢量图等每一个类内容实体都将被模型合理的表达，根据前述的定义，数据实体是内容实体组合成的复合体。根据成果地质资料的汇交规范，按数据类型的不同可分为正文、附图、附表、附件、审批、其他 6 类，这样成果地质数据的数据集将包含一个合并分卷后的正文数据实体（Z）、零到多个附图数据实体（T）、零到多个附表数据实体（B）、零到多个附件数据实体（J）、零到多个审批数据实体（S）、零到多个其他类数据实体（Q）。按照前述关于数据集的定义，可以将数据集形式化表示如下形式。

$$Ds = \{\{De_1^Z\}, \{De_{0*n}^T\}, \{De_{0*n}^B\}, \{De_{0*n}^J\}, \{De_{0*n}^S\}, \{De_{0*n}^Q\}\}。$$

为更直观说明建模实例，笔者采用 UML 模型形式化地表示地质成果数据的建模结果（图 5）。

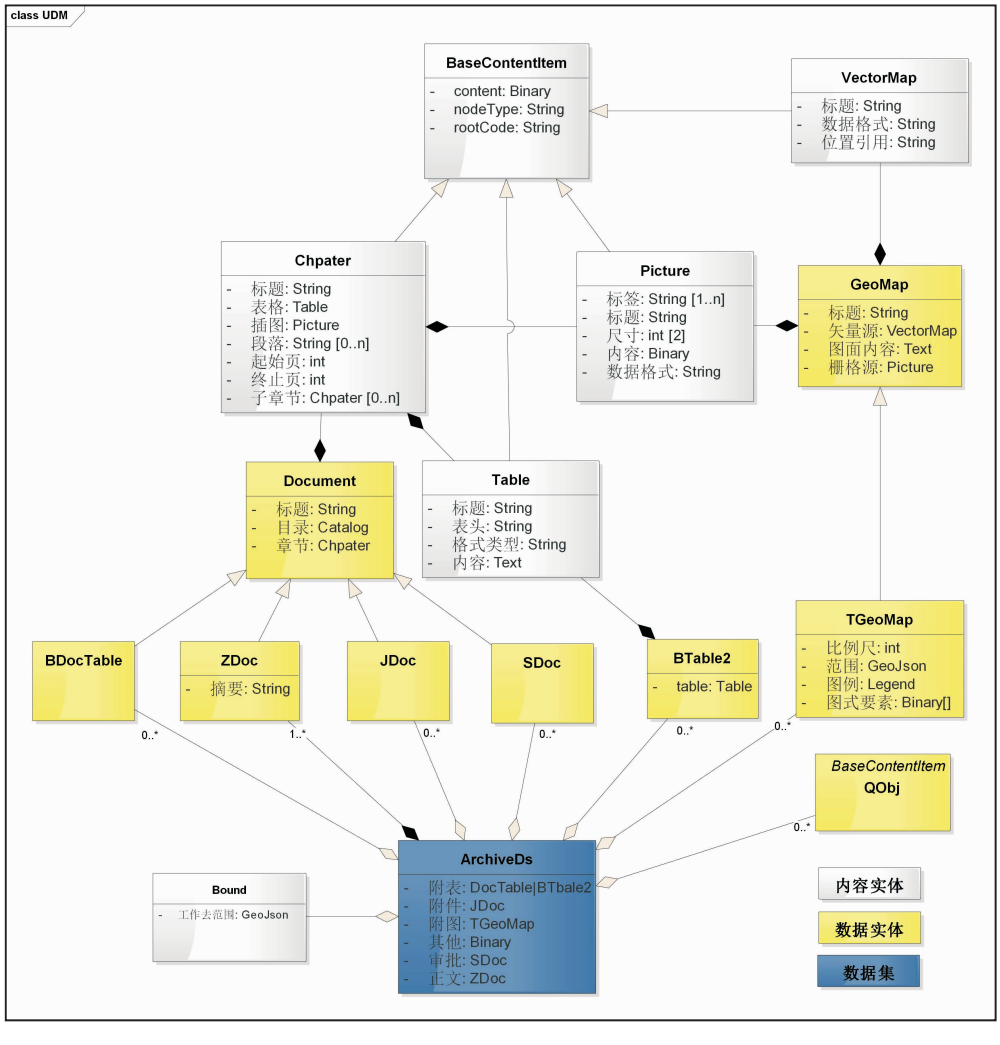

图 5 成果地质数据的建模图

Fig. 5 The modeling of the achievement geological data

在地学领域,由于问题的复杂性,人们关注“是什么”多过“为什么”。也就是说地学领域关心信息与信息之间的相关关系,而不是因果关系。例如,某种“区域范围”内会产生某些“矿产”,就是地质学家经常研究的问题。在历史积累下来的大量地质成果数据中记录有这 2 种内容的描述,将同一主题信息的内容从多样化的数据中分离并合理的汇聚是进行知识发现的基础。使用笔者数据模型和建模方法,可以将文档、图件等非结构化数据以内容实体为最小单元进行有效组织,再通过特征的演化方式逐步完善模型对数据的整体理解,以“大数据”的方式去分析计算、挖掘其内部的价值。

