

关联开放词表网络动态演变分析^{*}

贾君枝 李 衍

摘 要 本文选取关联开放词表数据集 LOV 作为研究对象,采用社会网络动态分析方法,使用 Rsiena 作为分析工具,对收集到的数据进行了统计描述分析和参数估计检验,探究影响 LOV 词表网络演化的因素及其演化规律。研究表明:词表网络的演化目前处于成熟期,网络规模增长相比之前较为缓慢;同质性、流行性、成熟度等因素对词表网络的演化产生正向积极影响;关联词表网络正朝着良性健康的方向发展,有助于词表生态系统的形成。图 4。表 4。参考文献 31。

关键词 关联数据 开放词表 社会网络 动态演变

Dynamic Evolution Analysis of Linked Open Vocabulary Network

Jia Junzhi Li Kan

Abstract: This paper selects LOV as the research object, uses the social network dynamic analysis method and Rsiena as the analysis tool, makes statistical description analysis and parameter estimation test on the linked data, and explores the factors affecting the evolution of LOV vocabularies network and its evolution law. The results show that: the evolution of vocabularies network is in a mature stage, and the growth of network scale is slower than before; homogeneity, popularity, maturity and other factors have a positive impact on the evolution of vocabularies network; the linked vocabularies network is developing in a healthy direction, which is conducive to the formation of vocabularies ecosystem. 4 figs. 4 tabs. 31 refs.

Keywords: Linked Data; Open Vocabulary; Social Network; Dynamic Evolution

1 引言

语义网发展着眼于对结构化或半结构化数据进行语义化的表达、关联和存储。无法关联的信息资源易成为封闭的孤岛,失去其数据价值。当特定的数据以关联数据形式发布时,首要任务应该是开发一个用于描述这些数据的词表^[1]。词表定义了类和属性的集合,分别表示概念和资源的属性或资源之间的关系,其作为关联数据中的纽带,是关联数据的底层控制基础。词表之间并不是孤立存在的,它们彼此之间互相联系、相互制约,共同形成了词表网络。开放知识基金会(Open Knowledge Foundation)主导的关联开放词表(Linked Open Vocabulary, LOV)项目始于 2011 年 3 月,作为数据提升研究项目的一部分,旨在促进可重用的关联词表生态系统的形成。据统计,

由词表与词表之间的连接所形成的 LOV 网络,有 97.61% 的节点度数不为零,即 97.61% 的词表至少和一个词表建立了 RDF 链接,有 44.94% 的节点度数超过 10^[2]。

LOV 体系结构主要由四个组件构成,分别是:跟踪分析、综合处理、数据访问和数据存储^[3]。其中,跟踪分析组件主要用于检查词表版本更新、存储词表本地版本、提取词表元数据并分析词表特性;综合处理组件为社区建议词表或编辑词表目录提供路径,包括验证词表元数据信息、插入及审查词表,以保证 LOV 数据集的高质量;数据访问组件主要提供对数据集的访问,有 LOV 搜索引擎、RDF 转储文档、SPARQL 端点和 LOV API 四种访问方法;数据存储组件提供了一种有效的数据存储方式。为确保 LOV 数据集的高质

^{*} 本文系国家社科基金项目“中文学术领域命名实体的知识图谱构建研究”(项目编号:18BTQ072)、中国人民大学科学研究基金项目“数据驱动下的知识组织发展研究”(项目编号:202230018)的成果之一。

量,插入词表需要满足一些严格的要求;使用语义本体 Web 语言进行描述,例如 RDF 或 OWL;在互联网上发布且可免费向用户提供;词表或术语均具有唯一 URI,并可用于检索;足够小,易于被其他词表部分或整体集成和重用。

LOV 专注于维护一个高质量的词表集合。为提升词表的可用性,使用 DCE、VOID、FOAF、DCAT 词表描述基本的元数据信息,并使用 VOAF (Vocabulary of a Friend) 表示词表间的依赖、扩展、指定、注释等关系。这些关系使得 LOV 词表集成为一个不断增长、相互关联构成的生态系统,并由同样不断增长的发布者、出版商、管理员等组成的社交网络提供维护和支持。

关联数据发布后并非处于静止状态,而是会进行修改、维护、增删等活动。相关研究显示,29 周以内,62.2%的资源不会发生改变,6.2%的资源会消失,23.2%的资源会以低频率改变,还有 8.4%的资源以高频率发生改变^[4]。因此,需要及时对发生改变的关联数据进行更新和维护,以确保其准确性及可靠性。由于词表的概念及关系所形成的词表网络处于不断演变中,再加上新词表不断建立和发布,使得词表网络的拓扑结构发生变化,词表网络的动态演变因素值得做进一步探究。

社会网络分析方法已经广泛应用于关联数据领域,利用该方法能够了解关联数据各节点之间的联系,揭示关系结构,反映节点之间的量化关系。但随着研究的深入,有两大问题亟待解决:一是关联数据之间互联关系的复杂性导致需要从不同的关系层面来分析数据网络;二是目前多从静态网络角度开展的数据网络分析指标单一,在分析网络演变及趋势方面具有很大的局限性。复杂网络的关键特点之一就在于节点可以根据不断产生和消失的连接而演化。鉴于 LOV 便于获取、质量高、元数据丰富等特点,本文选取 LOV 词表数据集为研究对象,采用动态社会网络研究方法,以 Rsiena 作为研究工具,对词表网络的演变规律进行探究,以期揭示词表间的动态关系,促进关联数据的良性发展,为开展标准化的语义网服务提供借鉴。

