

6-4-2020

Cloud Manufacturing Resource Discovery Method Based on EAV Model and Solr

Geng Chao

1. School of Management, Harbin Institute of Technology, Harbin, 150001;;

Shiyu Qu

1. School of Management, Harbin Institute of Technology, Harbin, 150001;;

Yingying Xiao

2. Beijing Complex Product Advanced Manufacturing Engineering Research Center, Beijing Simulation Center, Beijing 100854;;3. State Key Laboratory of Intelligent Manufacturing System Technology, Beijing Institute of Electronic System Engineering, Beijing 100854;;4. Science and Technology on Space System Simulation Laboratory, Beijing Simulation Center, Beijing 100854;;

Guoqiang Shi

2. Beijing Complex Product Advanced Manufacturing Engineering Research Center, Beijing Simulation Center, Beijing 100854;;3. State Key Laboratory of Intelligent Manufacturing System Technology, Beijing Institute of Electronic System Engineering, Beijing 100854;;4. Science and Technology on Space System Simulation Laboratory, Beijing Simulation Center, Beijing 100854;;

See next page for additional authors

Follow this and additional works at: <https://dc-china-simulation.researchcommons.org/journal>



Part of the [Artificial Intelligence and Robotics Commons](#), [Computer Engineering Commons](#), [Numerical Analysis and Scientific Computing Commons](#), [Operations Research](#), [Systems Engineering and Industrial Engineering Commons](#), and the [Systems Science Commons](#)

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

Cloud Manufacturing Resource Discovery Method Based on EAV Model and Solr

Abstract

Abstract: A resource discovery solution of cloud manufacturing based on text information processing technology was proposed. This solution converted the description model of cloud manufacturing resource to a formal model of text information processing via a mapping function, and *an EAV (Entity-Attribute-Value) derived model of cloud manufacturing resources was designed to solve the resource properties issues of heterogeneity, diversity, dynamic and so on. By introducing the Solr index schema, the text indexing system of resource properties data was built to improve query performance of related data. A prototype system was designed for discovering cloud manufacturing resources based on EAV model and Solr search engine, which applied to the cloud manufacturing platform, improving the discovery and matching capabilities of manufacturing resource.*

Keywords

EAV (Entity-Attribute-Value), Solr, resource discovery, cloud manufacturing

Authors

Geng Chao, Shiyou Qu, Yingying Xiao, Guoqiang Shi, Tingyu Lin, Liqin Guo, and Xiaowei Wang

Recommended Citation

Geng Chao, Qu Shiyou, Xiao Yingying, Shi Guoqiang, Lin Tingyu, Guo Liqin, Wang Xiaowei. Cloud Manufacturing Resource Discovery Method Based on EAV Model and Solr[J]. Journal of System Simulation, 2017, 29(10): 2496-2506.

基于 EAV 模型和 Solr 架构的云制造资源发现方法

耿超¹, 曲世友¹, 肖莹莹^{2,3,4}, 施国强^{2,3,4}, 林廷宇^{2,3,4}, 郭丽琴^{2,3,4}, 王啸威⁵

(1. 管理学院 哈尔滨工业大学 黑龙江 150001; 2 北京市复杂产品先进制造系统工程技术研究中心 北京仿真中心 北京 100854; 3 复杂产品智能制造系统技术国家重点实验室 北京电子工程总体研究所 北京 100854; 4 航天系统仿真重点实验室 北京仿真中心 北京 100854; 5 北京朗宥资产管理有限公司 北京 100102)

摘要: 提出了一种基于文本信息处理的云制造资源发现技术方案。该方案将云制造资源描述模型通过映射函数转化成文本信息处理中的形式化模型, 设计出了一种云制造资源的 EAV (Entity-Attribute-Value) 衍生模型, 解决了资源属性异构性、多样性、动态性等问题; 通过引入 Solr 索引架构, 实现了资源属性数据的文本索引体系构建, 提升了相关数据的查询性能; 设计实现了一个基于 EAV 模型和 Solr 搜索引擎的云制造资源发现服务原型系统应用在云制造平台上, 提升了制造资源的发现匹配能力。

关键词: EVA (Entity-Attribute-Value); Solr; 资源发现; 云制造

中图分类号: TP391.9 文献标识码: A 文章编号: 1004-731X (2017) 10-2496-11

DOI: 10.16182/j.issn1004731x.joss.201710034

Cloud Manufacturing Resource Discovery Method Based on EAV Model and Solr

Geng Chao¹, Qu Shiyu¹, Xiao Yingying^{2,3,4}, Shi Guoqiang^{2,3,4}, Lin Tingyu^{2,3,4}, Guo Liqin^{2,3,4}, Wang Xiaowei⁵

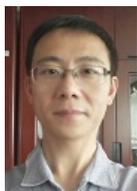
(1 School of Management, Harbin Institute of Technology, Harbin, 150001; 2 Beijing Complex Product Advanced Manufacturing Engineering Research Center, Beijing Simulation Center, Beijing 100854; 3 State Key Laboratory of Intelligent Manufacturing System Technology, Beijing Institute of Electronic System Engineering, Beijing 100854; 4 Science and Technology on Space System Simulation Laboratory, Beijing Simulation Center, Beijing 100854; 5 Beijing Langyou Asset Management Co., Ltd. Beijing 100102)

Abstract: A resource discovery solution of cloud manufacturing based on text information processing technology was proposed. This solution converted the description model of cloud manufacturing resource to a formal model of text information processing via a mapping function, and an EAV (Entity-Attribute-Value) derived model of cloud manufacturing resources was designed to solve the resource properties issues of heterogeneity, diversity, dynamic and so on. By introducing the Solr index schema, the text indexing system of resource properties data was built to improve query performance of related data. A prototype system was designed for discovering cloud manufacturing resources based on EAV model and Solr search engine, which applied to the cloud manufacturing platform, improving the discovery and matching capabilities of manufacturing resource.

