

基于关系数据库的模糊本体的存储方法

吕艳辉

(沈阳理工大学信息科学与工程学院 沈阳 110159)

摘 要 模糊本体被广泛用来描述语义 Web 上的模糊信息。如何对模糊本体进行有效的存储,逐渐成为语义 Web 领域一个重要的研究课题。鉴于关系数据库在数据组织与管理方面的优势,提出基于关系数据库的模糊本体的存储方法。首先,通过对 RDF 数据类型进行模糊扩展,给出模糊本体中模糊数据类型的表示方法;然后,提出了基于关系数据库的模糊本体的存储模式以及模糊本体的结构和实例在关系数据库中的存储方法;最后,证明了该存储方法的正确性。基于关系数据库的模糊本体的存储方法为语义 Web 中模糊数据的存储和管理提供了一个有效的解决方案。

关键词 语义网,模糊本体,关系数据库,存储,信息容量

中图法分类号 TP182 **文献标识码** A

Storage of Fuzzy Ontologies Based on Relational Databases

LU Yan-hui

(College of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China)

Abstract As fuzzy ontologies are widely used in describing fuzzy information in the Semantic Web, how to efficiently store fuzzy ontologies is becoming a more important research issue in the domain of the Semantic Web. An approach to store fuzzy ontologies based on relational databases was proposed by making use of the advantage of organizing and managing data of relational databases. Firstly, a method of representing fuzzy datatypes was introduced by fuzzy extending of the RDF datatypes. Then, the storage schema of fuzzy ontologies using relational databases was presented. On this basis, the storage processes of the fuzzy ontologies structure and instances were given. Finally, the correctness of the storage schema of fuzzy ontologies was proved. The approach of storing fuzzy ontologies in relational databases provides an efficient solution to the storage and management of fuzzy data in the Semantic Web.

Keywords Semantic Web, Fuzzy ontologies, Relational databases, Storage, Information capacity

本体是概念模型的明确规范说明,是实现语义 Web^[1] (Semantic Web)的关键技术之一。在很多应用领域中,由经典本体支持的形式化概念不足以表示和处理现实世界中广泛存在的不精确、不确定信息。通常的解决方案是将模糊逻辑^[2,3]引入本体,以处理这类模糊信息。近几年来,随着语义 Web 的发展,有大量研究致力于模糊本体^[4-7]的构建,并取得了一定的研究成果。但是,若将模糊本体应用到实际系统中,必须选择合适的方法将之存储起来。因此,如何有效存储现存的模糊本体,是语义 Web 研究领域一个很有意义的研究方向。

在经典环境下,本体的存储方法主要分为基于纯文本存储和基于关系数据库存储两类。基于纯文本存储方法是本体库以文件形式存储在本地文件系统中,这种存储形式的缺点在于不适合较大规模的本体库,因为它每次都需要读入内存操作,受到内存大小的限制。对于本体海量数据的存储和管理,利用关系数据库是一种较好的选择。文献[8]针对本体描述语言的特点,给出了本体存储模式的设计原则,并基于该原则提出了基于关系数据库的存储模式;文献[9]提出了一个

用于本体存储和推理的系统 Minerva,采用 WSM-LDL 语言完成了主要的描述逻辑推理任务,并给出了本体在关系数据库中的存储模式。上述文献研究的主要内容都是利用关系数据库来保存本体,即给定一个本体,将其结构和实例存储在关系数据库中,这不同于通过给出本体到关系数据库的转换(或映射)来实现本体的存储。后者是通过分析本体和关系数据库模式之间元素的对应关系,在本体和关系数据库之间建立一组语义转换(或映射)规则或算法,例如文献[10-13]涉及到的研究内容。

与经典本体的存储方法相比,模糊本体的存储不但要满足经典本体存储模式的要求,还要处理大量的模糊数据,这给模糊本体的存储带来巨大的挑战。目前,对模糊本体存储的研究工作还比较少。文献[14]针对 OWL 本体的一些主要构造子,给出了在关系数据库中的存储模式,没有考虑模糊本体实例的存储方法;而文献[15]通过分析模糊本体与关系数据库的对应关系,给出模糊本体结构与关系数据库的转换算法,并没有给出模糊本体的存储模式及相应的存储方法。此外,现有研究(模糊)本体存储方法的文献中,都没有给出相应存

到稿日期:2010-07-15 返修日期:2010-12-30 本文受国家自然科学基金(60873010),教育部新世纪优秀人才支持计划项目(NCET-05-0288),教育部高等学校博士学科点专项科研基金资助项目(20050145024)资助。

吕艳辉(1971-),女,博士,副教授,主要研究方向为语义 Web、模糊本体管理等,E-mail:yanhuilv@126.com。

储方法的正确性证明。

本文的研究工作首先从本体中模糊数据的表示方法入手,根据属性域是连续域还是离散域提出了两类模糊数据类型。在此基础上,研究了基于关系数据库的模糊本体的存储模式。然后,根据模糊本体的特点,分别讨论了模糊本体的结构和实例在关系数据库中的存储方法。最后,证明了模糊本体存储方法的正确性。

