基于超网络的维基百科内容知识本体演化研究①

倪子建，荣莉莉，刘 泉
（大连理工大学系统工程研究所，大连 116023）

摘要： Wiki（维基）作为巨大的知识库是知识共享的重要平台，从宏观角度研究 Wiki 中文章的知识属性和发展趋势，有利于了解科学技术和人文社会的发展方向. Wiki 中文章的标签和概要信息栏中的信息相结合可以有效地概括这篇文章的主要知识，而 Wiki 内容本体可以形式化描述这些语义知识. 通过实证研究发现 Wiki 内容本体有两个重要特点，即存在多元关系和本体的规模不断增长. 根据以上两个特点，提出以超网络模型描述 Wiki 内容本体，进而研究这个超网络模型的演化模式. 具体内容如下：以 Wiki 的内容知识本体 YAGO 为数据源，提出本体超网络的构建方法；构建了 Wiki 内容本体超网络；在对上述网络的重要拓扑属性进行分析的基础上，提出了 Wiki 内容本体超网络的演化模式；对演化模式进行解析和数值模拟，给出了结论.

关键词： 超网络；维基百科；本体；演化；多元关系；YAGO；超边；多元关系
中图分类号： C94 **文献标识码：** A **文章编号：** 1007－9807(2013)12－0068－11

0 引 言

在知识经济的背景下，企业无论是产品研发还是服务、创新都需要跨学科的知识支持[1]. 多领域知识是维基百科（以下简称 Wiki）的基础，从宏观上考察这些知识的脉络不仅有利于知识共享的进行，而且有利于了解科学技术和人文社会的发展方向. 目前此类研究主要基于网络模型对 Wiki 文章的主题进行建模[2-4]，从宏观上研究 Wiki 上知识的属性. Wiki 文章的主题，如 Chemistry、Politics 等，不仅是 Wiki 对文章进行分类组织的依据，而且反映了文章内容所涉及的主要知识概念和学科领域. 文献[2,4]试图通过这些主题的宏观统计特性揭示知识组织的一般模式；文献[2]把主题网络和人际网络相结合，通过对两种网络重叠性的分析，更有效地支持文章推荐，有利于知识的共享；文献[4]将主题看作知识树，然后通过算法分离隐藏在知识树中潜在的层次结构，并且对知识树中不同层次类别所包含的内容进行自动分类.

当前的研究取得了一些成果，但也有不完善之处，首先，主题仅仅是多个概念组成的集合，它不能反映文章内部概念与概念之间的语义关系，例如 Wiki 中介绍猫王（Elvis Presley）文章的主题之一是 Actor，但是这个主题无法反映猫王是在哪一年成为的演员，现在是否还是演员，而这种概念与时间的关系对于知识推荐往往是非常重要的. 第二，当前的研究一般只是对 Wiki 中的静态知识网络的属性进行研究，而 Wiki 文章的数量②却按照指数规模在不断增长[4]，相关的知识网络必须

① 收稿日期：2011－11－09；修订日期：2012－10－05.
基金项目：国家自然科学基金重大研究计划资助项目（91024003；91024031）；高等学校博士学科点专项科研基金资助项目（20090041110010）.
作者简介：倪子建（1981—），男，辽宁沈阳人，博士生. Email：nizijian@hotmail.com
② 文献[4]描述了 Wiki 文章数量与时间之间的函数可以分阶段地用两个指数函数描述.

从动态上把握这一特性.

需要指出的是 Wiki 文章内容的知识要点不仅包含在主题标签中,也体现在每篇文章的 infobox 中. infobox 一般出现在 Wiki 文章详细内容的右侧,其作用在于显示这篇文章的主要信息,本文把它翻译为概要信息栏. 图 1 所示就是猫王在 Wiki 中的概要信息栏,它和主题标签的信息相结合,可以很有效地概括出 Wiki 文章的内容知识.

图 1　猫王 Elvis 在 Wiki 文章中的概要信息栏

Fig. 1 The infobox of the Elvis in Wiki

注: 如果要形式化描述概要信息栏中最下面两行的完整语义,3 元关系是必须的.

基于本体的方法比简单的主题概念集合能更有效地对知识进行形式化描述,可以应用于 Wiki 知识的建模. 目前已有一些研究基于复杂网络的方法对领域本体进行建模,从宏观上研究本体的特性. 文献[5]通过研究两个节点之间的平均距离和节点的度来反映本体中概念的一致性和复杂性. 文献[6]通过有向非循环图对基因本体(GO)进行建模,通过每个概念的平均路径数及路径与概念比来衡量本体的复杂性. 这些研究通常以概念为节点,概念与概念之间的 2 元关系为边建立网络模型. 以上研究除了没有从演化的角度讨论本体模型的宏观特性以外,还有个重要的不足,即这些模型不能描述 Wiki 知识中广泛存在的多元关系.

当前的很多领域本体,其概念与概念之间的关系一般是 2 元的,即一个关系只能连接两个概念. 但是在 Wiki 文章中的多元关系比 2 元关系往往能更完整地反映语义知识,例如图 1 所示的概要信息栏中最下面两行的内容: Occupations: actor; Years active: 1953 – 77,其要表达的完整语义是: Elvis was active as the actor between 1953 and 1977. 如果仅用两个 2 元关系 ActiveBetween(Elvis, 1953 – 77) 和 HasOccupy(Elvis, Actor) 描述,则无法完整反映上述语义. 而使用 3 元关系 ActiveAsBetween(Elvis, Actor, 1953 – 77) 则可以描述. 这样由点和两点间连线组成的复杂网络模型就并不适合描述这种 3 元关系本体.

为此本文引入超网络[7]来构建这种由概念和多元关系构成的本体超网络模型. 超网络和普通网络模型最大的区别在于网络中的边. 普通网络模型中的边只能连接网络中的两个点,超网络中的边可以连接两个以上的点,一般把超网络的这种边叫做超边[8]. 另外建立超网络模型的过程中,根据实际情况的不同,点和超边可以有不同的含义. 例如在 Wiki 内容本体超网络中点表示本体中的概念,而超边则表示由这些概念所组成的多元关系.