Keywords: EAV (Entity-Attribute-Value); Solr; resource discovery; cloud manufacturing

引言

当前, 关于云制造资源发现技术的研究大部分



收稿日期: 2017-07-04 修回日期: 2017-08-15;
基金项目: 国家 863 计划 (2011AA040502);
作者简介: 耿超(1979-), 男, 北京, 博士生, 研究方向为云制造, 两化融合, 互联网+制造; 曲世友(1969-), 男, 黑龙江哈尔滨, 教授, 研究方向为国家创新系统。

集中在以基于本体的制造服务资源语义描述为基础, 采用本体语言的建模方式来构建 Web 服务。如文献[1]中引用了 WSMO 对外协加工资源进行了基于本体语言的建模和描述, 再通过 WSML 对资源进行形式化描述, 最终在 WSMX 执行层中实现 Web 服务。虽然目前关于本体的研究很多, 但其在使用领域上仍具有极大的局限性, 使用者必须对某

一类型的服务资源进行特定的本体建模和属性描述, 缺乏通用性和一般性。但是, 制造资源和制造能力的种类是具有多样性的, 其属性结构更是千变万化, 很难用一种通用的描述模型去对它们进行统一定义。同时, 基于本体语言的 Web 服务在具体系统的实现上难度很大, 它不仅要求开发人员具备专业的语义实现技术, 而且对实现框架中涉及到的各种工具有所了解, 而这些工具目前都还是处在研究开发阶段, 并不具备成熟的能够适用于企业级应用的体系架构。因此, 当前工程中采用较多的是传统数据结构模型(列模型)对服务资源进行存储管理, 并通过文本匹配来实现搜索。这种文本匹配的实现途径比较简单, 是一种简单的布尔型匹配, 属于自然语言处理领域^[2]中的文本信息处理^[3]。利用中文文本信息处理作为云制造资源发现技术理论基础来实现服务资源的统一管理和发现, 仍具有以下几个问题:

(1) 服务资源如何描述。文本信息处理针对的是一些“文本信息”, 通过语言学上的特征, 将文本对象转化成计算机能够理解的形式化模型, 进而进行处理。这种形式化模型通常来讲结构比较简单, 如布尔模型、空间向量模型、概率统计模型等, 也有相对比较复杂的神经网络模型, 但这些模型在 Web 信息检索中的应用都比较少。同时, 云制造平台中的资源模型具有复杂的描述结构, 有较强的层级关系、异构性、组合性和动态性, 目前相对比较成熟的描述有基于本体的语义描述, 针对制造资源的分类信息、对象信息和属性信息建立的三维模型等^[4-5]。

(2) 服务资源的结构化存储。不管服务资源的描述模型或者形式化模型是怎样定义的, 它们最终都是以数据的形式存储在计算机中。如何存储, 以什么结构和方式存储, 很大程度上决定了数据处理的效率和获取的效率。针对云制造资源而言, 其属性的复杂结构决定了它在数据存储上不同于传统的列模型, 而是具有数学概念上的稀疏矩阵的特性(动态性和异构型), 因而需要一种更加“先进”的数据模

型对资源的描述进行转化, 从而实现数据的结构化存储。

(3) 服务资源的查询体系。云制造资源服务作为一个特殊的 Web 服务, 最终是要能够被用户检索到从而对其进行使用。在这个检索过程中, 数据的存储形式、索引机制, 计算机对用户搜索语句的理解和处理能力, 系统对匹配资源的获取和优选等, 都直接影响到检索结果的查准率和查全率这两项重要的指标。

本文在现有的文本信息处理技术基础上, 针对上述问题, 将 EAV (Entity-Attribute-Value) 数据模型和大数据 Solr 搜索技术应用到云制造服务平台中, 以提升其对制造资源的发现能力。

1 研究现状

1.1 EAV 模型研究现状

T.Ganslandt 在文献[6]中采用 EAV 数据模型来描述各种临床试验数据之间潜在的数据结构关系, 把所有数据和应用统一到一个系统中, 可以方便快速地查询某个患者的所有类型的临床试验结果数据并呈现到 HTML 页面中。Prakash M. Nadkarni 等^[7]在将耶鲁大学 SENSELAB 项目开发的多个异构神经数据库进行数据库之间去耦合性时, 发现把那些数量庞大的数据记录整合到基于 EAV 模型的数据库中进行管理会更加有效。在开源社区, 基于 PHP 语言开发的 Magento 电子商务 B2C 平台系统^[8], 使用了 EAV 模型对客户、类目、产品、订单和发票等实体对象进行管理。

总的来说, EAV 模型的优势可以归结为以下几点:

(1) 扩展性强。在传统的数据库表中, 每增加一个实体对象, 就需要增加一张表; 每增加一个属性, 就需要增加一个表字段。当实体对象增多、属性个数巨量的时候, 数据库的扩展是复杂且有限的。在 EAV 结构中, 实体对象和属性的个数可以说是“无限”的(只要表容量允许), 因为它们都是用

有限列的一张表(或几张关联表)的行记录来表示。

(2) 稀疏性数据存储的空间效率高。EAV 结构数据几乎不存在空值,不需要实体对象在生成一个实例时为某些属性值预留空间。因此对于稀疏性较强的数据而言,存储空间得到了充分的利用。