2 国内外相关研究综述

社会网络的动态演化性与其扩散、合作等功能密切相关,对其动态演化影响因素及演化探测的研究,有助于观察社会网络的结构特征、预测网络的演化趋势、掌控网络态势、发现网络异常事件等。学者们针对社会网络演化的规律,提出了不同的演化模型^[5]。Leskovec 等人^[6]采用图论的方法探讨社会网络中节点数量与网络直径之间的动态关联及其随时间演化的特征,发现节点间边的建立频率与已存在边数呈反比。Chakrabarti 等人^[7]采用数据挖掘技术提出了一个考量时间与结构因素的虚拟网络演化模型。随着复杂网络技术的不断发展与成熟,学者们尝试运用网络演化模型进行相应的实证研究。例如,Hoekman 等^[8]在引力模型中引入了空间权重矩阵,从距离和边界两个角度考察网络演化发展问题。Broekel 等^[9,10]采用二次指派程序(QAP)和指数随机图(ERG)模型对网络的动态演化过程进行了分析,表明社会、地理邻近、认知、组织、制度等是网络演化的重要影响因素。但这些模型仍存在一定的局限性,主要体现在无法使用连续的数据来对网络演化进行观测。因此,Balland 等^[11]采用网络动态模型中的随机行动者模型(SAO, Stochastic Actor-Oriented)分析了商业和技术两种网络的演化影响因素。Brouwer 等^[12]采用 Rsiena 研究高等教育中影响小组教学的社会因素,构建了包括群体与个人属性因素的纵向社会网络。Hatch 等^[13]采用动态网络研究方法,研究了三年时间内三个学区中与教学、公平和种族相关的管理员社交网络的演变情况,并考察了网络演变与每个学区改善所有学生成绩的努力之间的关系。吴江等^[14]选择国内最大的糖尿病社区甜蜜家园为研究对象,采用基于随机行动者模型的社会网络动态分析方法,探究用户个体属性和网络结构属性对用户朋友关系网络动态演化的影响。王灿等^[15]应用随机行动者模型探索影响大学生朋友网络演化的潜在因素,研究中国大学生的社会属性、行为属性和认知属性对朋友网络的影响。张振刚等^[16]借鉴 RFM (Recency, Frequency, Monetary) 模型的思

路,使用K均值聚类法基于平均出现时间长度、出现频率和组合能力三个指标对知识元素进行聚类,并使用随机行动者导向模型对知识网络的演化进行分析。

LOV经常被作为LOD的补充,已成为国内外关联数据词表重用领域的研究热点。目前学者主要从词表建模和重用的角度对LOV进行研究,倾向于使用数理统计和社会网络分析方法。有研究专注于提升LOV词表质量,Schaible等^[17]提出了为LOV词表推荐类和属性的自动推荐方法,有助于本体工程师根据最佳实践对关联数据进行建模。还有学者重点研究LOV中的词表重用工作,如Halpin等^[18]研究owl:SameAs属性在LOD中的不同用途;Alexander等^[19]描述了互联数据集词表(Void)的实现和设计,在设计时重用了LOV中的词表,并以FOAF和都柏林核心作为Void的补充。有研究重点探讨LOV词表被LOD数据集使用的情况,如Schaible等^[20]对79名LOD用户进行了调查,分析他们在关联数据中用于建模和使用的词表频率;Poveda等^[21]提供了LOD中196个LOV词表重用的统计数据,包含重用元素、重用比率及词表之间的引用等;Dividino等^[22]研究表明,LOD中不同词表的使用是随着时间的推移而发生变化的;Nogales等^[23]为深入了解LOV词表特征及其在关联数据云中的使用情况,对LOV目录中收集的数据进行了定量研究,发现LOV是一个异构网络,词表领域范围广泛且更多使用的是用于语义建模的词表如RDF、RDFS、OWL等,以及一些广泛使用的元数据标准如SKOS、DCTERMS、DCE等;贾君枝等^[24]采用数理统计及社会网络分析方法对LOV中词表的基本特征分布、集中趋势、离散程度以及网络结构特征进行了分析,并验证LOV网络结构遵循无标度特性和小世界理论;贾君枝等^[25]选取LOV词表为研究对象,采用社会网络分析方法集中分析基于7种不同链接谓词形成的词表网络的统计特性及结构特性。总体来看,已有研究大多从单一时间节点静态网络角度进行分析,对长时间不同阶段的词表网络结构演变趋势和影响因素的关

注较少,词表重用更多从词表、术语和重用方式角度进行研究。而LOV词表元数据丰富,元数据作为词表属性必然影响着词表重用的过程,哪些元数据能够反映词表重用时的主动选择需求,哪些因素能够成为影响词表重用过程的显性因素,目前的研究并未充分解决这个问题。

对于关联数据中由词表重用关系形成的词表网络而言,网络中心性、密度、聚类系数、模块度等网络指标已经不足以了解其网络的动态结构和演化规律,其节点用元矩阵构造之后,在出入度的基础上还应包括如同质性、成熟度、丰富度、流行性、网络结构等指标。在其网络动态演化过程中,这些指标均能够相互作用从而影响网络的演化,因此应作为网络演化因素的考量。