下面将从模糊本体的形式化定义、基于关系数据库的模糊本体的存储模式、模糊本体的结构和实例的存储方法以及正确性证明几个方面进行介绍,并指出今后的研究方向。

1 模糊本体

1.1 模糊数据类型

语义 Web 本体应用中的模糊知识建立在大量模糊数据的基础之上。为准确描述这些模糊数据,需要根据模糊数据模糊性的表现形式为其赋予相应的数据类型。OWL(Web Ontology Language)^[16]作为描述语义 Web 本体的语言标准,建立在 W3C(World Wide Web Consortium)标准 RDF(Resource Description Framework)^[17]和 RDFS(RDF Semantics)^[18]之上,并且是基于 RDF 的数据类型。RDF 使用类型文字(Typed Literals)来提供数据类型信息,除预定义的数据类型 `rdf:XMLLiteral` 外,在 RDF 类型文字中使用的数据类型都需要在 RDF 外部定义并用 `URIref` 标识,然后通过 `URIref` 引用数据类型并与相应的文字进行关联,来描述 RDF 中的文字。因此,RDF 允许在其外部定义新的数据类型,并将该数据类型应用在 RDF 中^[19]。RDF 数据类型机制的这种特点为本体中模糊数据类型的定义提供了自然和有效的方法。

RDF 采用(资源 属性 值)这样的数据模型来表示 Web 上资源的属性的值,根据属性域是连续域还是离散域,将 RDF 的数据类型进行模糊扩展,提出两类模糊数据类型(Fuzzy Datatypes,FD)来表示本体中的模糊数据。

第一类模糊数据类型描述的是作用在连续域上的属性的模糊值,相应地称这样的属性为类型 1 模糊属性。这类数据类型包括 Interval, Approx, Tag, Trapezoidal, 还包括 Unknown, Undefined 和 Null 类型。其中,数据类型 Interval (Interval:[m, n])表示一个属性的取值为某一个区间范围[m, n],如“年龄”属性的取值可以是一个实数域中的区间[38,42](其函数表示如图 1 所示);数据类型 Approx (Approx: a)表示一个属性的取值可以是一个大约值 a ,如“身高”属性的取值为“180cm 左右”(其函数表示如图 2 所示);数据类型 Tag 表示一个属性的取值可以使用自然语言中的术语(即语言常量),如“速度”属性的取值可以是“慢速”、“中速”、“快速”等语言常量;数据类型 Trapezoidal (Trapezoidal:[$\alpha, \beta, \gamma, \delta$])表示一个梯形模糊数,其中 α, δ 的取值分别为梯形模糊数支集的最小值和最大值, β, γ 的取值分别为梯形模糊数核集的最小值和最大值。梯形模糊数常用于描述语言常量,如“年龄”属性的取值可以是语言常量“年轻”,这可以用梯形模糊数表示为[0,16,30,40](其函数表示如图 3 所示);数据类型 Unknown 表示属性的值未知,数据类型 Undefined 表示属性的值未定义,数据类型 Null 则表示属性的取值是空值。

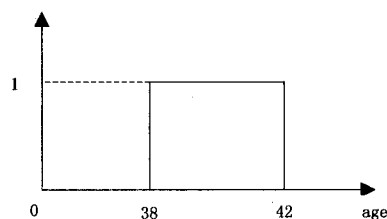


图 1 区间值(Interval:[38,42])的函数表示

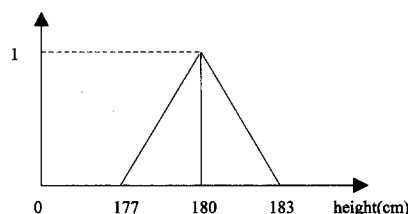


图 2 大约值(Approx:180)的函数表示

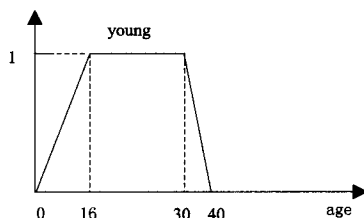


图 3 语言常量 young 的函数表示

第二类模糊数据类型描述的是作用在离散域上的属性的模糊值,相应地称这样的属性为类型 2 模糊属性。这类数据类型包括 Label 和 Posdis,其中数据类型 Label 同样表示具有该数据类型的属性的取值是语言常量,如“工作能力(Productivity)”属性可以取值“较差”、“一般”、“较好”、“优秀”等。由于“工作能力”属性定义在离散域上,因此要给出“工作能力”属性各个值之间的相似关系;数据类型 Posdis 表示属性的取值是一个可能性分布(Possibility Distribution),如“e-mail”属性由可能性分布{J_S@yahoo.com/0.9,J_Smith@msn.com/0.8,J_S@msn.com/0.85}给出,表示 e-mail 为“J_S@yahoo.com”的可能性是 0.9,为“J_Smith@msn.com”的可能性是 0.8,为“J_S@msn.com”的可能性是 0.85。第二类模糊数据类型也可以包括 Unknown, Undefined 以及 Null 类型。