本文的研究主要包括以下几个部分: 首先,提出基于 Wiki 内容知识本体的复杂超网络模型的构建方法; 然后,通过实际数据的分析,得出当前本体超网络的结构特点; 在此基础上,通过当前网络的结构特点,提出本体超网络的演化模型,并通过模拟仿真验证演化模型的解析结果.

1　超网络模型的构建

1.1　数据来源

目前已经有一些研究项目[9-11]通过自动或

者半自动的方法，从 Wiki 文章中抽取信息，建立 Wiki 内容知识本体. DBpedia 项目[11] 利用递归正则表达式方法，从概要信息栏中自动抽取概念及概念之间的关系，虽然其概念要比主题标签丰富，但是目前该本体所包含的关系主要是父类子类这样的继承关系. YAGO 项目[9] 将 WordNet 和 Wiki 相结合，使用启发式规则从 Wiki 概要信息栏和主题标签中抽取概念和概念之间的语义关系(不仅是继承关系). 此外 YAGO 的本体模型还可以描述文章中的多元关系[9].

Wiki 文章为了增加主题标签的区分度，往往把这些主题标签定义得很多很细. 例如 Wiki 中给作为个体的猫王定义了 21 个主题标签，如 TennesseeMusicians，AmericanCountrySingers 等，这些过于冗繁的概念会使对于个体的归类无所适从. 而 YAGO 本体通过 Wiki 主题标签和 WordNet 相结合的方法，把这些繁杂的底层概念归纳成 WordNet 中抽象度较高的概念，比较好地解决了这个问题. 例如在 YAGO 的概念层次中上面提到的 21 个概念都属于 Human 这个概念.

本文所建立的 Wiki 本体超网络模型就是依据 YAGO 数据源.

1.2 YAGO 语义模型

YAGO 语义模型类似于语义网中的 RDF 模型[12]，其基本结构是用 1 个 2 元关系连接两个实体或者概念. YAGO 把每一个这样的最基本的元组叫做 1 个事实(fact)，与 RDF 不同的是，YAGO 给每一个 fact 确定了 1 个唯一标示(fact identifier). 例如在 YAGO 模型中，Elvis's occupation was an actor 描述如下

#1: Elvis hasOccupy Actor

其中#1 是这个 fact 的唯一标示，Elvis 是这个 fact 的前项，Actor 是这个 fact 的后项，而 hasOccupy 表示前项和后项之间的语义关系.

YAGO 的 3 元关系模型依据如下假设，即每个 YAGO 中的 3 元关系都可以分解出语义 2 元关系[13]. 例如 3 元关系 ActiveAsBetween (Human，Actor，TimeInterval)，可以分解出两个 2 元语义关系 ActiveAs (Human，Actor) 和 ActiveBetween (Human，TimeInterval)，分别表示人的职业属性和时间属性，而 Actor 和 TimeInterval 间没有明显的语义关系.

在 YAGO 中 3 元关系模型一般是由两个 fact 组成，第 1 个 fact 表示这个 3 元关系中的 2 元语义关系，第 2 个 fact 的前项是第 1 个 fact 的标示符(fact identifier)，而后项则连接着 3 元关系中剩下的那一元. 例如对于如下叙述: Between 1953 and 1977 Elvis was active as an actor，其 3 元关系模型如下

#1: Elvis ActiveAs Actor

#2: #1 ActiveBetween 1953 – 77

其中#1 表示 2 元语义关系 ActiveAs (Elvis，Actor)，而#2 则把整个 fact 作为本身的前项.

可以证明 YAGO 模型[9] 是可推理可计算的.

1.3 YAGO 本体超网络模型的构建

在讨论了 YAGO 语义模型之后，本节将研究基于这个语义模型把 YAGO 的数据转换成超网络模型. 用 $H = \{V, HE\}$ 表示 Wiki 内容知识本体的超网络模型，其中 $V = \{v_1, v_2, \cdots, v_n\}$ 表示点的集合，不同的点对应本体中的不同概念; $HE = \{HE_1, HE_2, \cdots, HE_n\}$ 表示超边的集合，其中 $HE_i = \{v_{i1}, v_{i2}, \cdots, v_{in}\}$ 表示 1 条超边包含多个点，不同的超边对应本体中不同的多元关系. 需要注意的是在 H 中同样存在由两个点构成的普通边的集合 $E = \{e_1, e_2, \cdots, e_n\}$，$E$ 可以看做点集 V 笛卡尔积的子集，即 $E \subseteq V \times V$，在 E 中不同的边对应本体中不同的 2 元关系.

下面的例子中，将根据以上定义构建简单的超网络. 这个模型共包含 6 个 Wiki 内容本体中常见的概念(见图 2)，5 个语义 2 元关系和 3 个 3 元关系. 为了表达方便使用 3 元关系的一阶谓词描述方法，3 元关系如下

* InOfficeAsSince(Human，President，TimeInterval_S)
* InOfficeAsSince(Human，Senator，TimeInterval_S)
* ActiveAsBetween(Human，Singer，TimeInterval_B)