4　结束语

笔者通过研究非结构化地质数据集的内容组织和存储,详细论述了支持演化的数据模型的建立方法。这种方法的优势在于可以将格式不同的数据以一种自然合理的方式组织到一起,还可以通过特征的演化方式逐步完善模型对数据的整体理解。这种设计能够使非结构化地质数据以“大数据”的方式去分析计算,挖掘其内部的价值。这为以传统方式生产的数据的有效管理和知识发现提供了一个可行的方案。以此为基础,可建立基于大数据技术的地学信息深度分析与知识发现利用平台,构建地质成果数据的价值链;建设地学信息资源共享与增值示范服务,为全社会的信息整合、资源共享、知识创新创造有利的条件。

参考文献(References):

赵鹏大. 地质大数据特点及其合理开发利用[J]. 地学前缘,2019,26(4):1-5.

ZHAO P D. Characteristics and Rational Utilization of Geological Big Data[J]. Earth Science Frontiers, 2019,26(4):1-5.

陈建平,李靖,谢帅,等. 中国地质大数据研究现状[J]. 地质学刊,2017,41(03):353-366.

CHEN J P,LI J,XIE S,et al. China Geological Big Data Research Status[J]. Journal of Geology, 2017,41(03):353-366.

李超岭,李健强,张宏春,等. 智能地质调查大数据应用体系架构与关键技术[J]. 地质通报,2015,34(07):1288-1299.

LI C L, LI J Q,ZHANG H C, et al. Big Data Application Architecture and Key Technologies of Intelligent Geological Survey[J]. Geological Bulletin of China, 2015, 34(07):1288-1299.

王珊,王会举,覃雄派,等. 架构大数据:挑战、现状与展望[J]. 计算机学报,2011,34(10):1741-1752.

WANG S,WANG H J,QIN X P,et al. Architecting Big Data:Challenges, Studies and Forecasts[J]. Chinese Journal of Computers, 2011,34(10):1741-1752.

覃雄派,王会举,李芙蓉,等. 数据管理技术的新格局[J]. 软件学报,2013,24(02):175-197.

QIN X P, WANG H J, LI F R, et al. New Landscape of Data Management Technologies[J]. Journal of Software, 2013,24(2):175-197

王梅,周娇玲,乐嘉锦. 一种列存储数据仓库中的数据复用策略[J]. 计算机学报,2013,36(08):1626-1635.

WANG M,ZHOU J L,LE J J. A Data Reusing Strategy in Column-Store Data Warehouse[J]. Chinese Journal of Computers, 2013,36(08):1626-1635.

吴冲龙,刘刚,张夏林. 地质科学大数据及其利用的若干问题探讨[J]. 科学通报, 2016, 61(16):1797-1807.

WU C L,LIU G,ZHANG X L. Discussion on Geological Science Big Data and its Applications[J]. Chinese Science Bulletin,2016,61(16):1797-1807.

杨鹏,林俊晖. 一种基于 MongoDB 和 Hadoop 的海量非结构化物联网数据处理方案[J]. 微电子学与计算机,2018,35(04):68-72+78.

YANG P,LIN J H. A Scheme for Massive Unstructured Iot Data Processing Based on MongoDB and Hadoop[J]. Microelectronics & Computer,2018,35(04):68-72+78.

谢华成,陈向东. 面向云存储的非结构化数据存取[J]. 计算机应用,2012,32(07):1924-1928+1942.

XIE H C,CHEN X D. Cloud storage-oriented unstructured data storage[J]. Journal of Computer Applications, 2012, 32(07): 1924-1928+1942.

李玉坤,孟小峰,张相於. 数据空间技术研究[J]. 软件学报,2008(08):2018-2031.

LI Y K, MENG X F,ZHANG X Y. Research on Dataspace [J]. Journal of Software,,2008(08):2018-2031.

Biham E, Chen R, Joux A, et al. Collisions in SHA-0 and Reduced SHA-1[M]. Springer Berlin Heidelberg, 2005.

Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 2004, 51(1): 137-150.

Ashley I. Naimi, Daniel J. Westreich. Big Data: A Revolution That Will Transform How We Live, Work, and Think [J]. American Journal of Epidemiology, 2014, 179(9) Pages 1143-1144.

Cuzzocrea A, Song I Y, Davis K C. Analytics over Large-scale Multidimensional Data: the Big Data Revolution[A]// International Workshop on Dolap[C]. ACM,2011,101-104.

Franklin M, Halevy A, Maier D. From Databases to Dataspaces: A New Abstraction for Information Management[J]. Sigmod Record: Acm Sigmod (management of data),2005,34(4):27-33.

Chang F, Dean J, Ghemawat S, et al. Bigtable: A Distributed Storage System for Structured Data [J]. Acm Transactions on Computer Systems, 2008, 26(2):1-26.