3 研究假设

相关研究表明,社会网络的形成和演变会受到用户个人属性和网络结构特性等因素的影响^[26]。词表本身具有丰富的属性特征,如类、属性、语言、主题、发布者、发布时间、流行性、成熟度,这些显著特征一定程度上影响着词表之间的关联。其中,“类”定义了词表所涉及的主要概念集合,明确其定义域及值域,能够对所表达的资源进行限制;“属性”用于描述概念之间的关系,是区分类的标准。类数和属性数可作为衡量词表语义丰富程度的标准。语言指构建词表所用的语言,“语言”数越丰富,表明该词表使用地区越多,使用范围越广。“主题”指词表所在领域,LOV官网定义了43个词表主题类型,单个词表可被多个主题所标识。“流行性”指该词表被LOD数据集使用的次数,用以表明该词表与LOD数据集的关联紧密程度。“成熟度”主要由发布者类型和发布时间来决定,“发布者”包含机构和个人两种类型,通常由机构发布且发布时间较早的词表,其成熟度较高。词表之间以RDF链接的关联,提供了词表网络中已有的数据参数。为探究词表网络的形成因素,现从词表的属性和网络结构特性角度研究词表网络变化的影响因素。

3.1 同质性

同质性理论认为社会关系更容易建立在具有相似或相同属性的个体之间^[27]。词表的主题、语言、发布者等属性可能是影响词表关系建立的普遍性因素。因此,基于同质性理论,提出以下假设:

H1a:词表的主题对词表连边的形成具有显著影响,两个相同主题的词表更容易建立链接。

H1b:词表的发布者对词表连边的形成具有显著影响,两个相同发布者的词表更容易建立链接。

H1c:词表的语言对词表连边的形成具有显著影响,两个相同语言描述的词表更容易建立链接。

3.2 丰富性

词表的丰富性主要体现在词表的类、属性、实例、语言数量的多少,一般认为词表的语言、类、属性、实例数越多,词表描述资源的丰富性就越高。因此,将词表的语言数、类数、属性数、实例数作为丰富性的衡量指标,提出以下假设:

H2a:词表的语言数对词表连边的形成具有正向影响,词表更倾向于与使用较多种语言描述的词表建立链接。

H2b:词表的类数对词表连边的形成具有正向影响,词表更倾向于与定义了较多类的词表建立链接。

H2c:词表的属性数对词表连边的形成具有正向影响,词表更倾向于与定义了较多属性的词表建立链接。

H2d:词表的实例数对词表连边的形成具有正向影响,词表更倾向于与定义了较多实例的词表建立链接。

3.3 流行性

词表的流行性主要体现在词表的度、入度、出度的多少,一般认为词表的度值越高,词表的流行性就越高,就越易与其它词表建立链接。因此,将词表的度、入度、出度值作为流行性的衡量指标,提出以下假设:

H3a:词表的度数对词表连边的形成具有正向影响,词表更倾向于与度值较高的词表建立

链接。

H3b:词表的入度值对词表连边的形成具有正向影响,词表更倾向于与入度值较高的词表建立链接。

H3c:词表的出度值对词表连边的形成具有正向影响,词表更倾向于与出度值较高的词表建立链接。

H3d:词表被数据集使用的次数对词表连边的形成具有正向影响,词表更倾向于与被数据集使用次数较多的词表建立链接。

3.4 成熟度

词表的成熟度主要体现在发布者类型、发布时间等因素。机构发布的词表具有识别度和依赖度更高的发布主体,更为重要的是词表均具有URI的稳定性和Web上的可用性特征,并配备了适当的版本管理政策,提供了词表的标准格式以及高质量的元数据文档。而通过个人发布的词表,尽管术语仍可以在数据集中找到,但存在管理和维护不善的情况,例如存在URI缺失、词表内容滞后、更新频率较低等问题。因此,将词表的发布者、发布时间作为成熟度衡量指标,提出以下假设:

H4a:发布者是机构的词表对词表连边的形成具有正向影响,词表更倾向于与机构发布的词表建立链接。

H4b:发布时间较早的词表对词表连边的形成具有正向影响,词表更倾向于与发布时间较早的词表建立链接。

3.5 网络结构

词表网络的形成是一个动态的复杂过程,除了受词表的基本属性影响外,还会受到网络结构效应的影响,如互惠性、平衡性、三元闭包等。互惠性指在有向网络中随机选择一条有向边,从节点A指向节点B,那么存在从节点B指向节点A的有向边的概率^[28]。互惠性一般用在有向网络中,用于衡量网络中两个节点形成相互链接的程度。平衡性主要用于社交网络的社会关系中,社会关系可划分为正关系和负关系,由于词表之间

的连接并没有情感因素,因此不在本文进行探讨。三元闭包理论起源于现实生活中的经验性判断,即:与其他人相比,人们与朋友的朋友更可能建立朋友关系^[13]。在词表网络中,词表A与词表B之间建立了链接,词表B与词表C之间建立了链接,那么词表A与词表C之间也可能建立链接关系,通过词表B的中介作用,词表A与词表C可以快速建立链接。在无向网络中,三元闭包等同于关系传递,可以用来研究网络演化的倾向性。因此提出以下假设:

H5a:互惠性对词表连边的形成具有正向影响,词表A建立了与词表B的链接,那么词表B也倾向于与词表A建立链接。

H5b:三元闭包对词表连边的形成具有正向影响,词表更倾向于与链接词表的相关链接词表建立链接。

4 研究方法与数据采集

4.1 研究方法

本研究采用社会网络动态分析工具 SIENA 进行动态网络分析,该工具主要基于行动者导向模型(Actor-Oriented Dynamic Network Approach)分析网络的动态变化^[29]。对于需要两期以上连续观测的纵向网络数据来说,期间节点可以改变它们之间的链接,即节点可以建立或取消连边。SIENA 是一个适于分析这种网络动态变化的工具。由于网络演化的非确定性还掺杂了随机因素的影响,因此也将行动者导向模型称为随机行动者导向模型。网络节点之间连边的变化主要基于节点内生和外生两种原因,内生原因在于网络结构的演变,外生原因在于行动者属性的变化驱动^[30]。SIENA 在实际使用的过程中可随机产生马尔可夫过程(即随机效用模型),能够更好地预测社会网络的演化。该工具利用马尔可夫方法的仿真模拟来估计模型的参数,每一个要估计的模型参数 β 可以通过 t 统计量来判断显著性。每一个要估计的参数都将与实际观测的统计数据相对比,差值越小表明参数估计的收敛性越好,当 t 的绝对值小

于 0.1 时即代表观测值的参数估计是有效的^[31]。

行动者导向模型以网络变化作为因变量,以节点的属性、网络结构和其他随机变量作为节点改变连边的起因,主要包含速率函数和效应函数两种。速率函数主要是由网络中节点关系的变化频率来反映不同阶段下网络节点的演变速度,行动者导向模型主要运用逻辑回归模型来模拟节点对连边选择的概率。效应函数用来描述网络节点在某种网络结构下倾向的选择,所造成的对网络结构的影响。目标函数是各种效应函数的线性组合 $f_i(\beta, x) = \sum_k \beta_k s_{ik}(x)$, 其中 $f_i(\beta, x)$ 是目标节点 i 在网络结构 x 下的目标函数, $s_{ik}(k)$ 为各种效应函数。效应函数不仅取决于网络结构,还取决于节点的属性。基于此,网络变化和属性变化分别由下列不同的函数所决定^[13]:

$$\int_i^{net}(\beta, x) = \sum_k \beta_k^{net} s_k^{net}(x) \quad \text{公式(1)}$$

$$\int_i^{beh}(\beta, x) = \sum_k \beta_k^{beh} s_k^{beh}(x) \quad \text{公式(2)}$$

以上两个公式中, s^{net} 表示影响网络结构变化的效应函数, s^{beh} 表示影响属性变化的效应函数。 β 是统计参数,如果 β 等于 0,则表示相应的效应函数对网络演化没有任何影响;如果 β 为正,则表示效应函数对网络演化有正向影响;如果 β 为负,则表示效应函数对网络演化有负向影响。

SIENA 可通过 R 语言的 RSiena 包实现,根据需求设置相应的模型。而对模型的设定就是要选择将哪些效应纳入到模型中,可使用 `includeEffects()` 函数来进行设定。其效应主要可以分为两大类,即网络结构效应和属性效应。网络结构效应主要是用来预测节点之间关系的变化,即新关系的生成、已有关系的破裂或保持不变,常见的包括密度、互惠性、三元闭包、循环性等。属性效应主要是起到属性对关系的预测作用,即这种属性能否成为影响节点关系建立的因素。例如要探究三元闭包效应对模型的影响,可使用 `groupEffects <- includeEffects(groupEffects, transTrip)` 命令来实现。表 1 列举了几种基本的效应函数。

表1 Rsiena 中的部分效应函数

网络结构效应	计算公式	说明
互惠性	$\sum_j x_{ij}x_{ji}$	节点倾向于建立互惠联系
三元闭包	$\sum_{j,h} x_{ih}x_{ij}x_{jh}$	节点倾向于和目标词表已建立联系的词表建立联系
属性效应	计算公式	说明
alterX	$\sum_j x_{ij}v_j$	属性值高的节点更受欢迎
sameX	$\sum_j x_{ij}I\{v_i=v_j\}, \text{where } I\{v_i=v_j\} = 1, \text{ if } v_i=v_j, \text{ and } 0 \text{ otherwise}$	属性值相等节点之间更容易建立链接

其中,alterX 效应表示是否某种属性会使节点接受更多(参数为正)或更少(参数为负)的关系,sameX 效应表示节点是否更倾向于与跟自己在属性水平上完全相同的节点建立关系。除网络效应之外,本文衡量了 alter、same 两种属性效应。在进行参数估计前,需要对所使用的属性类别进行编码。以机构效应为例,将发布者为机构的词表编码为 1,将发布者为个人的词表编码为 0,如果属性“机构”的 alter 效应显著,则表明属性值编码高的发布者机构的词表更受欢迎,更容易被其他词表使用并与其建立链接关系。以语言效应为例,如果属性语言的 same 效应显著,则表示同语言的词表更容易建立链接关系。由于 LOV 中词表使用了 50 种语言进行描述,因此分别对 50 种语言按使用频率从 0 到 49 开始进行编码。由于具有多语言描述的词表数量较少,若出现多种语言,则选取应用最广的语言为该词表的描述语言。其他字符串类型属性的编码规则与此类同。

4.2 数据采集

截至 2020 年 4 月,LOV 平台已发布 700 个词表。本文的数据来自 LOV 官网中的 RDF dump 转储文件和 SPARQL 查询端口,查询词表列表及词表基本信息。LOV 包含一个名为 VOA 的词汇总表,旨在提供词表的一些共同特征,描述词表之间的关系。VOA 的属性可以分为两种类型:一种用于描述语言、类、属性、实例等常见特征属