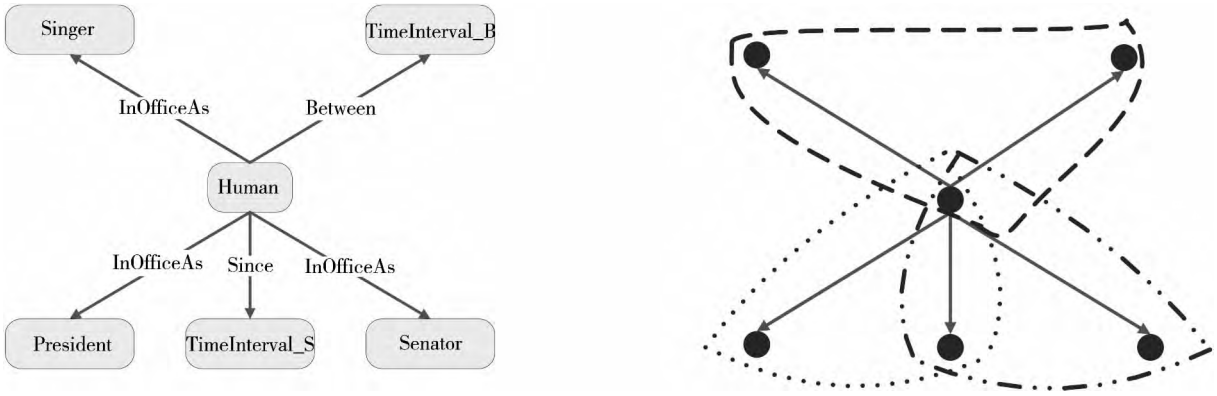

图 2　本体的超网络模型

Fig. 2 The hypernetwork of a ontology

注: 左侧图表示没有考虑 3 元关系的本体中概念和语义关系的示意图,TimeInterval_B 和 TimeInterval_S 是 TimeInterval 的子类; 右侧是本体网络被转换成超网络的示意图.

构建方法是首先把 3 元关系的 6 个概念建立 6 个点. 如果两个点之间存在 2 元语义关系,那么就建立 1 条从关系前项到后项的有向边. 此外每一个 3 元关系构成 1 条超边.

上述简单本体所转换的超网络模型如图 2 所示. 需要指出的是, 在 YAGO 数据中, TimeInterval_S 和 TimeInterval_B 用同一个概念 TimeInterval 表示,并且 TimeInterval 这个概念出现在很多 2 元语义关系中. 对于本体模型这并无不妥之处,但是对应于超网络模型,就会出现很多节点之间的连线不止一条的情况,这样会给网络模型相关属性的统计造成麻烦. 为此,使用不违背构建本体原则的方法解决上述问题,即定义概念 TimeInterval 的两个子类, 即 TimeInterval_B 和 TimeInterval_S. 这样就可以在不影响本体结构的基础上,减少重复连边的出现.

1.4　超网络模型的拓扑属性

需要说明的是本节讨论的属性都不考虑边的方向性. 根据 Wiki 内容本体超网络的静态特性和演化特性,本文重点讨论如下超网络属性:

1) 节点的度　节点 i 的度 k_i 表示与该节点直接相连的节点个数. 在本体超网络中,如果节点的度大,则表示这个概念可能出现在很多 Wiki 文章中,被广泛讨论. 显然,在动态变化的超网模型中,节点的度是不断变化的,所以主要讨论网络中节点的度分布 $P(k)$. $P(k)$ 表示一个随机选定节点的度恰好为 k 的概率. 通过这个属性可以反映不断增长的 Wiki 中热点概念演化的趋势.

2) 点的超度　节点 i 的超度 $d_{\mathrm{H}i}$ 表示包含该节点的超边条数,即一个概念出现在不同的多元关系中的次数. 在 Wiki 内容本体中,多元关系比 2 元关系更能完整地反映文章要表达的语义,所以点的超度比点的度更能反映语义概念被讨论的广泛程度.

3) 超边的度　为了研究超边的度,本文提出用本体超网络的超边映射图(简称超边映射图),来反映本体超网络中超边之间的关系.

图 3 显示的是超网络向超边映射图转换的示意图. 超边映射图中的点对应超网络中的超边,如果两个超边之间存在定义 1 所规定的关系,那么它们在超边映射图中对应的点之间就存在 1 条边.

定义 1　超边映射图中的两个点之间存在一条边,当且仅当这两个点对应于超网络中的超边共享一条边.

上述定义反映了本体中的多元关系间的一个特征,即如果两个 3 元关系都包含相同的语义 2 元关系,那么它们要表达的语义常常存在相似性. 例如 3 元关系 InOfficeAsSince(Human, President, TimeIntval) 和 InOfficeAsSince (Human, Senator, TimeInterval),两者都存在相同的语义 2 元关系 InOfficeSince(Human, TimeInterval). 下面讨论超边的度.

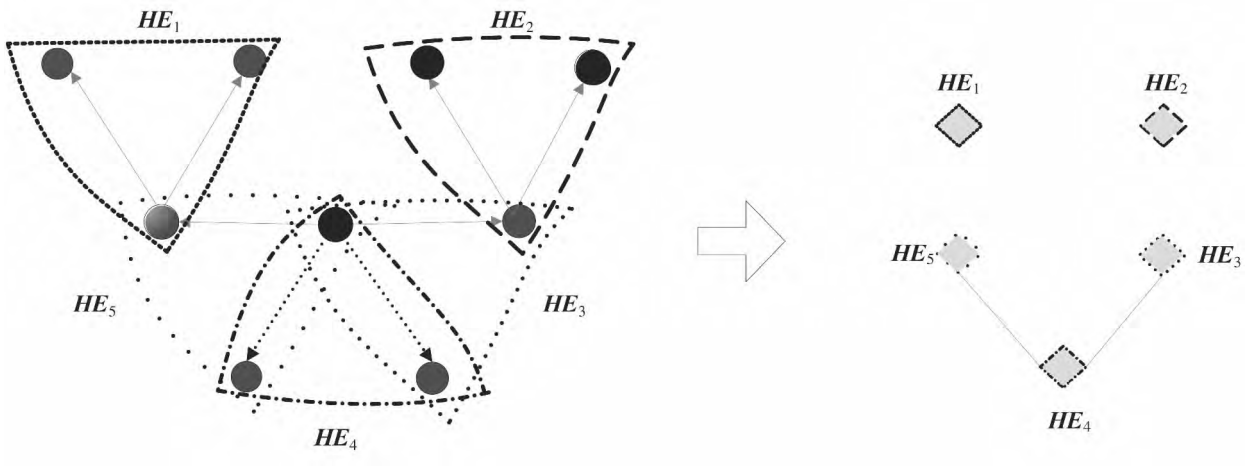

图 3 本体超网络的超边映射图

Fig. 3 The projected graph of the ontology hypergraph

注: 左侧超网络图中存在 5 个超边($HE1$,…,$HE5$) ,其中两对超边之间存在共享边(下部两条点线的边) . 根据定义对应于右侧的超边映射图就存在 5 个点,其中 3 个之间存在边,另有两个离散点.

