基于语义的亲属关系知识模型建模设计与实现

方 杨[1],罗 军[2]
(1. 同济大学,上海 201804;
2. 重庆大学,重庆 400044)

摘 要:随着互联网的高速发展,用户获取、查找信息的信息源也是丰富多样。在错综复杂的社会人物关系背景下,人与人之间因血缘关系或者婚姻的结合而产生出了庞杂的亲属关系。对已知的亲属关系进行推理并得出未知的亲属关系的问题在社会生活的各个方面经常会遇到,然而传统的查询技术缺乏对语义的理解,不能表达语义信息,返回的结果往往都不能准确地满足人们的需求。语义网概念的出现旨在帮助人们更精确、更全面地从海量的信息中查询到所需的信息,通过一种机器可理解的形式,而本体作为一种共享概念模型的明确的形式化描述,为信息提供了语义表示机制。文中将本体这种能为信息提供语义表示机制的技术应用于亲属关系知识领域,实现了对亲属关系知识领域本体的深层次推理查询。

关键词:语义;本体;亲属关系;推理查询
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2019)05-0001-05
doi:10.3969/j.issn.1673-629X.2019.05.001

Modeling and Implementation of Kinship Knowledge Model Based on Semantics

FANG Yang[1], LUO Jun[2]
(1. Tongji University, Shanghai 201804, China;
2. Chongqing University, Chongqing 400044, China)

Abstract: With the rapid development of Internet, the information source which users obtain and find information are rich and diverse. Under the intricate social background of the relationship among people, the relationship of blood or marriage among people give birth to a variety of complex kinship. In all aspects of social life, we may encounter the problem: reasoning about known kinship relations and getting the unknown kinship. However, the traditional search technologies can't express semantic information because lack of understanding of semantics. So the results returned often cannot meet people's needs accurately. The concept of semantic web aims to make people more accurate and comprehensive to find the needed information in the form of machine understandable. Ontology, as a clear formal description of the shared conceptual model, provides semantic representation mechanism for information. In this paper, the ontology technology, which can be used to provide the semantic representation mechanism of information, is applied in the domain of kinship knowledge, and realized the deep semantic query and reasoning about kinship field.

Key words: semantics; ontology; kinship; reasoning query

1 概 述

1.1 背景介绍

随着互联网的迅速发展,人们已经迈入信息时代,如今在互联网上不但聚集了海量的信息,而且这些信息的数量还在以指数级的速度增长。面对愈发庞大的信息量,仅仅依靠基于关键字的匹配查询技术会出现许多问题,诸如:检索结果过多、精度较低,无法智能获取隐含的知识以致查询结果常常与初衷相去甚远。因此用户只能靠自己浏览才能筛选出自己需要的信息,这就使得当前的网络查询技术已不能满足人们的需求。这些问题形成的主要原因在于:机器对信息是不能理解的,因而不能进行基于语义的查询。

收稿日期:2018-06-25 **修回日期**:2018-10-11 **网络出版时间**:2018-12-21
基金项目:国家自然科学基金(61672118)
作者简介:方 杨(1994-),女,在读研究生,研究方向为深度学习;罗 军,副教授,研究方向为网络及数据库、大型 MIS 系统建模及设计、基于数据库的应用系统平台的架构。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20181221.1625.064.html

语义网这个概念最先由万维网之父 Tim Berners-Lee 在 1998 年提出[1],其中本体[2]作为一种能在语义和知识层次上描述信息的概念模型建模工具[3],扮演了十分重要的角色。它可以描述某特定领域的知识并且可以实现对知识的分类,还能支持逻辑推理[4]。基于这些特点,本体可以更好地描述信息并挖掘出信息中隐含的知识,因而在科研教育、医疗保健、政府管理、企业、金融以及家庭中都有着广泛的应用前景。