为了使本体能够描述 Web 资源中的模糊数据,需要在本体的外部定义模糊数据类型 `rdf:FuzzyDatatype`,包括基础类型 `fd:interval`,`fd:approx`,`fd:tag`,`fd:trapezoidal`,`fd:unknown`,`fd:undefined`,`fd:null`,`fd:label` 以及 `fd:posdis`,并用 `URIref` 来标识。然后通过给本体中的文字关联一个模糊数据类型,形成模糊类型文字,就能够实现对模糊本体中各种类型的模糊数据的表示。由于在本体的外部对模糊数据类型的定义更倾向于纯技术方法,本文对此不予讨论。

1.2 模糊本体的形式化定义

下面基于模糊本体支持的模糊数据类型以及模糊本体的模糊性,给出模糊本体的形式化定义。

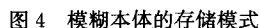
定义 1 一个模糊本体 FO 是一个四元组 $FO = \{C, I, P, X\}$,其中:

C 是类的集合, $Root \in C$ 为根类,一个模糊本体有且只有一个根类。模糊本体中所有的类通常构成一个分类层次,每个类 $c \in C$ 由作用到该类上的所有属性来描述。

I 是实例的集合,表示属于某个类的基本元素或具体实

由模糊本体的定义可知,模糊本体包括模糊本体结构和实例两部分,其中模糊本体的结构包括类、属性及其公理。

根据模糊本体的特点以及模糊数据类型的表示方法,给出模糊本体的存储模式,如图 4 所示。其中每个表的主键用下划线标识,虚线框内的部分用于组织和存储模糊本体中涉及到的模糊数据类型以及模糊数据。可以看出,图 4 所示的存储模式是按照模糊本体的资源类型及特点来进行分类存储。其中,资源表 Resource 存储了模糊本体中所有资源的标识符、URIref 和资源类型,并在此基础上按类、属性、实例分别进行存储。类的存储分别用类关系表 C_relation 以及用于复杂类存储的各构造子表来实现,包括表 C_intersection, C_union, C_complement, C_disjoint 和 C_oneof。属性的存储则分别用属性关系表 P_relation、属性定义域表 P_dom、属性值域表 P_ran、属性特性表 P_char 以及属性限制表 P_con 来实现。



(3)提高查询效率。在现有的基于关系数据库的存储模式中,除垂直模式外,其他存储模式的表结构都不稳定,在实际应用中具有很大的局限性。垂直模式可以使数据行变小,查询时能够减少 I/O 次数,但查询所有数据都需要连接操作。在数据库中,连接操作是比较难处理和优化的。因为多个表进行连接时可以有多种不同的次序,使得对应于查询执行计划的数目会随着该查询所含表的个数增加而呈指数级增长。

$$I = \{\text{staffid } 11001, \text{depid } 0206, \text{courid } 309\}$$

$X = \{ \text{subClassOf}(\text{AdminStaff}, \text{Staff}), \text{subClassOf}(\text{AcademicStaff}, \text{Staff}) \}$

由类 AcademicStaff 的实例 staffid_11001 的属性值可知 John Smith 教授工作在 depid_0206 系, 所教的课程是 courid_309, 正值中年(middle), 身高大约为 180cm (这里用符号“&.”表示大约的含义), 邮箱(email)为 {JS@yahoo. com/0. 8, JohnS@msn. com/0. 9}, 工作能力很强(excellent)。其中, 属性 staffname 和 title 为经典属性; 属性 work_in 和 teach 是对象属性, 其属性值为一个实例; 属性 age 和 height 为类型 1 模糊属性, 属性 email 和 ability 为类型 2 模糊属性。此外, 属性 age 的模糊值 middle 为语言常量, 属于 Tag 类型; 属性 height 的模糊值 &180cm 表示一个大约值, 属于 Approx 类型; 属性 email 的模糊值 {JS@yahoo. com/0. 8, JohnS@msn. com/0. 9} 表示一个可能性分布, 属于 Posdis 类型; 属性 ability 的模糊值 excellent 为语言常量, 属于 Label 类型。

下面结合图 5 所示的模糊本体, 基于图 4 给出的存储模式, 研究模糊本体结构在关系数据库中的存储方法。

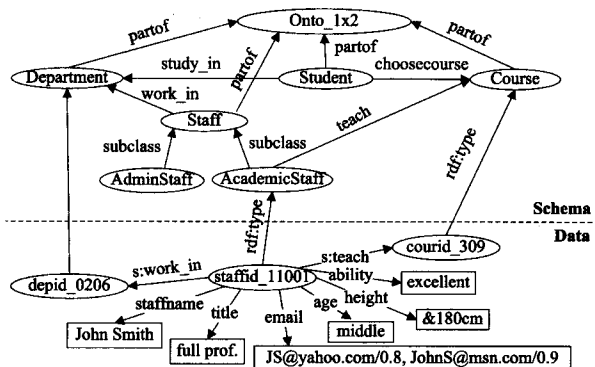


图 5 模糊本体的一个例子

模糊本体的结构主要指类、属性以及相应的公理。按模糊本体的存储模式, 需要将模糊本体的所有资源, 包括类、属性以及实例等保存在表 Resource 中, 如表 1 所列。其中, “onto_1x2”是图 5 所示的模糊本体名。

