基于属性坐标系框架下的 Freebase 语义库研究

王 斌,冯嘉礼

(上海海事大学 信息工程学院,上海 201306)

摘 要:属性坐标系是由 n 个不相关的属性组成的一个 $n-1$ 维坐标系。第 $n+1$ 个属性,可以由这 n 个属性做合取运算得到;这样在这个 $n-1$ 维坐标系中,就形成一个唯一点来表示这第 $n+1$ 个属性。2007 年 Freebase 数据库的建立,使得该属性坐标系理论得以验证,并为属性坐标系的建立提供可能,而且将在语义相关度计算中发挥重要作用。

关键词:属性坐标系;Freebase;语义相关度计算

中图分类号:TP311.13 **文献标识码**:A **文章编号**:1674-7720(2013)21-0059-03

Further study in Freebase within property coordinate system theory

Wang Bin, Feng Jiali

(College of Information and Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: The property coordinate system is a $n-1$ dimensional coordinate system builded up by n unrelated properties. In order that a new property (or an entity)could be located in the coordinate system at a unique coordinate point by performing logical conjunction on all (or part) of the properties. Freebase, which is available since 2007, has approved the property coordinate system theory, and makes the theory realizable, which will take advantage in semantic relevancy computing process.

Key words: propetry coordinate system; Freebase; semantic relevancy computing

哲学上讲,事物的质是事物的内在规定性,是区别于其他事物的规定性;质通过属性表现。在此引入集合论的概括公理可知:任给一个属性,存在一个由所有具有该属性的元素构成的集合。若引入特征公理可知:任何两不同事物,至少存在一属性,使得它是一事物区别于另一事物的特征。可证明:一个事物(或系统)可由它所具有的特征予以确定(或定义)。此处,特征包括事物的关系和结构。可见,"给定一(组)特征可确定其对应(或定义)的事物"是人类识别事物的基本原则[1-2]。

1 属性论及属性坐标系

在"属性论"的原则下,n 元关系 $R(x_1,x_2,\cdots,x_n)$ 可定义为序集 $A=(x_1,x_2,\cdots x_n)$ 作为一个整体而具有的某一个属性 $P_R(A)$,例如通常将对称关系 $S(a,b)$ 看做是序集 $A=(a,b)$ 所具有的对称性 $P_S(A)$;而对三元关系,直线 a 和 b 平行并且和 c 相交 $R(a,b,c)$ 则可看做是序集 $B=(a,b,c)$ 所具有的前两个元素 a 和 b 平行而与 c 相交的属性 $P_R(B)$。这种将论域 X 上的 n 元关系 $R(x_1,x_2,\cdots,x_n)$ 定义为幂集 $X(X)$ 中的一个点 $A=(x_1,x_2,\cdots x_n)$ 相应属性 $P_R(A)$ 的方式,可很方便地将谓词定义为个体的属性或个体序集的属性。故谓词逻辑演算等同于属性演算[3]。

1.1 属性集 PX 的基本结构

定义 1 设论域 X 上的幂集 $X(X)$,$P_A=\{P_i(A)\mid i\in I\}$ 是 X 的有序子集 $A=(x_1,x_2,\cdots x_n)$ 的属性集,$P_i(A)$ 表示集 A 的第 i 个类属性,I 是指标集。令 $P(A)=\bigwedge\limits_{i\in I}P_i(A)$ 为集 A 的合属性,$\wedge$ 为合取运算,故 $P(A)$ 是 A 作为一个整体时的内涵属性。同理,若 $A=\{x\}$ 是单元集,则 $P_A=\{P_i(A)\mid i\in I\}$ 是个体 x 的属性集,$P_i(A)$ 表示 x 的第 i 个属性,而 $P(x)=\bigwedge\limits_{i\in I}P_i(x)$ 为 x 的合属性。