在错综复杂的社会人物关系背景下,亲属关系是社会生活中一个重要的组成部分。汉族是一个智慧的民族,尽管社会人物关系错综复杂,但将亲属关系划分得条理分明,称谓也是尊卑有序:对长辈的称谓有伯伯、叔父、姑姑、舅舅、姨母,对同辈的称谓有哥哥、姐姐、弟弟、妹妹,对晚辈的称谓有侄子、外甥等等。另外,由于婚姻和血缘的联系而产生的庞杂的亲属关系也派生出了各种特定的亲属称谓,比如对各个辈分的称呼以及亲属称谓里都有着比较重要的限定词,如“堂”和“表”,以此来区分其他稍远的关系。这些都是中华文化的一大特征。亲属关系不仅与人类日常生活密不可分,也是人类社会的文化选择。

在日常生活的各个方面可能都会遇到这样的问题:已知 a 是 b 的伯母,c 是 b 的父亲,问 a 与 c 是什么关系? 但由于亲属关系表达形式多样化以及随着数据规模的提升,人与人之间关系复杂性的提高,往往很难理清这些复杂的人物关系。

比如在人口数据分析方面,在户籍、计生、公安、医院、民政、银行等各部门都建立了人口数据库,其中包含了诸如姓名、性别、出生日期等人物的基本信息,以及人物简单的亲属关系。但如果发生拆户事件,或者随着户中人员的迁移,那么原本具有亲属或血缘关系的人口群体就会逐渐分离,而在后台数据库中关于人物亲属关系的记录一般只有最近邻的父母或子女的较少的亲属信息,因此如果要得知更复杂的亲属关系信息(例如需查询“祖父”)就会相当困难,基于数据库的查询就难以满足用户的需求。

再比如,对于近亲联姻的问题,由于近亲联姻所生育的下一代遗传病发病率会相应增高,并且死亡率也比一般婚配要高的多。因此按照法律规定:具有直系血亲或三代以内旁系血亲关系的男女之间不能联姻。禁止近亲联姻是依据科学原理所采取的必要措施,如果能高精度高效率地查询出近亲的范围将能大大减小近亲联姻的概率,对人类社会将具有极大的意义。

因此,能够进行亲属关系的语义查询,实现对亲属关系的相互推导是信息化时代的必然要求,可以在诸如人口数据分析统计、家族追溯、医学遗传分析、公安刑事案件侦破等各方面为分析和应用人口之间的亲属关系提供有效的数据支撑。

1.2　研究现状

在已有的国内外关于亲属关系领域的研究中,有研究亲属关系专家系统[5]的,也有人设计了亲属关系推理模型并对亲属关系知识以及人物信息的表示与存储进行了研究[6]。在国外,因为文化不同,语言中表示亲属关系的词汇没有中国划分得那么细致,因此相对中国来说要少得多[7]。在国外,有对特定亲属范围内的一个家族的人物关系网进行的研究,主要研究知名人物的历史或家族。在国内,对家族人物关系的研究较多,比如有人对《红楼梦》建立了相关人物关系的专家问答系统[8],不过该研究针对的是特定家族[9-10],不具有普遍性,知识的表示能力也比较单一。

总的来说,在已有亲属关系的研究中,最常见的就是构造大型的亲属关系知识库,其中包括了若干类的知识库,比如亲属关系名词库、性质判断库、逆判断库、传递判断库等,当需要对亲属关系进行推理时,就根据需要查询相应的知识库,翻译成相应的表达式,然后对相关表达式进行合一,从而得出相应的关系。再采用一些亲属关系转换算法、化简运算等操作得到最简表达式,最后检索知识库得出答案。除此之外,还有一条重要的方法是:选取若干语义特征,用这些语体特征来定义任一亲属关系,即逻辑表达式,在进行推理时,把已知亲属关系的逻辑表达式联结起来,然后使用辅助运算规则找出表达式中蕴含的信息,再使用化简运算规则把表达式转化为最简表达式,最后检索定义库从而得出答案[11]。上述方案在推理过程中都大大增加了运算量,在程序实现上也增加了难度,并且都没能实现与本体技术相结合,不能进行基于语义推理查询。

2　亲属关系知识本体模型

2.1　亲属关系领域相关知识

亲属关系是一种社会关系,它是基于婚姻、血缘和法律拟制而形成的。现代汉语常见亲属关系的认知模型中有大约两百多个亲属关系名词,可以归入大约 70 余种亲属关系,可分为父党、母党、妻党、夫族、子族、女族、兄弟族、姐妹族八个部分。