超边的度 d_{He} 是指在超边映射图中,与超边 He_i 所对应的点直接相连点的个数. 可以对节点的超度和超边的度这两个概念的实际意义进行如下比较: 点的超度反映了单一概念在各个 3 元关系中的被讨论的广泛程度,而超边的度则反映了 3 元关系作为整体在超网络中的被讨论的广泛程度.

2 演化机制研究

首先对由 YAGO 数据生成的 Wiki 内容本体超网络的静态属性进行分析,然后,根据分析的结果提出演化机制.

本文所构建的 Wiki 内容知识本体超网络的数据来源于 YAGO 的 2008 数据集[14] ,该数据集提取了 Wiki 中截止到 2007 年 11 月的大约二百万篇文章. 本文对 2008 数据集中的概念和关系进行提升、整理、纠错之后,使用其中的 1 720 个概念, 2 743 个 3 元关系和 5 770 个 2 元关系构建初始网络.

这里需要说明的是,通过整理发现 YAGO 中的多元关系几乎都是 3 元关系,所以本文主要讨论 3 元关系.

2.1 超网络静态属性分析

本文主要讨论对 YAGO 网络中点的度分布和超边的度分布,这两个属性可以反映网络和超网络的宏观属性.

2.1.1 节点的度分布

当前,统计网络的度分布主要有 3 种方法[15] .

第 1 种方法,将网络的度和频率在双对数坐标上画图,然后进行线性拟合. 这种方法的缺点是度分布的图像尾部摆动范围很大,指数统计不易准确估计.

第 2 种方法,为了避免长尾引起的误差,可以将补分布 $F_c(k) = \sum_{l \geqslant k} P(l)$ 在双对数坐标上画图,然后进行线性拟合得到度指数 α,这个度指数和原始网络度指数 γ 之间的关系是 $\gamma = \alpha + 1$,由于本文篇幅有限,以上结论的详细证明请参考文献[15] .

第 3 种方法,使用对数盒子法,即分段求出度分布在双对数坐标系下的均值.

文献[15] 比较了上述 3 种方法,发现第 2 种方法最为准确可靠. 由本文所统计的点的度分布在双对数坐标下尾部波动很大,为了精确统计 γ 值,采用第 2 种方法. 图 4 就是从实际数据统计出点的度的补分布.

通过图 4 可以看出,超网络中点的度的补分布 $F_c(k)$ 符合幂律分布,即 $F_c(k) \sim K^{-\alpha}$. 对于图 4 中的点,采用 OriginPro 进行非线性拟合,得到

$\alpha = 1.1$,所以 $\gamma = \alpha + 1 = 2.1$. Barabási 和 Albert 在文献[7]中提出了 BA 模型,文献[6]对 BA 模型进行解析,得到在 BA 模型中 $\gamma = 3$. 而从实证数据得到的结果显示,与 BA 模型相比,Wiki 内容本体超网络中拥有较多连线的点更多,与文献[8]所统计的万维网入度的度分布相似.

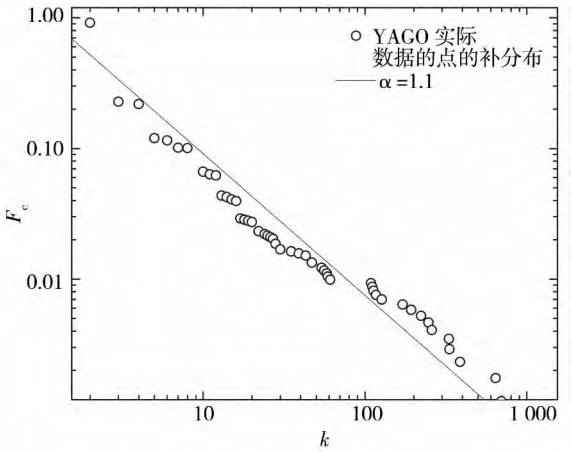

图 4　YAGO 数据反映的 Wiki 内容本体超网络点的度的补分布

Fig. 4 Complementary CDF of node degree in the ontology hypernet extracted from the YAGO for Wiki

虽然万维网和 Wiki 内容本体超网络都属于信息网络,且点的度分布很接近,但是导致点的度分布为幂律分布的原因却不同. 文献[9]认为互联网模型中点的幂律分布是由于互联网内容的发布者们仅仅拥有"有限权限"(local event),也就是说作为管理员(网站的编辑),他们更倾向于在新发布的页面中引用自己网站页面的链接.

而作为 Wiki 中的管理员(维客)却没有这方面的顾虑,他们可以发布和修改任何自己感兴趣的知识条目. 从表面来看 Wiki 中的这种机制可能使拥有较多连边的点更多,即热点概念更分散,但从实际网络中得到的数据并没有支持上述猜测. 本文认为其主要原因与 Wiki 中文章的内容和文章概要信息栏中的内容相关. 大家知道 Wiki 中主要以介绍人物和重要的事物、事件为主,这样就使在网络中类似于 Human 和 Event 等少数的概念拥有很多的连边,另外概要信息栏主要的内容是有关于人物的主要生平经历,或者事件、事物发展的历史,所以与时间有关的少数几个概念也拥有很多的连边,例如 TimeInterval.

2.1.2　超边的度分布

仅从点的度分布还难以抽取出超网络演化规律的主要特征,因为在以往研究所提出的多种网络演化模式中,不止一种模式稳态后其点的度分布指数可以小于 3,如文献[20]所提出的复制增长模型,文献[21]所提出的复制完全图模型等. 要反映 Wiki 内容本体超网络的演化特点,就必须要从现实中的网络增长方式中提取其可能的增长模式加以抽象,为此进一步讨论超边的度分布.