例如:设 X 为植物集,则其类属性 $P_i(X)$ 可认为是通常所说的"光合作用"、"根茎叶"等,而合性质 $P(X)$ 则是广义的"植物性";X 的子集 $A=\{$种子植物$\}$,则其类属性 $P_i(A)$ 则认为是"维管组织"、"开花结果"、"种子繁殖"等,其合属性 $P(A)$ 是"种子植物门";X 的个体 x,比如蒲公英,其单属性 $P_i(x)$ 例如"种子上有绒毛"以及其

他药用、食用特点，其合属性 $P(x)$ 即是“蒲公英科”。

可证明，事物 x 的属性集 P_x 连同合取 $\wedge$ 运算，构成可交换幺半群 $M(P_x, \wedge)$[3]。

1.2 属性集 P_x 的几何模型

由上文得到，P_x 连同合取 $\wedge$ 运算可构成幺半群 M，故所有属性可分为两类：(1) 由其他属性 $\wedge$ 运算得到的合属性；(2)不能由其他属性生成的属性。

定义 2 不能由其他属性合取得到，且不可再分的属性称为素属性。记为 P_x^*，则有 $P_x=P_x^* \cup (P_x^*)$，且幺半群 M 可由 P_x^* 生成，即 $M(P_x, \wedge)=<P_x^*>$。设素属性及 P_x^* 有 $n+1$ 个元素，即 $P_x^*=\{P_j | j=0, 1, 2 \cdots n\}$。其中 $P_0=1$，称为平凡素属性。

定义 3 属性多面体 K 及其重心剖分 $K^{(m)}$。设 n 维单纯形 $K=(P_0, P_1, P_2 \cdots P_n)$ 的顶点为素属性集 P_x^* 的 $n+1$ 个元素，即构成个体 x 的属性多面体。当第一次剖分 $K^{(1)}$，这 $r+1$ 个素属性构成的 r 维单纯形 $T_j^r=(P_{j0}, P_{j1}, \cdots P_{jr})$，其重心剖分点即为该 $r+1$ 个素属性的合属性。这里，j 为单形 T_j^r 作为单形 K 中的 r 维子单形的序号，$0 \leqslant r \leqslant n$。同理，在 $K^{(2)}$ 剖分中，l 个属性构成的单形 $S_i^l=(P_{i0}, P_{i1}, \cdots P_{il})$，其重心剖分点即为该 l 个属性的合属性。P_{il} 可能为素属性，也可能为前次剖分形成的合属性。按同样的方法，继续剖分 K 直到第 m 次剖分 $K^{(m)}$，使得第 $m+1$ 次剖分 $K^{(m+1)}$，形成的新的重心属性，不能在属性集 P_x 中找到对应元。

例如：群性质 G 由四个群公理 $\bigwedge_{i=0}^{4} G_i$ 所确定，由此，可将 G_i 看作是 G 的四个素属性（虽然 G 还有其他一些逻辑公理，在此暂不考虑）。由此可得到由 G_i 组成的三维多面体 $K_3(G)=(G_1, G_2, G_3, G_4)$。如图 1，第一次剖分，半群性 S 是一维子单形 $T_1^1=(G_2, G_3)$ 的重心剖分点上。以此类推，幺半群性 M 位于二维子单形 $T_2^2=(G_2, G_3, G_4)$ 的重心剖分点上；群性 G 落在该单形的重心剖分点上[2]。

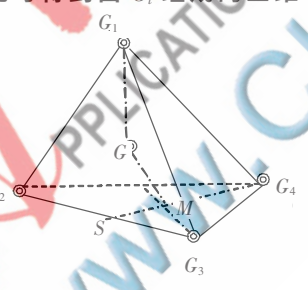

图 1 群性质 G 的重心剖分模型

2 语义数据库 Freebase

Graphd 与关系数据库以表的形式存储数据完全不同，Graphd 以节点以及节点之间的关系所形成的图结构来组织数据，以数组的方式对节点和其关系的元数据进行建模，以表格形式存储，表格中的每条数据对应一个节点关系数组，数组由源节点、属性、目标节点、源节点值组成。使用 MQL 语言作为查询语言，并通过 HTTP 标准的“请求/应答”机制发送请求[4-5]。