根据亲属关系发生的原因以及国内司法部的解释,亲属的具体范围包括:配偶、血亲、姻亲。

配偶:指夫妻双方因婚姻的结合而发生,是产生其他亲属关系(血亲、姻亲)的基础。

血亲:包括直系血亲和旁系血亲,直系血亲是指和自己有直接血缘关系的亲属,具有生育与被生育关系,源于父系和母系或同源于父母双方,包括生育自己父母及生育父母的祖父母、外祖父母以及更上的长辈,和自己生育的子女及子女生育的孙辈以及更下的晚辈;

旁系血亲指的是除直系血亲以外的、与自己同出一源的、与自己具有间接血缘关系的血亲。

姻亲:是因婚姻的结合而产生的亲属。具体又分为 3 类:

(1)血亲的配偶:指自己直系血亲或旁系血亲的配偶,如伯母、嫂嫂等。

(2)配偶的血亲:指自己配偶的直系血亲或旁系血亲,如岳父、大舅子等。

(3)配偶的血亲的配偶:指自己配偶的直系血亲或旁系血亲的丈夫或妻子,如伯嫂、大舅妇等。姻亲关系会因夫妻离婚或他方再婚而消失。

2.2　亲属关系领域本体的设计

本体是对某个领域知识的形式化的规范说明,是对客观存在的概念和关系的描述[12]。利用本体进行关于个体的推理,可以将隐含在信息中的知识表达出来。从知识共享的角度看,本体可以看作是感兴趣领域的形式化的规范说明,是对客观存在的概念和关系的描述。本体有 4 个重要的组成部分,分别如下:

(1)个体:可以理解成一个类的实例,或特定领域中感兴趣的对象。

(2)属性:代表个体之间的二元关系,它连接着两个个体。在本体语言中有两种主要的类型属性:对象属性和数据属性。其中对象属性连接两个个体,代表了个体之间的一种关系;数据属性连接的是个体和数据类型值或字面量,表示个体在某属性上具有的数据值。此外,属性还具有传递、对称、非对称、函数、反函数、逆、自反、非自反八大特性。

(3)类:一组包含了个体的集合,这个集合里的个体资源具有相似的特性。类具有超类和子类的层次结构,子类是相比超类对资源更细致的划分,因此子类继承了超类的性质。

(4)类公理:子类、等价类、不相交是类的三大公理,子类描述的是父类与子类的关系;等价类和推理密切相关;不相交则是为了让推理能够顺利进行的一个必要条件,没有它的显式声明,有些推理将无法运行。

由 2.1 可知,中国人家族关系的特点是类少,关系复杂,因此具体设计方案如下:

(1)类的设计。

三个非常重要的类:“人”,“男人”,“女人”。“人”包括且仅包括“男人”和“女人”,而且“男人”和“女人”不相交。

(2)属性的设计。

由于亲属关系如“父亲”、“母亲”等和人物有关,具有相对性,而且随着时间而变化,故采用属性建模比较合适。亲属关系中属性的数量很多,而且属性之间的关系复杂。由于对象属性能将两个个体联系起来,代表了个体之间的一种关系,因此亲属关系可以用对象属性来表示。亲属关系名词可以看作是一个逻辑谓词,反映的是称呼人与被称呼人之间的亲属关系,这里属性的命名原则是:“x 的属性是 y”,例如属性“父亲”,定义为:“father(a,b)”,个体 a 代表称呼人,个体 b 代表被称呼人,即读作“a 的父亲是 b”。有些属性具有函数性、对称性或传递性,有些属性之间具有互逆关系,有些属性之间具有层次关系。部分亲属关系属性的特性及约束定义举例如表 1 所示。

表 1　亲属关系对象属性创建举例

关系名称	定义域	值域	属性性质
丈夫	女人	男人	逆(与妻子互逆) 函数性 反函数性 反对称性 非自反性
祖父	人	男人	函数性 非自反性 反对称性
姐姐	人	女人	非自反性 反对称性 传递性

在属性创建的同时,还应考虑到属性之间的层次关系,比如属性“父亲”和“母亲”是属性“父母”的子属性,“父母”又是“直系血亲”的子属性,“直系血亲”又是“血亲”的子属性;而“血亲”是“直系血亲”和“旁系血亲”的超属性。