表 1 Resource 表

ontology	rid	namespace	localname	type
onto_1x2	c_1	http://www.syt.edu.cn/ailab/	department	class
onto_1x2	c_2	http://www.syt.edu.cn/ailab/	staff	class
onto_1x2	c_3	http://www.syt.edu.cn/ailab/	adminstaff	class
onto_1x2	c_4	http://www.syt.edu.cn/ailab/	academicstaff	class
onto_1x2	c_5	http://www.syt.edu.cn/ailab/	student	class
onto_1x2	c_6	http://www.syt.edu.cn/ailab/	course	class
onto_1x2	p_1	http://www.syt.edu.cn/ailab/	study_in	property
onto_1x2	p_2	http://www.syt.edu.cn/ailab/	work_in	property
onto_1x2	p_3	http://www.syt.edu.cn/ailab/	teach	property
onto_1x2	p_4	http://www.syt.edu.cn/ailab/	choosecourse	property
onto_1x2	p_5	http://www.syt.edu.cn/ailab/	staffname	property
onto_1x2	p_6	http://www.syt.edu.cn/ailab/	title	property
onto_1x2	p_7	http://www.syt.edu.cn/ailab/	email	property
onto_1x2	p_8	http://www.syt.edu.cn/ailab/	age	property
onto_1x2	p_9	http://www.syt.edu.cn/ailab/	height	property
onto_1x2	p_10	http://www.syt.edu.cn/ailab/	ability	property
onto_1x2	i_1	http://www.syt.edu.cn/ailab/	staffid_11001	instance
onto_1x2	i_2	http://www.syt.edu.cn/ailab/	depid_0206	instance
onto_1x2	i_3	http://www.syt.edu.cn/ailab/	courid_309	instance

表 2 存储了类 adminstaff(c_3)、类 academicstaff(c_4) 与类 staff(c_2) 的继承关系; 表 3 和表 4 存储了所有属性的定义域与值域。其中, 对象属性的定义域与值域都是类, 而对于数

据类型属性, 其定义域为类, 值域为数据类型, 按属性值取值范围分为 XML Schema 数据类型 (XSD) 和模糊数据类型 (FD)。

表 2 C_relation 表

cid1	cid2	relationship	μD
c_3	c_2	subclassof	1
c_4	c_2	subclassof	1

表 3 P_dom 表

pid	domain
p_1	c_5
p_2	c_2
p_3	c_4
p_4	c_5
p_5	c_4
p_6	c_4
p_7	c_4
p_8	c_4
p_9	c_4
p_10	c_4

表 4 P_ran 表

pid	range
p_1	c_1
p_2	c_1
p_3	c_6
p_4	c_6
p_5	XSD
p_6	XSD
p_7	FD
p_8	FD
p_9	FD
p_10	FD

图 5 所示的模糊本体并没有给出属性的约束。为说明属性约束的存储方法, 这里假定一个学生(student)或一个员工(Staff)属于且只能属于一个系(department), 一个教师(AcademicStaff)一个学期最多允许上 3 门课程, 一个学生(student)在一个学期最少选修 2 门课程(course), 最多选修 5 门课程(course)。这样就可以将涉及到的属性约束存储在表 5 中。

表 5 P_con 表

pid	constraint	pvalue
p_1	cardinality	1
p_2	cardinality	1
p_3	maxcardinality	3
p_4	mincardinality	2
p_4	maxcardinality	5

4 模糊本体实例的存储

相对于类和属性的存储, 实例的存储比较特殊。因为类是通过作用于它的属性来描述, 所以类的实例是通过相应的属性值来描述。这样, 就要区分属性值是精确值还是模糊值。如果一个类的属性值是精确值, 那么将该值存储在表 I_triple-value 中; 如果一个类的属性值是模糊值, 那么就要根据模糊值的数据类型使用相应的存储方法。

对于图 5 所示模糊本体, 实例 staffid_11001(i_1), depid_0206(i_2), courid_309(i_3) 分别是类 AcademicStaff(c_4), Department(c_1), Course(c_6) 的实例, 按模糊存储模式, 需要保存在表 I_of 中, 如表 6 所列。

表 6 I_of 表

iid	class	μD
i_1	c_4	1
i_2	c_1	1
i_3	c_6	1

对于实例 staffid_11001(i_1),其对象属性 work_in(p_2), teach(p_3)的值是一个实例,其数据类型属性 staffname(p_5), title(p_6)的值是精确值,需要保存在表 I_triplevalue 中,如表 7 所列。

表 7 I_triplevalue 表

iid	property	value
i_1	p_2	i_2
i_1	p_3	i_3
i_1	p_5	John Smith
i_1	p_6	full prof.