统计超边的度发现,有接近一半(45.6%)的超边的度为零,也就是说在超网络所转化的超边映射图中有大约一半的节点是离散的. 在映射图中,统计度大于零的节点度分布,同样采用补分布法进行拟合,结果如图 5 所示. 得到 $\alpha = 0.6$,$\gamma = \alpha + 1 = 1.6$.

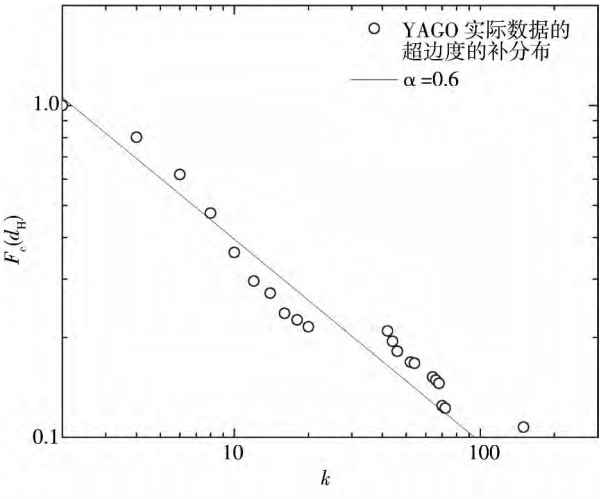

图 5　YAGO 数据反映超网络中超边度的补分布

Fig. 5 Complementary CDF of hyperedge degree in the ontology hypernet extracted from the YAGO for Wiki

注: 上述数据是统计非离散节点的结果,实际的超边映射图中的点有大约一半的节点是离散的.

对于超边映射图中的离散点,本文认为这恰恰能反映 Wiki 内容本体超网络的演化特点. 演化的模式需要综合考虑当前网络的两个特征,即节点的度分布的 γ 值小于 BA 模型和超边映射图中的点有一半是离散的.

2.2　超网络模型的演化模式

在给出具体演化模式之前,先讨论在 Wiki 文章不断增加的过程中其内容本体可能经常出现的现象. Wiki 中有很多介绍人物的文章,随着科学

技术的不断发展，很多有关人的职衔、称谓和职业等的新名词不断出现，如美国已经出现了 SpaceShipDriver（宇宙飞船驾驶员）这样的职业，也就是说 Wiki 英文版中已经出现了类似的概念. 未来的 Wiki 中文版中可能会出现“宇宙飞船驾驶员”之类的概念，这样就会出现 Wiki 内容本体的演化. 目前 Wiki 内容本体中有很多是关于人的生平履历方面的信息，如 RetireAsSince（Human，President，TimeIn-terval）. 如果在未来的演化过程中，在上述 3 元关系的基础上增加了一个关系 RetireAsSince（Human，宇宙飞船驾驶员，TimeInterval），那么相当于在当前的超网模型中增加了 1 个点、1 条边和 1 条超边，同时在超边映射图中增加了 1 个点和 1 条边.

在进一步的演化中，由于 Wiki 的多语言特性，多元关系中通常会出现 IsCalled_inLanguage（宇宙飞船驾驶员，SpaceShipDriver，English），即宇宙飞船驾驶员这个概念在英文中对应的翻译. 如上文所述，3 元关系中 SpaceShipDriver 和 English 是在 Wiki 内容本体中已经出现的概念. 这种演化会在当前的超网络模型中已有的 3 个节点新连接出两条边，和 1 条超边，同时在超边映射图中增加 1 个点，但是这个点是离散的，即这个点的度为 0. 具体过程如图 6 所示.

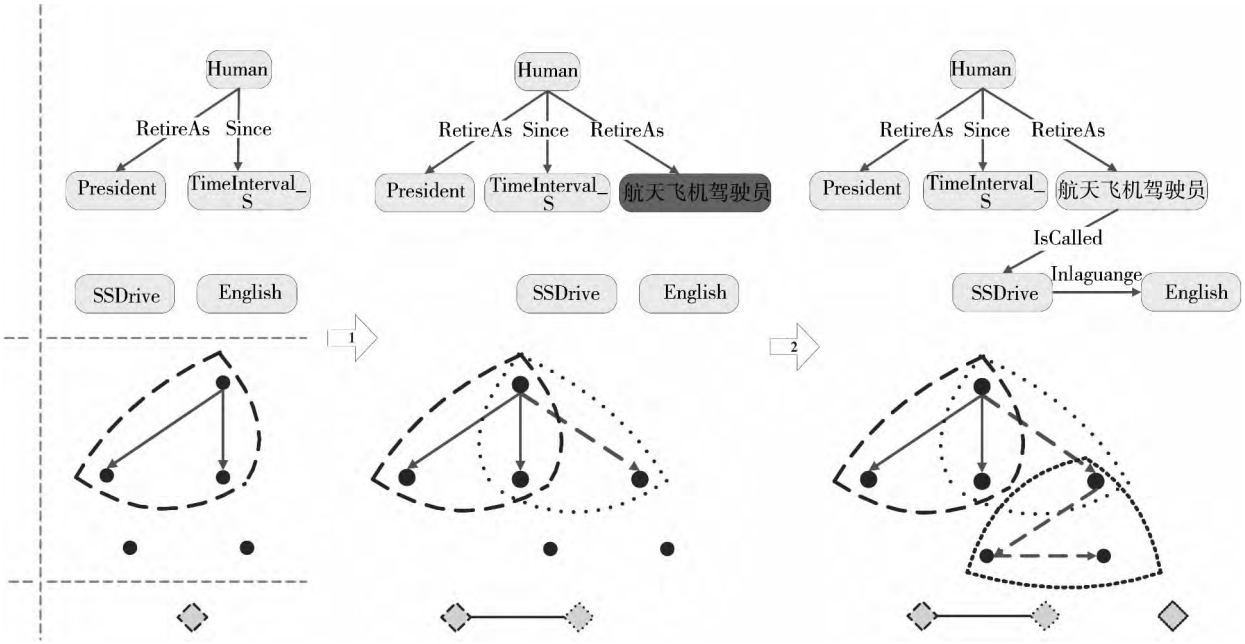

图 6 本体超网络演化示意图

Fig. 6 Scheme of the growing network of Wiki ontology

注：加入的概念在本体和超网模型中用内部填充黑色的图形表示，已有节点用黑色表示. SSDriver 是 Space Ship Driver 的简写.