用对象属性来表示个体之间的亲属关系以后,若要进行亲属关系之间的推理,还需要考虑个体的性质,比如:个体的性别;个体的出生日期,它表示了个体之间的出生先后,这样就能在同辈个体之间关系不确定的情况下判断出具体的关系。出生先后关系满足传递性。

因此还需创建两个数据属性“生日”、“性别”,这两个属性的定义域都是类“人”,值域是数据类型,其中生日是整形数据类型,性别是字符串数据类型,即:male/female。并且属性“生日”和“性别”都是单值的,即一个人只具有一个生日和性别,因此都具有函数性。

2.3　亲属关系知识模型推理规则的设计

婚姻与生育是一切亲属关系形成的核心,母亲生下孩子,但母亲不能单独生育,因此最核心的亲属概念有两个,即“母亲”和“丈夫”。从认知角度看,当暂不考虑一夫多妻、离异、再婚等改变亲属关系的因素时,所有亲属关系都可以用核心亲属概念来定义,“母亲”和“丈夫”是最核心的,其次是“父亲、妻子、女儿、儿子”4 个,它们由“母亲”和“丈夫”直接转化而来。再次就是“兄、弟、姐、妹”,他们是“父亲”或“母亲”的

“儿子”或“女儿”,由 2 个亲属关系构成。

而实际上,亲属关系的范围非常庞大,如:父亲、祖父、曾祖父……但是任何亲属可以分解成有限个最小亲属元的序列,这些亲属元不可再分。考虑到如果亲属基元很少,亲属关系的定义就会很复杂,因此文中将“父、母、夫、妻、子、女、兄、弟、姐、妹”这 10 个亲属关系视作亲属基元,其他亲属关系都可以看作是这 10 个亲属基元的扩展。因此如果把所有亲属关系看作成网络中的节点,则亲属基元就是半径为 1 的节点(即一维的节点);需要用 2 个亲属基元表示的亲属关系看作是半径为 2 的节点(即二维的节点),如:伯父可以表示成“父亲的哥哥”。而需要用 3 个亲属基元表示的亲属关系看作半径为 3 的节点(即三维的节点),如伯母可以表示成“父亲的哥哥的妻子”,以此类推。亲属关系基于基元的定义举例如表 2 所示。

表 2　亲属关系定义式举例

定义项	定义(亲属基元表达式)
祖父(x,y)	父亲(x,x_1)∧父亲(x_1,y)
姑姑(x,y)	父亲(x,x_1)∧[姐姐∨妹妹(x_1,y)]
舅舅(x,y)	母亲(x,x_1)∧[哥哥∨弟弟(x_1,y)]
表妹(x,y)	[母亲(x,x_1)∧[哥哥∨弟弟∨姐姐∨妹妹(x_1,x_2)]∧女儿(x_2,y)]∨[父亲(x,x_1)∧[姐姐∨妹妹(x_1,x_2)]∧女儿(x_2,y)]
姨父(x,y)	母亲(x,x_1)∧[姐姐∨妹妹(x_1,x_2)]∧丈夫x_2,y)

在表 2 定义式中,联结词“∧”、“∨”分别代表“并”和“或”。

因此,要实现对任意一对个体之间亲属关系的推理,就需要构造一套完备的运算推理规则。文中涉及到的亲属关系的运算有逆运算和复合运算。

(1)亲属的逆运算。

比如:已知 A 是 B 的 R 亲属,假设 B 是 A 的 S 亲属,求 R 的逆(S)。设 B 的性别 $S_B \in \{Male, Female\}$,以亲属基元为例,Reverse(S_B,R)函数完全可以用列表法枚举,如表 3 所示。

表 3　亲属基元逆运算

	丈夫	妻子	父亲	母亲	儿子	女儿	兄	弟	姐	妹
男人		丈夫	儿子	儿子	父亲	父亲	弟	兄	弟	兄
女人	妻子		女儿	女儿	母亲	母亲	妹	姐	妹	姐

(2)亲属的复合运算。

比如:已知 B 是 C 的 R_1 亲属,A 是 B 的 R_2 亲属,假设 A 是 C 的 S 亲属,求 S,即求 R_1 与 R_2 的复合 R_1R_2。设 A、C 间长幼关系 $G_{A,C} \in \{Elder, Younger\}$。表 4 为亲属基元复合运算的结果,其中第 1 列为 R_1 的取值,第 1 行为 R_2 的取值,复合值 R_1R_2 不是亲属基元的暂用“–”表示。