下面给出模糊本体中的模糊属性值在关系数据库中的存储方法。为便于说明,在图 5 的基础上补充类 Department 的 4 个实例,令其实例标识符分别为 i_4, i_5, i_6 和 i_7,其三元组形式如下所示:

i_4: (staffid_11002, staffname, Yuri Starks), (staffid_11002, title, lecturer), (staffid_11002, email, yuri@yahoo.com), (staffid_11002, age, young), (staffid_11002, height, high), (staffid_11002, ability, bad)

i_5: (staffid_11003, staffname, Ian Wright), (staffid_11003, title, lecturer), (staffid_11003, email, ianW@msn.com), (staffid_11003, age, 36), (staffid_11003, height, 178cm), (staffid_11003, ability, good)

i_6: (staffid_11004, staffname, Arthur Ashe), (staffid_11004, title, associate prof.), (staffid_11004, email, arthur_ashe@msn.com), (staffid_11004, age, * 38—42), (staffid_11004, height, low), (staffid_11004, ability, fair)

i_7: (staffid_11005, staffname, Joho Daly), (staffid_11005, title, full prof.), (staffid_11005, email, johodaly@yahoo.com), (staffid_11005, age, old), (staffid_11005, height, medium), (staffid_11005, ability, good)

其中,实例 i_6 的 age 属性值“* 38—42”中的符号“*”表示区间值的含义,即表示年龄在 38 岁至 42 岁之间。在上述实例中,没有考虑实例的对象属性的属性值,这是因为对象属性值是一个实例,故将此省略。下面以上述实例来说明不同类型的模糊数据如何在关系数据库中进行存储。

由图 4 所示的模糊本体的存储模式可知,使用表 Fuzzytype 存储模糊本体中所有的模糊数据类型,如表 8 所列,包括数据类型 Interval, Approx, Tag, Trapezoidal, Label, Posdis, Unknown, Undefined 和 Null,并分别用标识符 fd_01 至 fd_09 来标识。

表 8 Fuzzytype 表

fdid	fdname
fd_01	Interval
fd_02	Approx
fd_03	Tag
fd_04	Trapezoidal
fd_05	Label
fd_06	Posdis
fd_07	Unknown
fd_08	Undefined
fd_09	Null

为便于描述模糊本体实例的模糊属性值,要为每个模糊属性值赋予一个标识符,并将实例的标识符、实例的属性以及属性的模糊值标识符保存在表 Fuzzytrivalue 中,如表 9 所列。例如对于实例 staffid_11001(i_1),其属性 email(p_7)的模糊值标识符是 ftl_13,标识的是属性值 {JS@yahoo.com/0.8, JohnS@msn.com/0.9}。

表 9 Fuzzytrivalue 表

fiid	property	ftlid
i_1	p_7	ftl_13
i_1	p_8	ftl_2
i_1	p_9	ftl_8
i_1	p_10	ftl_12
i_4	p_8	ftl_1
i_4	p_9	ftl_6
i_4	p_10	ftl_9
i_5	p_10	ftl_11
i_6	p_8	ftl_7
i_6	p_9	ftl_4
i_6	p_10	ftl_10
i_7	p_8	ftl_3
i_7	p_9	ftl_5
i_7	p_10	ftl_11

每个模糊属性值的标识符、属性值以及相应的数据类型需要存储到表 Fuzzytypedlitteralist 中,如表 10 所列。例如对于属性值 young,其属性值标识符为 ftl_1,其数据类型为 Tag 类型(fd_03)。

表 10 Fuzzytypedlitteralist 表

ftlid	ftliteral	fdid
ftl_1	young	fd_03
ftl_2	middle	fd_03
ftl_3	old	fd_03
ftl_4	low	fd_03
ftl_5	medium	fd_03
ftl_6	high	fd_03
ftl_7	* 38—42	fd_01
ftl_8	&180	fd_02
ftl_9	bad	fd_05
ftl_10	fair	fd_05
ftl_11	good	fd_05
ftl_12	excellent	fd_05
ftl_13	JS@yahoo.com/0.8, JohnS@msn.com/0.9	fd_06

表 11 存储了类型 1 模糊属性的属性值,包括 Tag 类型(如 young(ftl_1), middle(ftl_2), old(ftl_3), low(ftl_4), medium(ftl_5), high(ftl_6)), Interval 类型(如 * 38—42(ftl_7)) 和 Approx 类型(如 &180(ftl_8)),其中模糊值 young(ftl_1)的 4 个参数分别为 0, 16, 30, 40, 模糊值 middle(ftl_2)的 4 个参数分别为 30, 40, 50, 60, 其它模糊值的梯形模糊数参数如表 11 中所列。