图 6 中显示了这个演化过程，图中的第 1 行显示了本体中概念和关系的增加，第 2 行显示了超网络模型的演化，第 3 行显示了超边映射图的演化.

进一步观察上述演化过程，可以发现这个过程实际包括两个步骤，即新节点连接已有网络和网络已有的老节点之间的连接. 对于 Wiki 来说，新节点的连接代表外延式的发展，即随着新知识的不断发展，新概念层出不穷；而老节点的连接则代表内涵式的发展，即随着对已有知识研究的不断深入，发现已有知识间的一些新关联.

在不考虑超边增加和超边之间连接的情况下，本文提出如下以增长和择优为原则的 Wiki 内容本体的超网络的演化模式：

1）初始条件. 网络中初始存在 m_0 个节点，m_0 很小.

2）边的增长. 在 $t = 1$ 时刻内按以下两步依次进行. 第 1 步，增加 1 个度为 $m(m \leqslant m_0)$ 的节点，并连接 m 个老节点；第 2 步，在图中的 m_0 个老节点产生 p 条边（$p \leqslant m_0(m_0 - 1)/2$）. 在以后每

一个时间间隔内,向网络增加 1 个度为 m 的新节点,其 m 条边择优连接到网络中已经存在的不同节点上,并且现有网络老节点间择优产生 p 条边. 这样经过 t 个时间间隔,便形成具有 $N = m_0 + t$ 个节点,$(m + p)t$ 条边的网络.

3) 边的择优增长. 第 2 步中两种情况选择的老节点均按照以下择优概率选择旧节点 i 与之连线

$$\prod(k_i) = \frac{k_i}{\sum_j k_j} \tag{1}$$

式中 k_i 是旧节点 i 的度,$\sum_j k_j$ 表示所有旧节点度的和.

在下一节中,将通过解析方法计算上述演化模式.

2.3　理论解析和结果讨论

采用文献 [22] 所提出的平均场法进行解析. 令 $k_i(t)$ 表示在 t 步结束时节点 i 的度数,而节点 i 是在第 i 步增加的新节点. 假设 $k_i(t)$ 连续,由于新增节点与老节点间的连接,以及老节点之间的连接均服从择优连接规则,对已有节点 i,有

$$\frac{\partial k_i(t)}{\partial t} = \Delta k(t)\, g \prod(k_i) \tag{2}$$

式中 $\Delta k(t)$ 表示 t 时刻网络度的增量. 在 1 个时间间隔网络度的变化 $\Delta k(t)$ 由两部分构成,一是由新增节点增加的 m 条边导致老网络的新增度 m,二是由老节点产生 p 条边导致的新增度 $2p$,因此有 $\Delta k(t) = m + 2p$. 又考虑到 $\sum_j k_j = 2(m + p)t$,则有

$$\frac{\partial k_i(t)}{\partial t} = \frac{m + 2p}{2(m + p)} \frac{k_i(t)}{t} \tag{3}$$

设节点 i 是 t_i 时刻添加到网络的,则连通度 $k_i(t_i) = m$,此即为式(3) 中微分方程的初始条件. 通过计算,可得方程(3) 的解

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{m+2p}{2(m+p)}} \tag{4}$$

考虑节点连通度 $k_i(t)$ 小于 k 的概率 $P(k_i(t) < k)$ 可表示为

$$P(k_i(t) < k) = P\left(t_i > \left(\frac{1}{k}\right)^{\frac{2(m+p)}{m+2p}} t\right) \tag{5}$$

假设时间 t 服从均匀分布,则

$$P(k_i(t) < k) = 1 - \frac{1}{m_0 + t}\left(\frac{m}{k}\right)^{\frac{2(m+p)}{m+2p}} \tag{6}$$

则节点连通度的分布函数 $P(k)$ 为

$$P(k) = \frac{\partial P(k_i(t) < k)}{\partial k} = \frac{t}{m_0 + t} \frac{2(m + p)}{m + 2p} m^{\frac{2(m+p)}{m+2p}} k^{-\left(\frac{2(m+p)}{m+2p}+1\right)} \tag{7}$$

从上式可得节点度分布呈幂率分布,其幂指数为

$$\gamma = \frac{2(m + p)}{m + 2p} + 1 = 2 + \frac{1}{1 + \frac{2p}{m}} \tag{8}$$

p 与 m 取值不同时 γ 的变化如下:

1) 当 $p = 0$ 时,$\gamma = 3$,即转化为 BA 模型;

2) 当 $p \ll m$ 时,即老节点连接的边远小于新节点连接的边时,γ 接近 3;

3) 当 $m \ll p$ 时,即新节点连接的边远小于老节点连接的边时,γ 接近 2.

下面讨论解析结果中涉及到的变量的实际意义、变量取值和超边选择策略:

1) 当 $p \ll m$ 时,说明新知识产生得非常多,通常对应于科学界的重大发现,这时容易把有限的科研力量吸引到新领域中. 通过解析可以发现这时研究讨论的热点更为集中.

2) 当 $m \ll p$ 时,说明开创性的知识发现速度减慢,这时研究界可能转换为对现有知识进行领域交叉. 从解析的结果来看这时热点相对比较分散.