表 4　亲属基元复合运算

	丈夫	妻子	父亲	母亲	儿子	女儿	兄	弟	姐	妹
丈夫			—	—	儿子	女儿	—	—	—	—
妻子			—	—	儿子	女儿	—	—	—	—
父亲		母亲	—	—	兄/弟姐/妹		—	—	—	—
母亲	父亲		—	—	兄/弟姐/妹		—	—	—	—
儿子		—	丈夫	妻子	—	—	儿子	儿子	女儿	女儿
女儿	—		丈夫	妻子	—	—	儿子	儿子	女儿	女儿
兄		—	父亲	母亲	—	—	兄	兄/弟	姐	姐/妹
弟		—	父亲	母亲	—	—	兄/弟	弟	姐/妹	妹
姐	—		父亲	母亲	—	—	兄	兄/弟	姐	姐/妹
妹	—		父亲	母亲	—	—	兄/弟	弟	姐/妹	妹

在上述生成的规则中,大部分会涉及到对性别和年龄的判定,因此,这里还需要定义一条重要的规则来根据人物的出生日期判断长幼关系。由 2.2 可知,有些属性具有对称性或传递性,并且有些属性之间具有互逆关系,有些属性之间还具有层次关系。除了对具体亲属关系进行推理查询外,还可能会对亲属关系的范围进行推理查询,因此还需定义公理规则来表达属性层次间的关系以及属性之间的互逆特性。有了不同维度亲属关系的定义,求逆、复合规则以及公理规则,这样便可生成一套完备的并具有丰富语义的规则集作为亲属关系知识领域的知识库,基于这个知识库可以实现对任意两个个体之间亲属关系的推理。

3　亲属关系知识模型推理查询的实现

3.1　实验环境

文中首先使用本体编辑工具 Protégé 构建亲属关系知识本体,然后采用 Pellet 推理机对本体进行一致性检查[13]和推理。并结合 Eclipse 开发平台以及本体开发工具 Jena[14]框架,建立了一系列推理规则作为本体推理的基础[15],从而实现对亲属关系本体模型的解析、查询和推理。

3.2　亲属关系推理查询

(1)查询 b 和 m 之间的关系。已知 a 的父亲是 b,a 的伯母是 m,则应该推出 m 是 b 的嫂嫂,查询结果如图 1 所示。

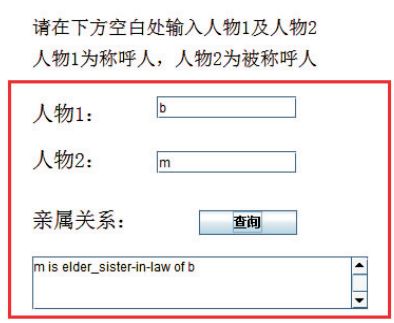

图 1　查询 b 和 m 之间的关系

(2)查询 a 的外祖母是谁。已知 a 的父亲是 b,b 的妻子是 d,d 的母亲是 f,故应该推出 a 的外祖母是 f,查询结果如图 2 所示。

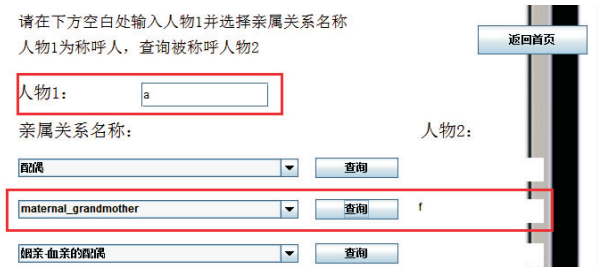

图 2　查询 a 的外祖母

(3)查询 a 的直系血亲有哪些,查询结果如图 3 所示。

图 3　查询 a 的直系血亲

已知的是 a 的父亲是 b,爷爷是 c,儿子是 h,推出的是 a 的母亲是 d,外祖母是 f,由父亲、爷爷、儿子、母亲、外祖母属于直系血亲范畴,故 a 的直系血亲有 d,f,b,c,h。