(3) 简单的 A-V 键值对格式使得数据的可访问性强。EAV 模型中,属性和属性值的保存形式,类似于一般数据结构中的键值对格式。这种独立于实体对象的数据格式更有利于第三方组件的访问,它们只需知道某个属性的唯一标识码,便可以获取相关属性值。这也正是很多大型系统中常用的数据格式,如 Windows 的注册表,浏览器的 Cookie 等。

(4) 二次开发成本低。正如 Magento 平台系统所呈现出的特点,这种数据结构使得二次开发人员不需要为每个实体对象开发特定的 ORM,只需执行几条简单的 SQL 语句即可。

1.2 云制造资源模型的特点

文献[1]中对云制造资源中的外协加工类资源的属性构成及特点进行了详细论述,可以看出其属性多而复杂,具有分散性、自治性、异构性、多样性、动态性和组合性的特点。

(1) 分散性、自治性、异构性和多样性首先在某种程度上其实是反应出了该类资源属性繁杂、数据稀疏等特点。资源来自不同的地域,在主体上具有独立性,在行为上具有自治性,类型多样同时属性构成复杂、异构问题突出等等这些特点,从数据模型的角度,用比较极端的方式阐述,便是“每种资源实体对象拥有各自独立的属性集,在这些属性集合中,根据具体的需求,每个资源实例拥有各自不同的属性子集(即有属性值的集合)”。这正是 EAV 模型所能实现的。

(2) 动态性指出了资源是不断变化的,资源的状态、属性集和属性值在时间上并非如系统初始化预设的一样,而是需要一种能够动态变化的管理机制。EAV 模型将实体对象的属性和属性值作为数据库表的一条一条可操作的行记录来进行管理,并不

是如传统的列模型那样将属性作为对象表创建时定义的一个个固定字段,因此在动态性上很好地满足了这类资源的需求。

(3) 组合性要求不同种类的资源能够进行组合服务。在这点上, EAV 的元数据(Metadata)首先能够在一定程度上对资源的组合服务进行定义和解释,其次 A-V 键值对的数据格式使得第三方集成组件在方便快捷地获得资源属性特征,然后将资源进行更加有效、准确的组合。

最后,要求在云制造环境下,这些资源能够快速地被发现并匹配复杂的需求信息,而对于 EAV 模型来说,这个要求在性能上有点力不从心。然而,同样基于 EAV 模型数据的可访问性,我们可以为整个资源数据构建索引体系,来曲线解决查询性能上的劣势问题。

通过上述的案例分析可知, EAV 数据模型用在云制造资源上面是合理的。虽然外协加工资源只是其中的一类资源,但各类文献资料显示,分散性、自治性、异构性、多样性、动态性、组合性等这些特性是几乎所有云制造资源的共性,因此本文使用 EAV 模型解决云制造资源能力的数据库模型表示问题。

2 制造资源能力的 EAV 模型表示

2.1 制造资源能力的描述模型

本文参照文献[9]中提到的制造能力服务组织层次,从 3 种粒度描述制造资源能力服务,以期通过一种映射函数将其转化成文本信息处理领域的形式模型。

(1) 元子能力表达单元

元子能力表达单元是元子能力的基本表达单元,主要针对具体的制造操作,目的是通过结构化的描述模型将制造操作抽象为具有特征属性的操作能力,便于系统管理和用户查找。元子能力表达单元(Atomic Capability Unit, 简称 ACU)用六元组来抽象表示:

ACU = {BasicInfo, FunctionInfo,
Tools, Devices, Operators, FeaturInfo}

(2) 零部件能力表达单元

零部件能力表达单元是零部件能力的基本单元, 零部件能力表达单元描述的是企业通过一些技术手段将材料, 配件等进行加工最终形成具有流通价值的零部件的能力。对零件, 组件与部件的描述不做明显区分, 表达单元统称为 Component Capability Unit, 即 CCU。

CCU = {BasicInfo, Designers,
FunctionInfo, FeatureInfo, OtherUnitSet}

(3) 产品能力表达单元

产品能力表达单元是产品能力的基本单元, 产品能力单元描述的是企业通过装配等技术手段将材料, 零部件等进行组装与加工, 最终形成具有更高附加值的产品的能力。产品能力表达单元为 Product Capability Unit, 即 PCU。

PCU = {BasicInfo, Designers, FunctionInfo,
FeatureInfo, OtherUnitSet}

由上可知 3 类描述形式具有明显分层结构, 且每层结构有其特定的参数名和参数类型, 如对某一 ACU 的 description 属性(存在于 BasicInfo 描述单元组中)而言, 它的层次可理解为 ACU – BasicInfo – Description。因此本文引入参数路径的概念 attrPath 来表示某一制造能力属性的层级关系。以 ACU 的 description 属性为例, 定义

$$\text{attrPath}(\text{ACU}, \text{description}) =$$

$$\langle \text{acu.basicInfo.description} \rangle$$

并将不同数据类型 dataType 与参数路径结合, 则用向量表达形式定义为 $F = \langle \text{dataType}, \text{attrPath} \rangle$, 即 ACU 的 description 属性

$$F(\text{ACU}, \text{description}) =$$

$$\langle \text{text}, \text{acu.basicInfo.description} \rangle$$

由于 attrPath 具有唯一性, 很容易证明 F 为制造能力属性的一种映射函数且结果集为一个空间向量集合。

以制造能力中的 ACU 这种类型为例, 其属性结构如表 1 所示, 其中打勾表示其所在列的制造能

力在其所在行的属性上有非空的属性值。

表 1 制造能力类型 ACU 的属性结构简化表
Tab. 1 Attribute structure simplified table of ACU

Attribute		Ability					...
		ACU1	ACU2	ACU3	ACU4	ACU5	
Basic Info	名称	✓	✓	✓	✓	✓	
	描述	✓		✓	✓		
	...						
Feature Info	尺寸		✓	✓			
	材料			✓		✓	
	精度	✓			✓		
.....							
...							