性,另一种用于描述词表的相关性。本研究在第一类属性的基础上增加发布者、发布时间、主题和被 LOD 数据集使用次数四个变量,针对第二类属性构建一个社会网络分析(SNA)图,将词表定义为节点,词表关系属性定义为边。由于词表之间的 RDF 链接是有向的,因此 LOV 网络是一个有向网络。最终得到词表名称、URI、语言、语言数、属性数、类数、实例数、发布者、机构类型、发布时间、主题、被 LOD 数据集使用次数、入度、出度、度数等 15 个特征值。通过对数据进行整理,部分属性情况如表 2 所示。

表2 部分属性数据统计结果

变量	最小值	最大值	均值
类数	0	2853	43.27
属性数	0	7033	54.96
语言数	1	41	1.45
被 LOD 数据集使用数	0	504	4.74
入度	0	1471	9.46
出度	0	43	8.31

将数据导入到 R 软件中,部分属性数据如图 1 所示。

利用 det:issued 和 dc:issued 谓词在 LOV 的 SPARQL 查询端口查询词表的发布日期,对结果进行去重和整理后得到 623 条记录,查询表达式和部分结果如图 2 所示。由于某些词表并未提供发布日期元数据,因此在这些词表的版本描述中提取首次发布时间作为发布日期。

language_number	classes	properties	datatype	instances	in.degree	out.degree	expressivity	expressivity_number	LOD	publication.l
1	12	15	0	0	0	14	RDF/RDFS	2	0	20160415
2	20	26	0	6	2	8	RDF/RDFS/OWL	3	0	20120614
3	6	8	0	0	5	5	RDF/RDFS	2	0	20120703
4	1469	0	0	0	0	3	RDF/RDFS/OWL	3	0	20130424
5	1	7	0	7	0	7	RDF/RDFS	2	0	20140123
6	4	13	0	0	28	11	RDF/RDFS	2	3	20120522
7	49	16	0	8	0	13	RDF/RDFS/OWL	3	0	20080623
8	14	13	5	15	0	7	RDF/RDFS	2	0	20020418
9	4	9	0	10	0	7	RDF/RDFS	2	0	20171009
10	48	83	0	0	2	10	RDF/RDFS	2	0	20110630
11	15	10	0	0	0	11	RDF/RDFS/OWL	3	8	20080514
12	18	76	0	0	0	6	RDF/RDFS	2	0	20140626
13	152	122	0	62	5	7	RDF/RDFS/OWL	3	40	20030210
14	17	15	0	13	2	5	RDF/RDFS/OWL	3	33	20030210
15	11	4	0	23	1	9	RDF/RDFS	2	0	20050731
16	2	14	0	0	0	3	RDF/RDFS	2	0	20120824
17	17	87	0	19	0	13	RDF/RDFS	2	0	20130323
18	1	13	0	0	5	16	RDF/RDFS/OWL	3	0	20100913
19	45	50	0	235	0	7	RDF/RDFS/OWL	3	0	20121115
20	19	30	0	10	0	7	RDF/RDFS	2	6	20100104

图1 部分属性数据

Query Examples		
1 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>		
2 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>		
3 PREFIX owl: <http://www.w3.org/2002/07/owl#>		
4 PREFIX vann: <http://purl.org/vocab/vann/>		
5 PREFIX voaf: <http://purl.org/voccommons/voaf#>		
6 PREFIX dcterms: <http://purl.org/dc/terms/>		
7 PREFIX doe: <http://purl.org/dc/elements/1.1/>		
8 ### SKOS properties used in vocabularies to describe classes		
9 SELECT DISTINCT ?z ?y {		
10 GRAPH ?g{		
11 ?x dcterms:issued ?y.		
12 ?x a voaf:Vocabulary.		
13 ?x vann:preferredNamespacePrefix ?z.		
14 }} ORDER BY ?x		
Table	Raw Response	Search:
z	y	
1	cl	"2014-03-26"^^xsd:date
2	eac-cpf	"2011-01-28"^^xsd:date
3	bag	"2013-11-05"^^xsd:date
4	awol	"2005-10-22"^^xsd:date
5	cbs	"2018-05-02"^^xsd:date
6	bimerr-op	"2020-02-01"^^xsd:date
7	faldo	"2012-09-07"^^xsd:date
8	caresses	"2019-06-10"^^xsd:date

图2 词表发布日期 SPARQL 查询代码及部分结果

以发布年份为 x 轴,发布个数为 y 轴绘制词表发布数量统计柱状图,如图 3 所示。可以看出,

LOV 中词表最早发布时间为 1999 年,是 rdf 词表。自 2010 年开始 LOV 中词表发布个数迅速增

长,在 2012 年达到了顶峰,其后发布数量开始呈现下滑趋势,但平均发布数量仍高于 1999 至 2009 年 11 年间的平均值。按照 2010 年以前、

2010 至 2014 年和 2015 至 2020 年进行划分,分别为网络一、网络二和网络三,在图 3 中分别用黑色、深灰色、浅灰色显示。

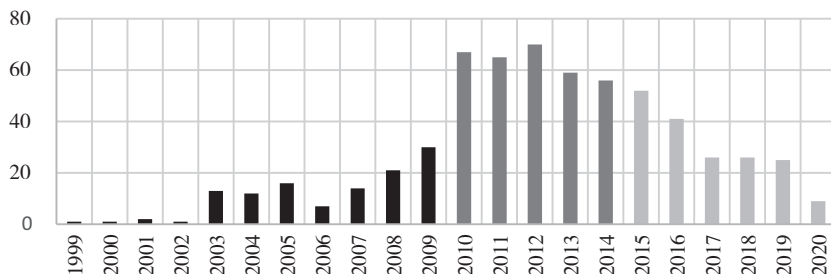


图3 词表发布数量统计

由于 Rsiena 只能识别矩阵形式的网络数据,而收集到的是列表形式的关系数据,不能用 Rsiena 进行计算,需要对数据格式进行转换。因此,基于每期数据,将关系数据转变为关系矩阵导入到 R 中。若 A 和 B 之间具有链接,则其矩阵中相应位置的数值为 1,否则为 0。