4　结束语

从语义的角度入手,将自提出以来就引起国内外众多科研人员关注的本体技术应用于亲属关系知识领域,在对亲属社会关系深入了解的基础上,通过构建亲属知识本体以及利用亲属称谓之间存在的相互推导的规律,实现了对亲属关系领域包括配偶、血亲、姻亲范围内的基于语义的推理查询,并验证其正确性。

可以得出,本体作为一种共享概念模型的明确的形式化描述,为信息提供了语义表示机制,并且具有如下优势:可以实现把具有相同特性的数据进行分类、定义、约束,从而减少了查询结果的重复,并且不会遗漏,使查询结果更为全面、准确;本体的推理功能可以推理出隐含的知识,实现基于语义的查询,从而可查询出与查询条件具有相同语义的信息,提高了查准率、查全率。因此相比数据库系统这种传统的对知识和数据管理的方法,本体更加关注语义,从而可以帮助人们更准确、更全面地对知识进行推理查询。

中国从传统社会以来就十分注重血缘的联系,具有亲属关系的人们往往也生活在同一个地方,因而也有了家谱、族谱的概念,而家谱、族谱也是中国传统文化中一个重要的组成部分。然而随着社会的发展,社会结构发生了变化,原来聚居在一起生活的人口也逐渐发生了迁移,对家谱、族谱的概念也逐渐淡化。如今的互联网技术虽然不能将具有亲属关系的人们从地理位置上联系起来,但是可以将人们从逻辑意义上联系起来。而将语义网中的本体技术应用于亲属关系知识领域,对亲属关系领域进行更深层次的研究,并利用亲属关系之间存在的相互推导的规律,实现对亲属关系领域基于语义的推理,从而准确地提取亲属关系知识。这将在人口领域、家族关系分析和社会管理等社会生活的各个方面具有极大的现实意义。

参考文献:

[1] STUDER R,BENJAMINS V R,FENSEL D. Knowledge engineering,principles and methods[J]. Data and Knowledge Engineering,1998,25(1-2):161-197.

[2] 杜小勇,李　曼,王　珊. 本体学习研究综述[J]. 软件学报,2006,17(9):1837-1847.

[3] CHANDRASEKARAN B,JOSEPHSON J R,BENJAMINS VR. What are ontologies,and why do we need them? [J]. IEEE Intelligent Systems and Their Applications,1999,14(1):20-26.

[4] 万　捷,滕至阳. 本体论在基于内容信息检索中的应用[J]. 计算机工程,2003,29(4):122-123.

[5] PATTERSON D W. Introduction to artificial intelligence and expert systems[M]. Upper Saddle River,NJ,USA:Prentice-Hall,Inc. ,1990.

[6] 葛　强. 亲属关系逻辑推理专家系统的研究[D]. 郑州:河南大学,2005.

[7] 夏孝才. 亲属关系词中的文化差异[J]. 湖南大学学报:社会科学版,2001,15(2):154-155.

[8] 王树西,刘　群,白　硕. 一个人物关系问答的专家系统[J]. 广西师范大学学报:自然科学版,2003,21(1):31-36.

[9] 王树西,刘　群,白　硕. 自然语言界面的专家系统的研究[J]. 计算机工程与应用,2003,39(17):35-37.

[10] 郑实福,刘　挺,秦　兵,等. 自动问答综述[J]. 中文信息学报,2002,16(6):46-52.

[11] 陈振宇,袁毓林. 汉语亲属关系的语义表示和自动推理[J]. 中国语文,2010(1):44-56.

[12] 马文峰,杜小勇. 领域本体进化研究[J]. 图书情报工作,2006,50(6):71-75.

[13] 许　勇,王智学,李宗勇. 领域本体的一致性检查[J]. 计算机工程,2009,35(1):55-57.

[14] MCBRIDE B. Jena:a semantic web toolkit[J]. IEEE Internet Computing,2002,6(6):55-59.

[15] 向　阳,王　敏,马　强. 基于 Jena 的本体构建方法研究[J]. 计算机工程,2007,33(14):59-61.