从表格中可以很直观地看出, ACU 的属性数据呈现稀疏性的特征, 同时 ACU、CCU 和 PCU 这 3 种类型的制造能力实体对象具有的属性集是不尽相同的, 具有异构特征。

2.2 制造资源能力的 EVA 形式化模型

基于上述定义, 本文使用 EAV 模型实现制造资源能力的形式化存储, 主要包含: 数据模型表结构的设计、元数据表的定义和数据接口代码的实现。

(1) 数据模型表结构的设计

如图 1 所示, 基于 EAV 模型实现制造资源能力的形式化存储包含 3 类数据表:

Entity 表: 实体对象是资源能力 Ability, 由表 ability_entity 表示, 包含 entity_id (唯一标识), status (状态), create_time (创建时间) 和 last_modified_time (最后修改时间), sku (编码), type_id (分类 ID, 具体定义在 ability_type 表中) 基本属性。

Attribute 表: 制造能力的属性由表 ability_attribute 表示, 包含 attribute_id (唯一标识), attribute_code (属性编码), attribute_label (属性名称), attribute_type_id (分类 ID, 具体定义在 ability_attribute_type 表中), data_type (属性的数据类型), path (属性的参数路径), is_search (该属性是否用于搜索)。其中, path 是制造能力属性表中记录属性参数路径 attrPath 的字段; data_type 可以

用以表示制造能力属性值的五种基本数据类型 (varchar、text、int、decimal 和 datetime)。此外，系统还支持三种特殊的数据类型：static、option 和 custom。

Value 表：针对 varchar、text、int、decimal 和

datetime 五种基本数据类型，分别使用一张属性值表与之对应，命名规则是 ability_entity_*。每张表包含 value_id (唯一标识)，attribute_id (对应的属性 ID)，entity_id (对应的实例对象 ID) 和 value (属性值) 字段。