5 结果分析

5.1 词表网络演化描述性统计分析

本研究涉及三期网络数据,如表 3 所示,网络一

的连边数量为 616 条,网络二的连边数量为 1200 条,网络三的连边数量为 1326 条,词表互联关系连边稳步增长。利用社交网络分析工具 Gephi 对三期的网络数据进行可视化后得到网络拓扑图(见图 4),可以看出词表网络在不断地运动和变化,且网络一中词表节点和边数明显少于网络二,而网络二与网络三之间的节点数与边数的增幅均小于之前。整体而言,根据词表发布数量的变化可将词表网络演化的过程划分为三个阶段,即初始期、极速发展期和成熟期,分别与三期网络相对应。

表3 三期网络数据基本特征

时间	节点数	关系数	新增连边数	平均度	网络密度	平均路径长度	平均聚类系数
2010 年以前	126	616	无	4.889	0.039	1.313	0.296
2010—2014 年	1001	4089	3473	4.085	0.004	6.262	0.140
2015—2020 年	1557	6777	2688	4.353	0.003	6.423	0.172

由表 3 可知,第一,词表网络演化保持活跃。三个时间段均有新的词表发布,并通过一定的重用关系与其它词表之间建立了链接,形成新的词表关系,词表网络规模不断扩大。其中初始期对应于词表网络发展的起步阶段,词表数量和词表链接均较少;词表网络结构非常紧密,产生明显的集群性;词表之间的链接多呈现“一对一”的单链网络模式与“一对多”的星型网络模式,复杂网络结构较少。在极速发展期,词表数量和词表链

接的增长速度明显加快;词表网络整体结构松散,网络密度明显降低;此时除了在初始期存在的两种网络模式外,还增加了“多对多”的复杂网络组合模式,词表重用范围较之前有所扩大。在此之后,词表网络集群逐渐进入成熟期,网络规模增长趋缓,词表网络紧密程度提高。第二,新词表的构建注重对已发布词表的重用,词表网络逐渐优化。成熟期的词表网络,其节点数、关系数等网络指标的数值呈现递增趋势,表明各界对关联数据词表

发布的重视程度提高,越来越多的机构发布了自己的词表。平均度、网络密度、平均路径长度及平均聚类系数则呈现出先增后减的趋势,表明初始期词表较少与外部资源建立关联;极速发展期词表在增加数量的同时也注重了与外部资源之间的互联,关联化程度提高,但由于短时间内词表发布数量的增速高于关系数量的增速,而导致了平均度等网络指标的下降;而在成熟期,词表在发布的过程中更加注重资源之间的互联程度,体现出词表关联网络在逐渐优化。第三,词表网络具有小世界网络特性,即词表网络具有高聚类系数与低平均路径。高聚类系数表明词表间的链接并不随机,与词表A相连的两个词表B和C之间的链接可能性也很高;低平均路径指即便词表数量不断增加,由于存在核心节点,词表间距离依然较近,并且词表网络的小世界网络特性愈加明显。三期网络数据的平均路径长度逐渐增大,表明节点之间平均需要的中介节点数量越来越多。大多数的词表都会重用标准词表如rdf、rdfs、

owl、foaf等,因此,具有高度数节点的词表度数越来越高;同时也有越来越多的词表很少被其它词表重用,低度数节点的词表数越来越多,造成了词表网络向小世界网络演化的局面。作为网络中介主体的度中心性和中介中心性高的高度数节点,是词表网络中的关键节点。这有助于缩短词表之间链接的平均路径长度,提高词表重用效率,促进数据关联和知识共享。

由此,可以梳理出词表重用网络演化的基本过程和规律。从演化过程来说,词表重用网络经历了由少量词表作为核心的桥梁模式并最终发展为多资源互联的成长模式的过程。从空间特征来看,随着词表重用模式从简单到复杂的逐渐升级演化,词表重用的数量和类型不断丰富,而由于词表间重用关系的不断建立,词表网络最终成为一个节点数量庞大、合作关系紧密、网络结构脆弱性较低的稳定网络,呈现出多元化动态增长并以知识内在关联为基础的网状结构。



图4 三期词表网络拓扑图

5.2 模型参数估计分析

为了更好地解释词表网络结构的演化,设定尽可能多的参数进行分析,利用RSiena进行模型参数估计和显著性检验,所有参数拟合结果如表4所示。其中, t -统计量=(观察值-样本均值)/标准误差, t -统计量绝对值小于0.1表明参数在估计过程中收敛,即得到的参数估计结果是有效的, t 绝对值大于2时表明结果显著。