表 11 Typel 表

ftlid	alpha	beta	gamma	delta
ftl_1	0	16	30	40
ftl_2	30	40	50	60
ftl_3	50	60	75	90
ftl_4	50	110	150	160
ftl_5	150	160	170	180
ftl_6	170	180	190	200
ftl_7	38	38	42	42
ftl_8	177	180	180	183

表 12 存储了属性 ability 的模糊值,即语言常量值 bad,

fair, good, excellent。因为属性 ability 作用在离散域上,所以要给出属性值 bad, fair, good 及 excellent 之间的相似度。例如,属性值 bad(ftl_9)和属性值 fair(ftl_10)的相似度为 0.8。

表 12 Label 表

ftlid1	ftlid2	degree
ftl_9	ftl_10	0.8
ftl_9	ftl_11	0.5
ftl_9	ftl_12	0.1
ftl_10	ftl_11	0.7
ftl_10	ftl_12	0.5
ftl_11	ftl_12	0.8

表 13 存储了属性 height 的 Approx 数据类型值 &180 (ftl_8)中 margin 常量的值,这里常量 margin 的值为 3。

表 13 Approx 表

ftlid	margin
ftl_8	3

表 14 对属性 email 的 Posdis 数据类型值 {JS@yahoo.com/0.8, JohnS@msn.com/0.9} (ftl_13)进行了存储。

表 14 Posdis 表

ftlid	f_1	f_d1	f_2	f_d2	f_3	f_d3	f_4	f_d4	f_5	f_d5
ftl_13	JS@yahoo.com	0.8	JohnS@msn.com	0.9	null	null	null	null	null	null

5 正确性证明

利用关系数据库来存储模糊本体,形式上是将模糊本体模型转化为关系数据库模型,实现了从一个模式到另外一个模式的转换,因此要考虑模式转换的正确性。下面首先从信息容量^[20]角度,给出一种判断模式转换正确性的标准^[21-23],然后证明基于关系数据库的模糊本体存储方法的正确性。

定义 2 模式 S 所能表示的所有有效实例的集合称为模式 S 的信息容量^[20],记为 $I(S)$ 。

定义 3 设 S_1 和 S_2 是两个模式, $I(S)$ 表示模式 S 的信息容量,如果 $f: I(S_1) \rightarrow I(S_2)$ 是一个在 S_1 上定义的单射函数,那么称 f 是从 S_1 到 S_2 的保持信息容量的映射(尽管在理论上 f 可以采用不同的映射方法,但在本文中特指取相同实例的映射)。

定义 4 如果 $f: I(S_1) \rightarrow I(S_2)$ 是一个双射函数,那么称 f 是从 S_1 到 S_2 的保持等价性的映射。

定义 5 如果 $f: I(S_1) \rightarrow I(S_2)$ 是一个从 S_1 到 S_2 的保持信息容量的映射,那么称 S_2 通过 f 和 S_1 保持信息容量。

定义 6 如果 $f: I(S_1) \rightarrow I(S_2)$ 是一个从 S_1 到 S_2 的保持等价性的映射,那么称 S_2 通过 f 和 S_1 等价。

目前还没有一个统一的标准用于判断模式转换的正确性。理想情况下,希望模式转换是保持等价性的,但是,一般情况下,保持信息容量的模式转换就认为是正确的模式转换^[21-23]。

定理 1 利用图 4 所示的存储模式,模糊本体 FO 在关系数据库 R 中的存储是保持信息容量的存储。

要证明模糊本体在关系数据库 R 中的存储是保持信息容量的存储,也就是要证明模式 FO 到模式 R 的转换是保持信息容量的转换。首先,按图 4 所示的存储模式中模糊本体实例的存储方法,建立一个从 $I(FO)$ 到 $I(R)$ 的关系 $f: I(FO) \rightarrow$

$I(R)$ 。设 $\mathfrak{I} \in I(FO)$ 是模式 FO 的一个实例,即模糊本体的任意一个类的一组属性值,由 f 得到的 $f(\mathfrak{I}) \in I(R)$ 是将模糊本体的各个属性值以三元组的形式保存在表 $I_triplevalue$ 或表 $Fuzzytrivalvalue$ 中,其中表 $I_triplevalue$ 中存储精确值,表 $Fuzzytrivalvalue$ 中存储模糊值。形式上 f 定义如下(以表 $I_triplevalue$ 为例):

For $i=1$ to $m // m$ 个类

$c_i = \{p_i^1, p_i^2, \dots, p_i^n\} //$ 每个类有 n 个属性

For $k=1$ to $s //$ 每个类有 s 个实例

For $j=1$ to n

$f(\mathfrak{I})[iid_i^k] \leftarrow iid_i^k // iid_i^k$ 是第 i 个类的第 k 个实例的标识符,将其存储在字段 iid_i^k 中

$f(\mathfrak{I})[property_i^j] \leftarrow p_i^j //$ 将模糊本体属性 p_i^j 存储在字段 $property_i^j$ 中