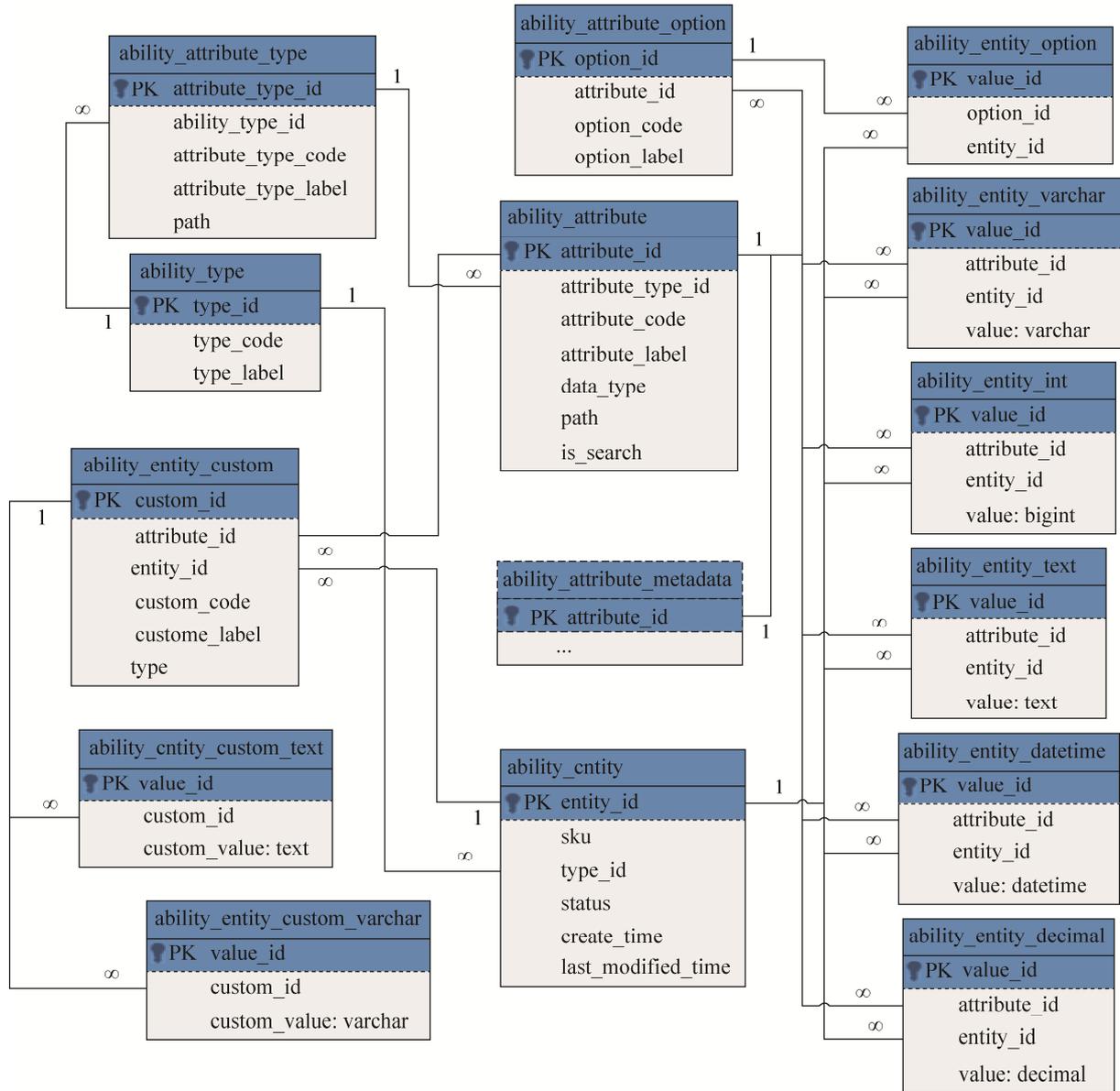


图 1 制造能力 EAV 模型数据库 UML 图
Fig. 1 UML diagram of EAV models

(2) 元数据表的定义

元数据(Metadata)表作为属性的附属表，针对不同的信息需求(分组、验证、展示等)，有各种各样的定义。本文着重讨论的是如何将制造资源的描

述模型转化成数据模型存储在数据库，对于数据在 Web 服务中如何与前端 UI 交互、数据的操作界面如何实现等问题，不做详细的论述。

(3) 数据接口代码的实现

为实现将各平台中基于列模型的制造资源数据转化成本文系统所需的 EVA 数据模型存储, 还需实现基本数据访问对象接口、EAV 模型对象类和数据库的动态创建及数据的迁移 3 类功能, 相关类图如图 2 所示。

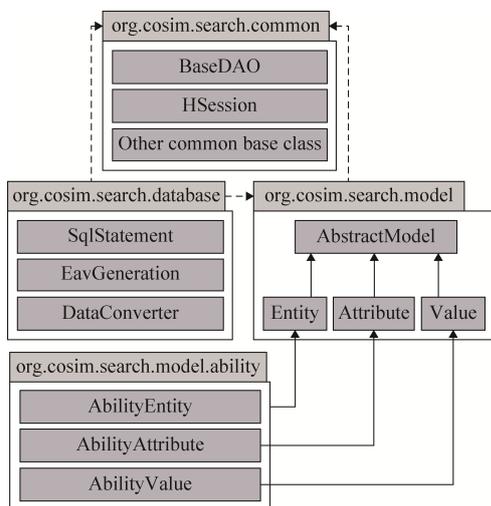


图 2 制造资源/能力 EAV 模型数据交互操作接口实现类图
Fig. 2 Data interworking interface implementation class diagram of manufacturing resource/capacity EAV models

基本数据访问对象接口: 是系统业务逻辑层与数据库之间数据交互的桥梁。本文在 common 类包中创建了两个类(HSession 和 BaseDAO), 前者可通过配置实现多个数据库连接池的创建和会话, 后者封装了 Hibernate 和 JDBC 两种方式的数据操作方法。这两个类是其他所有需要和数据库数据进行交互的类所依赖的封装类。

EAV 模型对象类的实现: 本文用 Entity、Attribute 和 Value 三个基本类表示, 它们同时继承 AbstractModel 抽象类。当需要表示具体的实体对象时, 则分别继承对应的基本类, 如 AbilityEntity 继承了 Entity, AbilityAttribute 继承了 Attribute, AbilityValue 集成了 Value。

数据库的动态创建及数据的迁移: 本文使用 XML 文件配置的方式实现 EAV 数据库相关表的动态创建。对于平台管理者而言, 只需修改 XML 文件即可定义制造能力实体对象及它的属性集。同时通过 SqlStatement 和 EavGeneration 两个类实现了

基于 XML 制造能力描述文件的 EAV 数据库表的动态创建。另外, 为了验证原型系统的可使用性, 本论文通过 DataConverter 类把所有用传统列模型存储的数据转化成 EAV 模型存储。

3 基于 Solr 的制造资源搜索框架

如上所述, EVA 模型能够解决云制造资源异构性、数据稀疏性、动态性等问题, 但批量查询和多属性条件查询会造成整个系统性能下降, 使得云制造平台的资源发现服务效率很低。本节将引入 Solr 索引机制, 同时对它进行基于 EAV 模型的架构设计和二次开发, 用以解决云制造平台中资源的发现问题。

3.1 Solr 搜索框架

Solr 是一个基于 Apache Lucene 的搜索引擎服务器, 它的两大最重要功能是“数据索引”和“全文本搜索”, 其在底层数据存储上使用了 ZooKeeper 分布式服务框架, 用以管理大量的各种类型的索引数据等。Solr4 的主要组件架构如图 3 所示^[10]。

从图 3 中可看到, Solr 是一个 Java Web 应用, 它可部署运行在任何主流的 Java 应用引擎(如 Tomcat, Jetty, JBoss 等)中。同时, Solr 在索引管理上提出了 Multi-core 的概念, 并且在 Solr4 中提出了 SolrCloud 的概念: Multi-core 使得 Solr 的一个服务实例能够同时管理多个类型的独立的索引库; SolrCloud 是基于 Solr 框架和 ZooKeeper 框架的分布式搜索方案, 实现了 Solr 服务节点的集群管理, 拥有集中式的配置信息、自动容错、近实时搜索、查询时自动负载均衡等功能特点。

3.2 基于 Solr 的制造资源 EVA 模型搜索方法

上节实现了云制造资源的 EAV 模型结构表示, 通过 XML 格式文件配置的方式将制造资源能力的描述模型转化成了 EVA 存储结构, 并且将所有制造资源能力数据迁移至数据库中。但是在 EAV 模型中, 虽然属性和属性值的管理是集中的, 但针