其中速率函数Rate反映了不同阶段的词表网络演变速度,分别对应从词表网络一到网络二

再到网络三演变的两个时期里,每个词表节点连边改变的程度。由表4可知,网络一到网络二的时期里,词表连边改变程度更加剧烈,也对应了在2010至2014年间是词表网络极速增长的时期,大量的词表被发布,并在词表构建过程中增加了对已有词表的重用,词表互联程度显著提高。2015至2020年间,词表连接速率减弱,主要是由于词表发布的速率减少,但增加了与外部资源的链接,总体而言,词表网络的连通性也在提高。

表4 参数拟合结果

变量名	参数估计	标准误差	t	t-统计量	结论
Rate parameter period1	4.7682	0.0942	50.6178	—	
Rate parameter period2	3.1045	0.0716	43.3589	—	
Same Theme	1.1033	0.0888	12.42	-0.03	H1a 假设成立
Same Agent	3.0533	0.137	22.28	0.0046	H1b 假设成立
Same Language	1.9294	0.3941	4.8957	0.099	H1c 假设成立
Language alter	-0.6609	0.1232	-5.3644	-0.0149	H2a 假设不成立
Class alter	-0.3412	0.0353	-9.6657	0.0255	H2b 假设不成立
Property alter	0.1522	0.0343	4.4373	0.0211	H2c 假设成立
Instance alter	0.6415	0.0976	6.5727	0.0133	H2d 假设成立
Degree alter	0.0037	0.0006	6.1666	-0.0467	H3a 假设成立
In-degree alter	0.009	0.0012	7.5	-0.01	H3b 假设成立
Out-degree alter	-0.1973	0.101	-1.95	0.041	H3c 假设不成立
LOD alter	1.4701	0.0785	18.7273	0.046	H3d 假设成立
Agent	0.5708	0.0928	6.1508	-0.0126	H4a 假设成立
Publication Time	-0.1422	0.0083	-17.1325	0.0172	H4b 假设成立
Reciprocity Preference	0.5991	0.3698	1.62	-0.0303	H5a 假设不成立
Transitive triads	0.8222	0.1449	4.7	0.0218	H5b 假设成立

同质性对词表网络演化的影响。其中 Same Theme、Same Agent、Same Language 参数的 t-统计量绝对值均小于 0.1, t 绝对值大于 2, 表明词表更倾向于与相同主题、相同机构的词表建立链接。这主要由于相同主题的词表代表了同一领域, 能够有效地集成领域知识; 相同机构的词表能够促进词表之间的互操作, 简化词表构建过程; 由于当前词表大多使用英语作为描述语言, 且相同机构出版的词表大多采用同一语言, 因此词表更倾向于与相同语言的词表建立链接。

丰富性对词表网络演化的影响。语言数和类数的参数估计值为负, 因此类数和语言数并不对词表的互连情况形成正向显著影响。属性数和实例数对词表的互连情况形成正向显著影响。这主要是由于词表的语言数、类数并不作为词表重用时候的选择因素, 由于词表重用中对属性的重用远高于对类的重用, 属性数、实例数高的词表更容易被其它词表重用。

流行性对词表网络演化的影响。其中 Degree、In-degree、被 LOD 数据集使用的次数的 t-统计量绝对值均小于 0.1, t 绝对值大于 2, 说明度

数、入度和词表被 LOD 数据集使用的次数对词表连边的形成具有显著影响, 词表更倾向于与度数高、入度高、在 LOD 云中使用较多的词表建立链接。这主要是由于高度数节点大多是通用词表, 这些词表较为稳定、发布较早、发布者权威, 容易被新发布的词表重用。其中出度并未对词表的连边形成显著正向影响, 主要由于词表出度均较低, 差别并不明显, 且出度表现的是词表对其它词表的需求程度, 并不作为词表重用时的倾向。

成熟度对词表网络演化的影响。其中 Agent 和 Publication Time 的 t-统计量绝对值均小于 0.1, 且 t 绝对值大于 2, 表明机构和发布时间属性均对词表网络的演化具有显著影响。Agent 的参数估计值为正, 表明词表更倾向于与机构发布的词表进行链接, 而不是个人代理发布的词表。这与机构发布的词表更稳定、更新频率更高有关。这些机构大多致力于为知识共享而服务, 其发布的词表一般具有较好的可维护性和管理性, 能够保证长期 URI 稳定性, 可定期进行主动性词表维护。Publication Time 的参数估计值为负, 表明词表更倾向于与发布时间较早的词表进行链接。

网络结构对于词表网络演化的影响。Reciprocity Preference 和 Transitive triads 的 t -统计量绝对值均小于 0.1, 但 Reciprocity Preference 的 t 绝对值小于 2, 表明互惠性并不对词表连边形成显著影响, 词表链接具有方向性且具有方向偏好, 后发布词表较多与发布时间较早的标准词表建立链接, 而发布时间较早的词表却较少与后发布词表进行链接。Transitive triads 的 t 绝对值大于 2, 表明三元闭包对词表连边具有显著影响。即词表网络具有明显的关系传递作用, 词表间的链接并不随机, 词表更倾向于与链接词表建立关系的词表进行链接, 由于中介词表的作用, 会促进相关词表进行重用。这样的结构使描述同一实体的资源互相连接, 丰富了描述实体的多样性。