3) 当 $p = 2$, $m = 1$ 时,$\gamma = 2.2$ 与实证数据统计出的结果最接近. 此时 $p/m = 2$,从数值看两者差距不大.

4) 当 $p = 2$, $m = 1$ 时,边和超边在初始 m_0 个节点的网络中增长策略如下: 第 1 步,增加 1 个节点,并择优连接 1 个老节点 a,然后随机选择 1 个与 a 相连的老节点构成 1 个超边; 第 2 步,择优选择当前网络中的 3 个老节点,用两条边连接它们,在超边映射图中构成 1 个离散的点. 这样经过 t 个时间间隔,便形成一个具有 $N = m_0 + t$ 个节点,$3t$

条边的网络.

2.4　超网络演化模拟

采用模拟方法[23]，对简单初始网络按照本文所提出的演化模式进行 10 次实验，每次模拟按照前文的演化规则重复 10 000 次. 图 7 就是所得到的 10 次模拟结果平均后的点度的补分布，

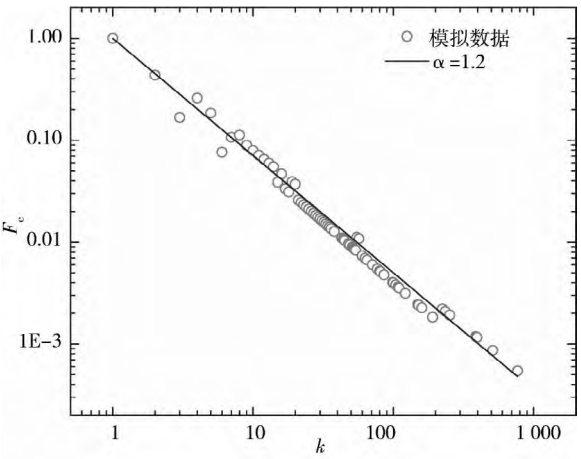

图 7　模拟数据与理论推导的点度的补分布的比较

Fig. 7 The comparison of complementary CDF of node and theoretical Complementary CDF

从图 7 可以看出，按照本文所提出的演化规则所生成网络的 $\gamma = 2.2$. 与解析出的度分布值相吻合.

采用上面 10 组数据中的结果，统计超边的度分布，得到有将近一半的超边度为 0，即这些超边在超边映射图中是离散的. 对于非离散超边的度的补分布的统计结果如图 8 所示.

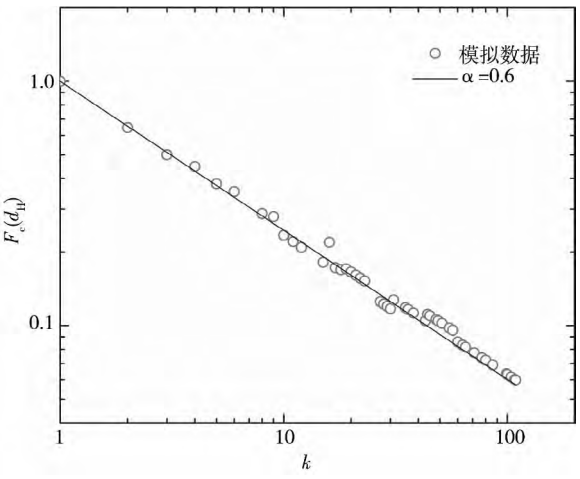

图 8　模拟数据与理论推导的超边的度的补分布的比较

Fig. 8 The comparison of complementary CDF of hyperedge and theoretical

从图 8 可以看出，模拟数据得到的超边度分布与解析的结果相吻合.

本节采用解析和模拟的方法验证了本文提出的 Wiki 内容本体超网络的演化模式. 从解析的结果来看，本文所提出的增长模式可以有效地描述实证数据网络静态属性的两个特点，即拥有大量连边的点更少和超边映射图中存在相当数量的离散点. 此外，模拟的结果既验证了增长模式，又验证了解析结果的正确性.

3　结束语

本文在复杂性的视角下，研究了对 Wiki 内容本体. 在总结以往研究的基础上，发现了 Wiki 内容本体的两个特点: 存在多元关系和本体规模不断增长，对应于网络的属性就是存在超边和不断演化. 为此本文提出使用超网模型描述 Wiki 内容本体的这些性质: 首先给出了基于 YAGO 本体构建 Wiki 内容超网络模型的方法，然后分析了当前网络中主要的拓扑属性，提出了超网络模型的演化模式，最后对演化模式进行了解析和数值模拟，给出了结论.

Wiki 本身是个非常活跃的海量知识库，其内容和模式也在不断地变化中. 随着相关算法的不断进步，从中自动挖掘知识的内容和数量都会不断增长.

本文所提出的方法为研究类似复杂知识系统的建模和演化问题提供了可借鉴的思路. 例如在应急管理领域，应急预案被认为是应对突发事件知识的集合. 我国从 2003 年开始应急预案体系的建设，虽然还没研究统计过每年增加的预案数目，但是目前各级政府部门公布的各类预案超过 150 万件[24]，可见速度相当惊人.

应急预案中涉及到不同的应急主体(主要是各级政府部门) ，例如“国家地震预案”中就有 180 多个不同的主体. 而预案就是要为不同的主体提供不同应急响应信息. 这些信息一般模式是对于某个主体，在一定的灾情状况下应该进行哪类应急响应. 预案中所包含的知识概念主要涉及

到这3种信息,即主体,灾情和响应. 把这3种概念联系起来需要 3 元关系,本文的超网建模方法可以应用于此,进而分析预案中知识超网的静态属性.

各级各类预案的不断颁布既是应急救援范围不断扩大的过程(应对突发事件种类不断增加),也是应急响应不断细化的过程(从战略层到操作层),本文的演化策略对研究这种具有双重驱动力的演化过程有借鉴意义.