对某一实例对象的属性集而言, 它们的分布是碎片化的, 这种存储方式使得批量查询和多属性条件语句查询性能低, 不利于整个系统搜索资源速度的提升. 因此, 本节基于 Solr 中索引基于文档(Document)管理的倒排索引将制造资源能力 EVA 模型的搜索属性统一管理, 提高其搜索效率; 并使用 SolrCloud 分布式搜索框架为存储在不同平台中的制造资源能力提供统一的管理。

基于 Solr 的制造资源 EVA 模型搜索方法, 需要完成数据库连接/会话池的创建、索引对象的获取、索引对象向文档的转化和 Solr 索引库的映射和分发 4 个步骤(如图 4)。并采用队列的方式实现

多线程操作, 以及使用“池”和“空间”的概念实现可分布式框架。

(1) 数据库连接/会话池的创建

EAV 模型针对的是具有异构型和数据稀疏性的实体对象, 在整个系统平台中, 对其他特征的实体对象而言, 用的较多的还是传统数据库模型, 特殊的实体对象甚至会用到 NoSQL 数据库。另外, 就算指对 EAV 模型的数据创建索引, 当数据量达到一定程度时, 通常会采用分库、甚至是分布式存储。因此, 在创建索引时, 要实现一个数据库连接和会话池, 使得索引系统能够快速地与任何服务器上的任何类型的数据库会话, 从而进行数据交互。