$f(\mathfrak{I})[value_i^j] \leftarrow \mathfrak{I}[p_i^j] // \mathfrak{I}[p_i^j]$ 表示实例 $\mathfrak{I}$ 在属性列 p_i^j 上的分量,将其存储在字段 $value_i^j$ 中

证明 f 是一个函数。由 f 的定义可知,模糊本体的每个实例的每个属性值在表 $I_triplevalue$ 中都对应着一个元组,因为表 $I_triplevalue$ 的主键能够保证 $f(\mathfrak{I})$ 是模式 R 的一个有效实例(即一个元组),所以 f 是一个从模式 FO 到模式 R 的函数。

证明 f 是单射。设 $\mathfrak{I}_1 = (\mathfrak{I}_1[p_i^1], \mathfrak{I}_1[p_i^2], \dots, \mathfrak{I}_1[p_i^n])$ 和 $\mathfrak{I}_2 = (\mathfrak{I}_2[p_i^1], \mathfrak{I}_2[p_i^2], \dots, \mathfrak{I}_2[p_i^n])$ 是模糊本体类 c_i 的两个不同实例,则至少存在一个 $j(j \in \{1, 2, \dots, n\})$ 使得 $\mathfrak{I}_1[p_i^j] \neq \mathfrak{I}_2[p_i^j]$ 。由 f 的定义,表中存有元组 $f(\mathfrak{I}_1) = (f(\mathfrak{I}_1)[iid_i^1], f(\mathfrak{I}_1)[property_i^1], f(\mathfrak{I}_1)[value_i^1])$ 和 $f(\mathfrak{I}_2) = (f(\mathfrak{I}_2)[iid_i^1], f(\mathfrak{I}_2)[property_i^1], f(\mathfrak{I}_2)[value_i^1])$, 其中 $j=1, 2, \dots, n$ 。由至少存在一个 $j(j \in \{1, 2, \dots, n\})$ 使得 $\mathfrak{I}_1[p_i^j] \neq \mathfrak{I}_2[p_i^j]$, 可得至少存在一个 $j(j \in \{1, 2, \dots, n\})$ 使得 $f(\mathfrak{I}_1)[value_i^j] \neq f(\mathfrak{I}_2)[value_i^j]$, 因此 $f(\mathfrak{I}_1) \neq f(\mathfrak{I}_2)$ 。所以 f 是单射。

上面 f 的定义是以表 $I_triplevalue$ 为例。对于含有模糊值的实例在表 $Fuzzytrivalvalue$ 中的存储,其证明可用类似方法得到,在此不再赘述。

综上, $f: I(FO) \rightarrow I(R)$ 是一个从模糊本体 FO 到关系数据库 R 的单射函数,所以从模糊本体到关系数据的转换是一个保持信息容量的转换,即模糊本体在关系数据库中的存储是保持信息容量的存储,涵盖了模糊本体的结构和所有实例。

结束语 模糊本体被广泛用于描述语义 Web 上的模糊信息。为了更好地使用和管理模糊本体,需要对模糊本体进行合理、有效的存储。鉴于关系数据库的优势以及模糊本体的特点,提出了基于关系数据库的模糊本体的存储模式和相应的存储方法。理论证明,本存储方法是保持模糊本体信息容量的存储。

在未来的工作中,将对模糊本体的存储结果进行标准化表示和定量评价。此外,在模糊本体存储模式的基础上,我们将考虑对模糊数据类型更加细致地划分,从而不断改进和完善模糊数据的存储方法。

参考文献

- [1] Berners Lee T, Hendler J A, Lassila O. The semantic Web[J]. Scientific American, 2001, 284(5): 34-43

(下转第 245 页)

结束语 网络书写纹识别是一个具有研究价值与实际意义的方向,特别是针对汉语环境,目前还有许多问题需要进一步解决。针对 N-gram 字符特征具有高维、冗余等特点,提出一种基于特征空间划分的集成学习方法,并通过遗传算法对特征空间的划分进行优化。实验表明,该方法可有效提高网络书写纹识别的性能。下一步将就划分粒度对识别性能影响的理论依据以及特征空间划分的其他方法等问题展开进一步研究。另外,将充分利用集成学习与遗传算法的内在并行性,设计网络书写纹识别的并行算法,以提高其处理效率。