2.1 Freebase 知识表示和组织机制

Freebase 的结构分为 3 层：Domain->Type->Topic。以 Arnold Schwarzenegger 为例，解释 Freebase 中的知识结构。如图 2 所示。其中，椭圆框表示 Topic，方形框表示 Type[6]。

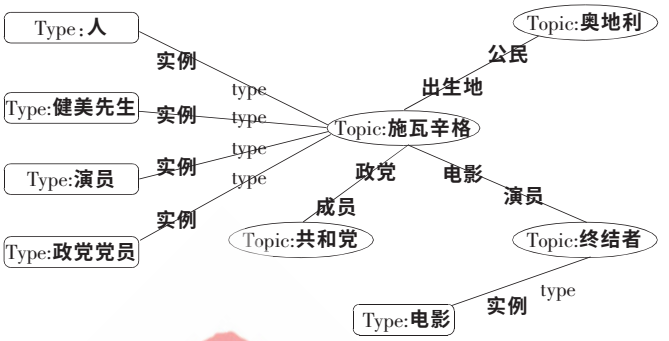

图 2 Freebase 中知识结构示例

以 Arnold Schwarzenegger 为原点进行讨论，首先是一个 Topic，对应于现实中的一个对象。它有 4 个 Type（可理解为定义）：Person、Body Builder、Actor、Politician。

Type：Person 下有一个属性：country of birth，其值为 Topic：Austria。这样，就把 Arnold Schwarzenegger（对象）与 Austria（对象）建立了联系。以此类推。

同样反过来看 Topic：Terminator 有一个 Type：File，其下有一个属性 cast，其值为：Schwarzenegger。这样反向也建立了联系[6]。

2.2 MQL 查询语言

由于 Freebase 后台使用自己设计的 Graphd，所以也摒弃了传统的 select 等 SQL 语言。设计 GQL（graph query language）作为请求处理语言；为了用户使用方便，在此基础上，设计 MQL（Metaweb Query Language）作为数据查询语言。其完全符合 HTTP 协议标准的“request/response”机制，可在浏览器地址栏，直接输入 MQL 查询语句。例：

https://api.freebase.com/api/service/mqlread?query={"query":{"type":"/music/artist","name":"The Police","album":[]}}

该例可在地址栏中直接输入并在页面中返回结果[7]。

3 属性坐标系框架下的 Freebase 研究

通过对 Freebase 的数据组织机制、知识表示的再研究，发现其暗含了属性坐标系的建库理念。从侧面证明了“属性论”及“属性坐标系”等理论在知识处理领域的独到性、前瞻性和实用性。

3.1 Freebase 中对对象的定义

还是以 Arnold Schwarzenegger 为例，对人脑思维来说，首先该对象是一个人，对应在 Freebase 存在一个 Type：person；以此类推其他（比如 Type：actor）。

如一个谜语：有一个人，是健美先生，是演员，演过《终结者》，这个人是谁？如果人们脑中有这个定义，马上就能得到谜底，也就是该对象 Arnold Schwarzenegger。

综上所述事实，恰好印证了属性论中阐述：一个事物（或系统）可由它所具有的特征予以确定（或定义）；

人脑感觉则仅对其敏感的事物属性作出反应。用属性坐标系表示，如图3所示。

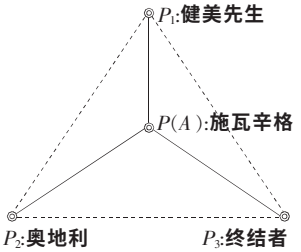

图3 对象属性坐标系表示

不难看出 $P(A)=P_1 \wedge P_2 \wedge P_3$。这正是人们人脑中的反应。

3.2 对象属性坐标的建立及相关度计算

Freebase 在对象的定义、上下位关系的定义过程中有独到的见解，提供了一个有效地途径。比如 Arnold Schwarzenegger 对象和 Austria、Terminator、Republican 等对象之间的关系，不难看出，这些对象之间有明显的包含和被包含关系。