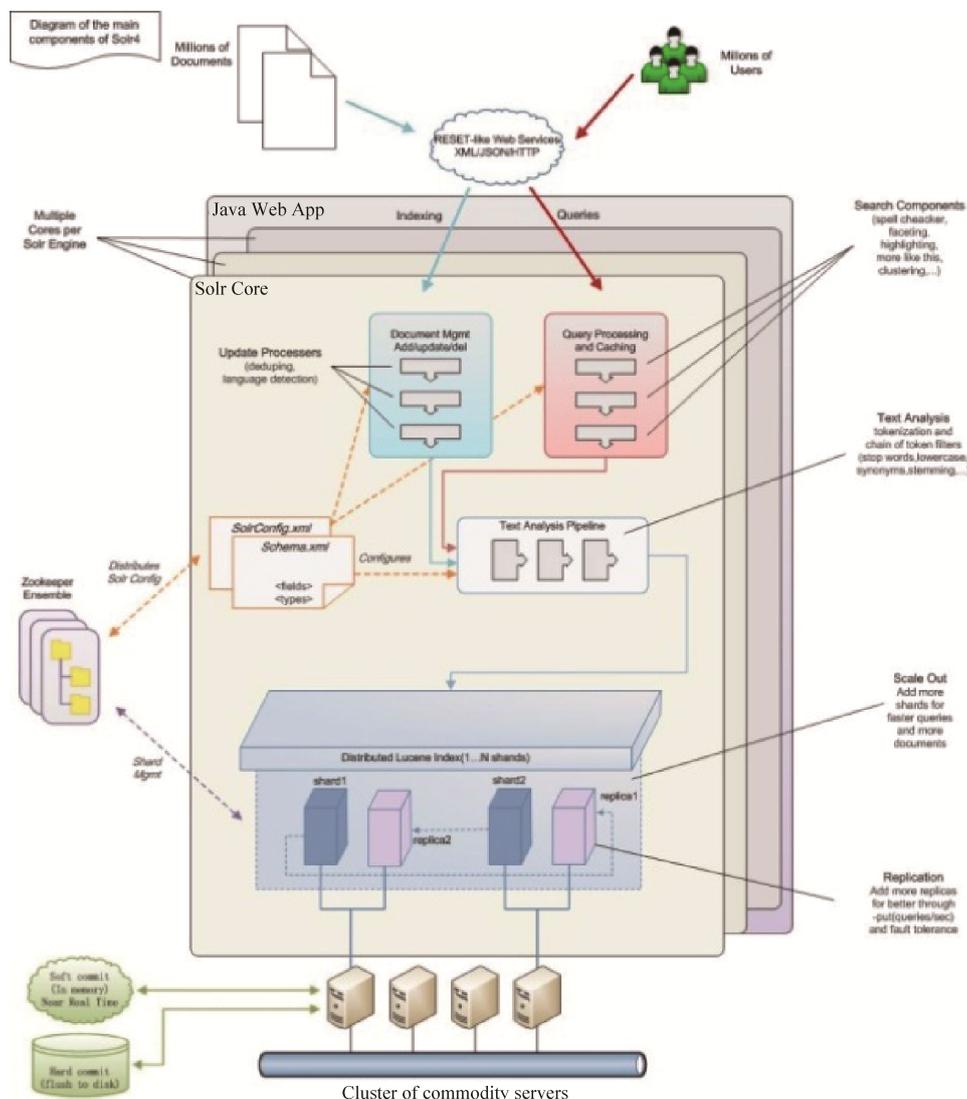


图 3 Solr4 的主要组件架构图
Fig. 3 Main components of Solr4

<http://www.china-simulation.com>

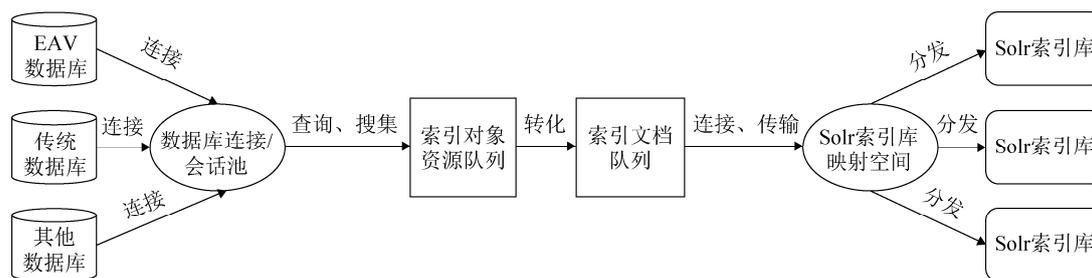


图 4 云制造平台基于 Solr 的索引架构图

Fig. 4 Solr based index architecture of Cloud Manufacturing platform

(2) 索引对象的获取

在正式调用 Solr 的 API 创建索引之前, 我们需要从数据库中查询并搜集需要创建索引的数据。在云制造平台中获取某一制造能力相关属性和属性值有两种方式: 1) 通过表 `ability_attribute` 中字段 `Is Search` 的值直接判断。2) 通过基本数据类型和属性路径判断。在制造能力描述模型转化成形式化模型的过程中, 我们定义了属性的二维映射函数, 其中一个维度是基本数据类型, 另一个维度是具有唯一性的属性路径。因此, 我们可以根据基本数据类型和属性路径来判断并寻找那些在模型中定义为“需要创建索引”的属性。

(3) 索引对象向文档的转化

在 Solr 索引库中, 对象是以文档(Document)的数据格式存储于倒排索引中的, 在获取完索引对象及其属性集之后, 我们要进行数据格式的转化。文档的数据结构实际上是一种拥有指定标签规则的 XML 格式。

实体对象数据转化成这种文档格式的字符流之后, 便可以调用 Solr 的 API, 以 HTTP POST 的方式将这些索引数据传输到 Solr 搜索引擎服务器中。

(4) Solr 索引库的映射和分发

索引的文档队列, 并没直接和 Solr 索引库进行直接的交互。在一次索引的创建过程中, 可以对不同的实体对象创建各自的索引, 基于 Solr 的 `Multi-core` 概念, 存储在不同的 Solr Core 里面。当索引文档进入文档队列时, 它并不知道应该去往哪个索引库里。因此, 需要在文档队列和 Solr 索引

库之间, 创建一个 Solr 索引库映射空间, 实现文档的自动分发和定向路由功能。包 `index` 中的类 `IndexSpace`, 可通过配置 XML 文件映射关系从而初始化实体对象和 Solr 索引库对应关系。在包 `index.doc` 的类 `DocDispatch` 中实现文档根据不同的实体对象分发、路由至不同的索引库。

4 系统实现

4.1 系统架构

整个系统分为 2 个部分: 数据部分和搜索部分。数据模型部分关注的是数据模型和数据库交互等功能, 搜索部分关注的是索引的创建、查询以及用户搜索体验等功能。如图 5 所示, 本文中的系统共有 7 大模块和 2 个资源池/空间。下面根据图中最主要的 10 大步骤为该系统的做详细的介绍。

读取模型文件。为每类制造资源能力制定一套 XML 元模型描述文件, 数据模型转化模块能够识别这种特定规则的 XML 文件内容, 从而动态解析出文件中包含的模型结构。

初始化 EAV 数据库结构。数据模型转化模块根据从 XML 文件中解析出来的实体对象模型结构, 通过 EAV 模型数据交互接口模块, 与 EAV 数据库进行交互动态生成对应的实体对象 EAV 表结构。

获取传统数据库资源数据。由于目前大部分的云制造资源已经使用传统的列模型表结构存储, 因此数据模型转化模块通过 DAO 模块从传统数据库中取出所有的资源数据, 将原来的资源对象通过属性字段对应关系, 拆解、重组成 EAV 资源对象。