参考文献

- [1] Holmes D I. The analysis of literary style — A review[J]. Journal of the Royal Statistical Society, Series A, 1985, 148(4): 328-341
- [2] Abbasi A, Chen H. Applying authorship analysis to extremist-group web forum messages[J]. IEEE Intelligent Systems, 2005, 20(5): 67-75
- [3] 武晓春, 黄萱菁, 吴立德. 基于语义分析的作者身份识别方法研究[J]. 中文信息学报, 2006, 20(6): 61-68
- [4] Stamatatos E. A survey of modern authorship attribution methods[J]. Journal of the American Society of Information Science and Technology, 2009, 60(3): 538-556
- [5] Dietterich T. Machine-learning research: Four current directions [J]. AI Magazine, 1997, 18(4): 97-136
- [6] Goldberg D E. Genetic Algorithms in Search, Optimization and Machine Learning[M]. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 1989
- [7] 任江涛, 孙婧昊, 黄焕宇, 等. 一种基于信息增益及遗传算法的特征选择算法[J]. 计算机科学, 2006, 33(10): 193-196
- [8] Grefenstette J J, Gopal R, Rosmaita B, et al. Genetic Algorithms for the Traveling Salesman Problem[C]// Proc. 1st Int. Conf. Genetic Algorithms and Their Applications. Lawrence Erlbaum Ass., 1985: 160-168
- [9] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification [J]. The Journal of Machine Learning Research, 2008, 9: 1871-1874
- [10] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update[J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10-18
- (上接第 222 页)
- [2] Stoilos G, Nikos S, Stamou G, et al. Uncertainty and the semantic Web[J]. IEEE Trans on Intelligent Systems, 2006, 21(5): 83-87
- [3] Straccia U. Towards a fuzzy description logic for the semantic Web (preliminary report)[C]// Proceedings of 2nd European Semantic Web Conference, LNCS 3532. Heraklion, Greece, 2005: 167-181
- [4] Quan T T, Hui S C, Fong A C M, et al. Automatic fuzzy ontology generation for semantic Web [J]. IEEE Transaction on Knowledge and Data Engineering, 2006, 18(6): 842-856
- [5] Sanchez E, Yamanot T. Fuzzy ontologies for the semantic Web [C]// Proceedings of 7th International Conference on Flexible Query Answering Systems, LNAI 4027. Milan, Italy, 2006: 691-699
- [6] Lee C, Jian Z, Huang L. Fuzzy ontology and its application to news summarization[J]. IEEE Transactions on Systems, Man, and Cybernetics- Part B: Cybernetics, 2005, 35(5): 859-880
- [7] Ma Zong-min, Lv Yan-hui, Yan Li. A fuzzy ontology generation framework from fuzzy relational databases [J]. International Journal on Semantic Web and Information Systems, 2008, 4(3): 1-15
- [8] 李曼, 王琰, 赵益宇, 等. 基于关系数据库的大规模本体的存储模式研究[J]. 华中科技大学学报: 自然科学版, 2005, 33(sup.): 217-220
- [9] Zhou Jian, Ma Li, Liu Qiao-ling, et al. Minerva: a scalable OWL ontology storage and inference system[C]// Proceedings of 1st Asian Semantic Web Conference, LNCS 4185. Beijing, China, 2006: 429-443
- [10] Anuradha G, Cindy X C, Kajal T C, et al. From ontology to relational databases[C]// Proceedings of 2004 International Workshop on Conceptual Model, LNCS 3289. 2004: 278-289
- [11] Vysniauskas E, Nemuraite L. Transforming ontology representation from OWL to relational database[J]. Information Technology and Control, 2006, 35(3): 333-343
- [12] 许卓明, 黄永菁. 从 OWL 本体到关系数据库模式的转换[J]. 河海大学学报, 2006, 34(1): 95-99
- [13] 朱姬凤, 马宗民, 吕艳辉. OWL 本体到关系数据库模式的映射[J]. 计算机科学, 2008, 35(8): 165-169, 205
- [14] Barranco C D, Campana J R, Medina J M, et al. On storing ontologies including fuzzy datatypes in relational databases [C]// Proceedings of 16th IEEE International Conference on Fuzzy Systems. London, UK, 2007: 1-6
- [15] Lv Yan-hui, Ma Zong-min, Zhang Xu-hui. Fuzzy ontology storage in fuzzy relational database [C] // Proceedings of 6th International Conference on Fuzzy Systems and Knowledge Discovery. Tianjin, China, 2009, 2: 242-246
- [16] Hitzler P, Krotzsch M, Parsia B, et al. OWL 2 web ontology language primer [M/OL]. <http://www.w3.org/TR/2009/REC-owl2-primer-20091027/>, 2009-10-27
- [17] Klyne G, Carroll J J. Resource description framework (RDF): concepts and abstract syntax [M/OL]. <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>, 2004-02-10
- [18] Hayes P. RDF semantics [M/OL]. <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/>, 2004-02-10
- [19] Manola F. RDF primer [M/OL]. <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>, 2004-02-10
- [20] Hull R. Relative information capacity of simple relational database schemata[J]. SIAM Journal of Computing, 1986, 15(3): 856-886
- [21] Miller R J, Ioannidis Y E, Ramakrishnan R. The use of information capacity in schema integration and translation [C]// Proceedings of 19th International Conference on Very Large Data Bases. Dublin, Ireland, 1993: 120-133
- [22] Miller R J, Ioannidis Y E, Ramakrishnan R. Schema equivalence in heterogeneous systems: bridging theory and practice [J]. Information Systems, 1994, 19(1): 3-31
- [23] Qian Xiao-lei. Correct schema transformations [C]// Proceedings of 5th International Conference on Extending Database Technology (EDBT), LNCS 1057. Avignon, France, 1996: 114-128