参 考 文 献:

[1] 杨　青,乔志刚,黄丽华,等. 动态联盟中企业建模的 Meta-Model [J]. 管理科学学报,2001,4(6): 31 - 38.
Yang Qing, Qiao Zhigang, Huang Lihua, et al. Meta-model of enterprise modeling in dynamic alignment [J]. Journal of Management Sciences in China, 2001, 4(6): 31 - 38. (in Chinese)

[2] Capocci A, Rao F, Caldarelli G. Taxonomy and clustering in collaborative systems: The case of the online encyclopedia Wikipedia [J]. EPL, 2008, 81(2): 28 - 38.

[3] Yule G U. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis [J]. F. R. S, 1925, 213(3): 21 - 87.

[4] Muchnik L, Royi I, Sorin S. Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies [J]. Physical Review E, 2007, 76 (15): 16 - 106.

[5] Rodrigo B, Barzan M. On the evolution of Wikipedia [C] // Proceedings of the International Conference on Weblogs and Social Media, 2007: 1 - 8.

[6] His I. Analyzing the conceptual coherence of computing applications through ontological excavation [D]. Atlanta: Georgia Institute of Techonlogy, 2004.

[7] Mungall C. Increased complexity in the GO [EB/OL]. 2005, http://www.fruitfly.org/~cjm/obol/doc/go-complexity.html

[8] 王众托,王志平. 超网络初探 [J]. 管理学报,2008,5(1): 1 - 8.
Wang Zhongtuo, Wang Zhiping. Elemen tary study of supernetworks [J]. Chinese Journal of Management, 2008, 5(1): 1 - 8. (in Chinese)

[9] Suehanek F, Kasnee M. YAGO: A core of semantic knowledge [C] //Proeess of 16th WWW Conference, 2007.

[10] Nastase V, Strube M. Decoding Wikipedia categories for knowledge acquisition [C] //Procees of the AAAI' 08 Conferenee, 2008: 1219 - 1224.

[11] Auer S, Bizer C. DBpedia: A nucleus for a web of open data [C] //6th Int. Semantic Web Conference, Busan, Korea, 2007: 11 - 15.

[12] Suchanek F M, Kasneci G, Weikum G. YAGO: A large ontology from Wikipedia and WordNet [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6 (3): 203 - 217.

[13] Frank M, Eric M. RDF Primer [EB/OL]. Web: 2004, http://www.w3.org/TR/rdf-primer/.

[14] Fabian M, Suchanek. Automated Construction and Growth of a Large Ontology [D]. Faculties of Natural Sciences and Technology of Saarland University, 2009.

[15] 史定华. 网络度分布理论 [M]. 北京: 高等教育出版社,2011.
Shi Dinghua. Theory of Network Degree Distributions [M]. Beijing: Higher Education Press, 2011. (in Chinese)

[16] Alderson D, Doyle J C, Li L, et al. Towards a theory of scale-free graphs: Definition, properties, and implications [J]. Internet Mathematics, 2005, 2 (4): 431 - 523.

[17] Barabási A L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286 (1999): 509 - 512.

[18] Dorogovtsev S N, Mendes J F F, Samukhin A N. Structure of growth networks with preferential linking [J]. Physical Review Letters, 2000, 85(21): 4633 - 4636.

[19] Broder A, Kumar R, Maghoul F, et al. Graph structure in the Web [C] //Proceedings of the 9th International World Wide Web Conference on Computer Networks: The International Journal of Computer and Telecommunications Networking, North-

Holland Publishing Co, 2000: 309 – 320.

[20] Capocci A, Servedio V, Colaiori F. Preferential attachment in the growth of social networks: The case of Wikipedia [J]. Physical Review E, 2006, 74(3): 036111 – 036116.

[21] Krapivsky P L, Redner S. Network growth by copying [J]. Physical Review E, 2005, 71(3): 036118(7).

[22] 章忠志, 荣莉莉, 周 涛. 一类无标度合作网络的演化模型 [J]. 系统工程理论与实践, 2005, 25(11): 55 – 60.
Zhang Zhongzhi, Rong Lili, Zhou Tao. An evolving model for scale-free collaboration networks [J]. Systems Engineering-Theory & Practice, 2005, 25(11): 55 – 60. (in Chinese)

[23] 王建伟, 荣莉莉, 王 铎. 基于节点局域特征的复杂网络上相继故障模型 [J]. 管理科学学报, 2010, 13(8): 42 – 50.
Wang Jianwei, Rong Lili, Wang Duo. Model for cascading failures on complex networks based on local characteristics of nodes [J]. Journal of Management Sciences in China, 2010, 13(8): 42 – 50. (in Chinese)

[24] 闪淳昌, 黄 敏. 中国应急管理及运行模式 [J]. 北京航空航天大学学报(社会科学版), 2010, 23 (2): 22 – 26.
Shan Chunchang, Huang Min. Organization and operation of China's emergency management [J]. Journal of Beijing University of Aeronautics and Astronautics(Social Sciences Edition), 2010, 23(2): 22 – 26. (in Chinese)

Study on evolving hypernetwork model of Wiki ontology

NI Zi-jian, RONG Li-li, LIU Quan

Institute of Systems Engineering, Dalian University of Technology, Dalian 116023, China

Abstract: Wikipedia as a largest knowledge base is an important platform for sharing knowledge in the world. It can help understand the developing trend in the science, technology and society through the macro-level statistical properties of knowledge and their evolution regularities. The semantic of the article in Wiki can be derived from the labels and the info box in the pages, which can be formally represented by ontology. Two important characters of Wiki ontology are proposed. One is there are n-ary relations in it. The other is the size of this ontology is enlarging. Moreover, we propose the evolution mechanism of the ontology of Wiki. The main contents of the paper are shown as follows. 1) With the data from YAGO, the method of constructing hypernetwork of Wiki ontology is proposed. 2) After the analyzing the basic topological quantities of the hypernetwork, we propose an evolving mechanisms of the hyperedge growth. 3) The results of analyzing and numerical simulations are given at last.

Key words: hypernetworks; Wikipedia; ontology; evolution mechanism; n-ary relations; YAGO; hyperedges, n-ary relations