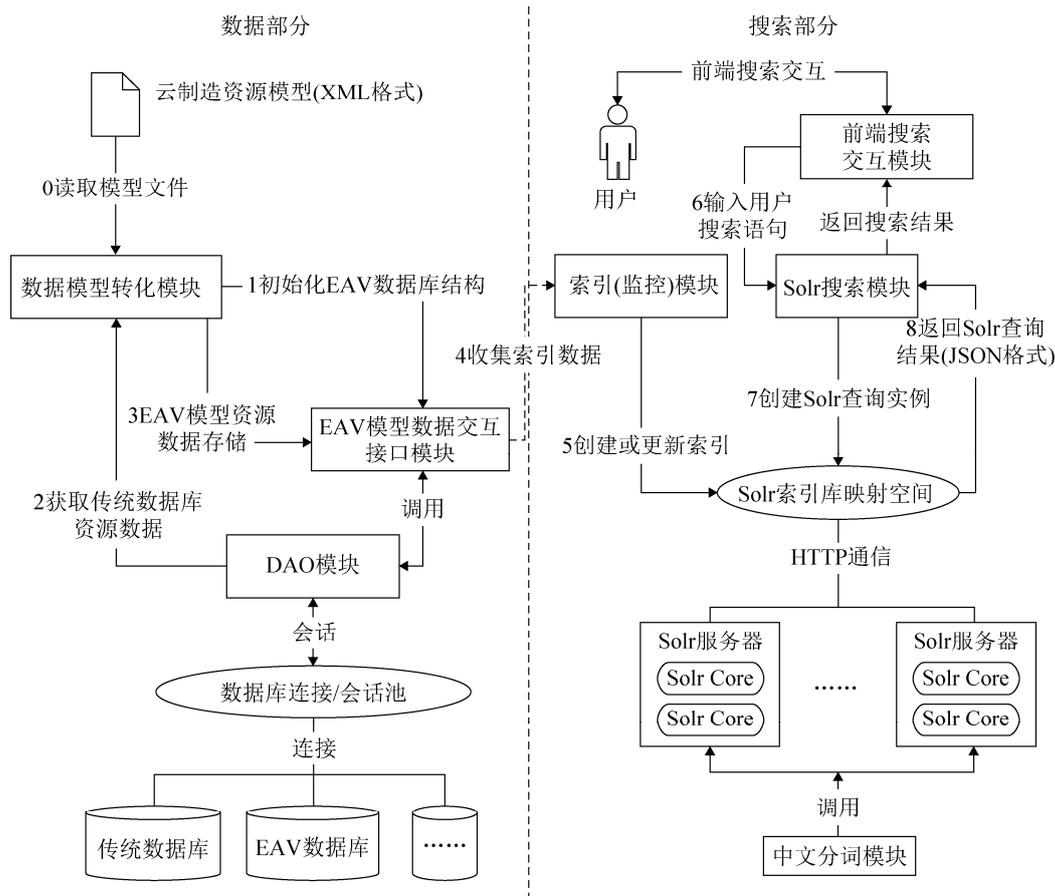


图 5 云制造平台发现服务系统架构图

Fig. 5 Service discovery system architecture of Cloud manufacturing platform

EAV 模型资源数据存储。通过 EAV 模型数据交互接口模块，将所有新的资源对象存储至 EAV 数据库中。

搜集索引数据。索引模块能够通过 EAV 模型数据交互接口模块从数据库中获取需要创建索引的数据，形成文档(Document)格式的字符串流，放入索引队列中，等待与 Solr 服务器的通信。同时，索引模块还能监控索引的更新状态，定时将 Solr 索引与数据库进行数据同步。

创建或更新索引。通过 Solr 索引库映射空间中初始化的实体对象-Solr 索引库对应关系，索引模块调用 Solr 基于 HTTP 协议的 REST-like 服务 API，用 POST 的传输方式，将文档格式的字符串流以及相关命令传输至 Solr 服务器，实现索引的创建和更新。

输入用户搜索语句。前端搜索交互模块接收到

用户的搜索需求之后，经过简单的预处理，将搜索参数传入搜索模块中。

创建 Solr 查询实例。Solr 搜索模块在接收到搜索请求参数之后，将搜索语句和其他参数组装成符合 Solr 查询规则的 URL，创建查询实例，然后通过 HTTP 协议与 Solr 搜索引擎服务器通信，实现云制造资源基于关键词的发现和匹配。

返回 Solr 查询结果。Solr 搜索引擎返回的查询结构有两种格式——XML 和 JSON。本论文采用 JSON 格式获取查询结果，然后将其在 Solr 搜索模块中解析，形成能够和前端交互的搜索结果对象。

返回搜索结果。将最终的搜索结果用 HTML 页面呈现的形式返回给用户。

4.2 原型系统案例

本文将上述研究成果应用于“天智网”云制造

平台中,以制造能力为例,实现了资源的数据存储和搜索服务。原型系统使用了 Java 编程语言和 MySQL 数据库,部署在 Tomcat 服务容器中进行测试。将原系统的制造能力数据全部进行迁移,并与之前的搜索服务相比较,本文的系统在以下的几个方面有了功能和性能上提升或保持(图 6 所示)。



(a) 关键词“纺织”的返回结果



(b) 搜索“高性能浆料纺织技术”等较复杂语句的返回结果

图 6 原型系统搜索案例

Fig. 6 Prototype system search case

(1) 提升了关键词搜索结果的查全率

原系统只是将搜索关键词与资源名称等主要的简单字段进行布尔型匹配,而该系统则是对资源的所有文本型属性(也可通过配置自定义选择)创建基于关键词分析的倒排索引,通过权重计算返回

匹配的有序结果,具有较高的查全率。比如原系统搜索关键词“纺织”,只能得到 4 个匹配结果;在本论文系统里,共有 9 个匹配结果(图 6(a)所示)。

(2) 搜索语句的简单分析处理

在原系统中,如果用户在搜索框中输入“高性能浆料纺织技术”、“high performance 浆料纺织技术”、“HP 浆料纺织技术”等较长语句、英文关键词、英文缩写词汇时,返回结果数量是 0。而在本论文的原型系统中,这些搜索语句的返回结果如图 6(b)所示。

5 结论

针对现有文本信息处理技术在处理具有分散性、自治性、异构性、多样性、动态性和组合性特点的制造资源能力服务搜索时的不足,将 EVA 模型和 Solr 搜索引擎应用到云制造服务平台中,以提升其对制造资源的发现能力。主要工作包括:

(1) 云制造资源描述模型和自然语言形式化模型之间的转化。在现有的资源描述模型研究成果的基础上,找出了云制造资源描述模型和文本信息处理形式化模型之间的某种映射关系。

(2) 资源数据的结构化存储。本文基于 EAV (Entity-Attribute-Value) 数据模型进行适应于云制造平台的改造和升级,使得资源的描述模型转化成数据结构模型后,能够动态地为模型添加或删除属性,而不用去修改数据表结构本身以及数据操作接口,很好地满足了云制造资源属性多样性、动态性和异构性的要求。

(3) 索引体系的构建。EAV 数据结构虽然在模型上解决了云制造资源属性的各种特性需求,但会因此在数据库衍生出很多张表,造成数据分布零散,不易获取。因而本文在 EAV 模型的基础上,引入了 Solr 企业级索引架构,提升了服务的查询效率。

参考文献:

- [1] 尹胜,尹超,刘飞,等. 云制造环境下外协加工资源集成服务模式及语义描述 [J]. 计算机集成制造系统,