但是，同级对象之间的关系比较，比如 Arnold Schwarzenegger 和 Sylvester Stallone 之间关系的定义，Freebase 就稍显不足。可不妨换一种思路，通过建立对象的属性坐标，对象坐标做笛卡尔积，从而得到两个同级对象之间关系。

首先建立 Arnold Schwarzenegger 的对象坐标，根据上文的图示坐标可得（该对象有很多属性，但该部分属性已可以定义该对象）：

Arnold Schwarzenegger=[
{person [{nationality:Austria,USA}…]},
{actor[{film:Terminator},…]},
{bodybuilder[]},…
] ……$\vec{V}_1$

这里采用二级坐标表示：person 表示对象特征定义；nationality 表示特征属性；Austria 表示特征属性值。因为 nationality 等一些列特征属性合取可得到 person 这个定义。

再来建立 Sylvester Stallone 的对象坐标，根据前文的构造语句，可以得到 Stallone 的 Type 及属性。

Sylvester Stallone=[
{person[{nationality:USA},…]},
{actor[{film:The First Blood},…]…},
] ……$\vec{V}_2$

然后将两个向量相乘，首先：

person*person=1

表示这两个对象有特征定义 person，可以进一步比较特征属性 nationality：

Austria*USA=0

USA*USA=1

综上可认为：nationality*nationality=0.5

表示，这两个对象的特征属性 nationality 有部分相关，都有 USA 这个特征属性值。

其次，person*actor=0，表示这两个对象的特征定义不存在相关关系。

再次，actor*actor=1，表示这两个对象有相关的特征定义 actor，进一步比较特征属性 film，由前文提供计算方法可得 film*film=0，表示没有特征属性值的相关关系。

综上计算方法及结果，可得如下结论：Arnold Schwarzenegger 和 Sylvester Stallone 两个对象存在相关关系。即：这两个对象是 person，国籍都是 USA，都是 actor，但是没有合演过电影。这个结论与现实相符。

前文所述的对象相关度计算方法，仅是表面上粗浅的计算，也仅是对语义相关度问题提供一种方法参考，但计算结果让人满意，实际意义值得期待。而且完全可以对该方法进一步加以研究，比如：每一个对象按特征定义的贡献度，为特征定义赋予相关权值；细化计算粒度，可以用小数表示相关度；为结果定义实际意义，例如两个向量的乘积等于1有什么意义，等于0.5时有什么意义。

人工智能发展至今，最根本的问题是知识的表示问题。各种知识库如 WordNet、Freebase 数据库、各种本体数据库，都力求将大千世界中的对象有机组织起来。但是，不管是本体数据库还是 Freebase 数据库，都是利用了对象之间天然的上下位关系进行层次关系定义，而且经验证，这种方法是可行的，但仅是在纵向关系研究中。但是属性坐标系理论以其独到性，提出了一种对象横向关系（相关性）的表示方法，对本体数据库或是 Freebase 数据库中的对象横向关系研究提供了宝贵的理论基础。

参考文献

[1] 冯嘉礼.思维智能与属性论方法[J].广西师范大学学报（自然科学版），1997，15(3)：1-6.

[2] 冯嘉礼.思维智能与属性论方法（续）[J].广西师范大学学报（自然科学版），1997，15(4)：1-6.

[3] 冯嘉礼，冯嘉仁，詹增修.以属性为基础的知识库建库原则[J].计算机研究与发展，1987，24(11)：56-61.

[4] 李俊.语义数据库 Freebase 研究[J]，现代图书情报技术，2011(10)：18-23.

[5] A Brief Tour of Graphd [EB/OL]. http://wiki.freebase.com/wiki/Graphd.

[6] 阮一峰.Freebase 再研究[EB/OL]. http://www.ruanyifeng.com/blog/2008/04/freebase_reloaded.html.

[7] Query Editor[EB/OL]. http://www.freebase.com/view/queryeditor/.

（收稿日期：2013-08-27）

作者简介：

王斌，男，1987年生，硕士研究生，主要研究方向：人工智能的属性论。

冯嘉礼，男，1948年生，教授，博士，博士生导师，主要研究方向：人工智能的属性论。