6 结语

本文利用社会网络动态分析方法, 基于三期网络数据, 提出可能作为影响词表网络演化因素的假设, 对词表网络演化过程进行统计描述分析, 并对假设进行参数检验, 明确词表网络演化的动态及影响因素。根据研究结果, 词表网络的演化目前处于成熟期, 网络规模增长相比之前较为缓慢, 但还是有新词表在不断发布, 表明各界对词表发布的重视程度较高。同质性、流行性、成熟度等因素对词表网络的演化产生正向积极影响, 词表更倾向与具有成熟度高、流行性高、同领域等特点的词表建立链接关系, 因此在未来进行词表构建时, 应尽量关注这些因素, 减少用户在词表建模过程中产生的数据冗余, 促进词表重用、知识共享。研究结果也表明, LOV 提供了在词表生态系统中进行搜索和质量评估的功能, 有助于推动词表不断优化; 基于词表重用而形成的关联词表网络正在朝着良性健康的方向发展, 有助于词表生态系统的形成。

参考文献

- 1 PovedaVillalón M, et al. The landscape of ontology reuse in linkeddata [EB/OL]. [2023-03-10]. <https://oa.upm.es/14385/1/povedaEtAl-OEDW>
- 2 In LOV at a glance... [EB/OL]. [2023-03-01]. <https://lov.linkeddata.es/dataset/lov/about>.
- 3 Vandenbussche P Y, et al. Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web[J]. Semantic Web, 2017, 8(3): 437-452.
- 4 郭少友, 闫振斌. 动态关联数据: 概念、要求与实现机制[J]. 图书馆学研究, 2014(9): 7-10, 59.
- 5 方滨兴, 等. 社交网络分析核心科学问题、研究现状及未来展望[J]. 中国科学院院刊, 2015(2): 187-199.
- 6 Jure Leskovec, et al. Graphs over time: densification laws, shrinking diameters and possible explanations[EB/OL]. [2023-03-10]. <https://dl.acm.org/doi/10.1145/1081870.1081893>.
- 7 Deepayan Chakrabarti, et al. Evolutionary clustering[EB/OL]. [2023-03-10]. <https://faculty.mcombs.utexas.edu/deepayan.chakrabarti/my-www/papers/kdd06-evolutionary.pdf>.
- 8 Hoekman J, et al. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe[J]. Research Policy, 2010, 39(5): 662-673.
- 9 Broekel T, Boschma R. Knowledge networks in the Dutch aviation industry: the proximity paradox[J]. Papers in Evolutionary Economic Geography, 2012, 12(2): 409-433.
- 10 Broekel T, Hartog M. Explaining the structure of inter-organizational networks using exponential random graph models: does proximity matter[J]. Industry & Innovation, 2013, 20(3): 277-295.
- 11 Balland, et al. The Dynamics of Technical and Business Knowledge Networks in Industrial Clusters: Embeddedness, Status, or Proximity[J]. Economic Geography, 2016, 92(1): 35-60.

- 12 Jasperina Brouwer, et al. Longitudinal Peer Network Data in Higher Education [EB/OL]. [2023-03-10]. <https://research.rug.nl/en/publications/longitudinal-peer-network-data-in-higher-education>.
- 13 Hatch T, et al. Instruction, equity, and social networks in district-wide improvement[J]. Journal of Professional Capital and Community, 2020, 5(1): 72-91.
- 14 吴江,等. 基于随机行动者模型的在线医疗社区用户关系网络动态演化研究[J]. 情报学报, 2017(2): 213-220.
- 15 王灿,唐锡晋. 基于随机行动者模型及 iView 分析的大学生朋友关系研究[J]. 系统科学与数学, 2018(2): 163-176.
- 16 张振刚,罗泰晔. 基于 RFM 模型和随机行动者导向模型的技术机会识别[J]. 情报学报, 2021(1): 53-61.
- 17 Johann Schaible, et al. LOVER: Support for Modeling Data Using Linked Open Vocabularies [EB/OL]. [2023-03-10]. <https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=02ED2770BDE9C35C6286C89FD68B68ABoi=10.1.1.428.9027&rep=rep1&type=pdf>.
- 18 Harry Halpin, et al. When owl: sameas isn't the same: An analysis of identity in linked data [EB/OL]. [2023-03-10]. https://link.springer.com/chapter/10.1007/978-3-642-17746-0_20.
- 19 Alexander K, et al. Describing linked datasets [EB/OL]. [2023-03-10]. [https://www.research.lancs.ac.uk/portal/en/publications/describing-linked-datasets\(e57de88f-4176-4de1-98b8-68754353806f\)/export.html](https://www.research.lancs.ac.uk/portal/en/publications/describing-linked-datasets(e57de88f-4176-4de1-98b8-68754353806f)/export.html).
- 20 Johann Schaible, et al. Survey on common strategies of vocabulary reuse in linked open data-modeling [EB/OL]. [2023-03-10]. https://link.springer.com/chapter/10.1007/978-3-319-07443-6_31.
- 21 Poveda, et al. The Landscape of Ontology Reuse in Linked Data [EB/OL]. [2023-03-10]. <https://oa.upm.es/14385/>.
- 22 Dividino, et al. Change-a-LOD: Does the Schema on the Linked Data Cloud Change or Not [EB/OL]. [2023-03-10]. https://ceur-ws.org/Vol-1034/DividinoEtAl_COLD2013.pdf.
- 23 Nogales A, et al. Measuring vocabulary use in the Linked Data Cloud [J]. Online Information Review, 2017, 41(2): 252-271.
- 24 贾君枝,李衍. LOV 词表特征及网络结构分析[J]. 情报科学, 2021(3): 166-171.
- 25 贾君枝,李衍. 关联开放词表的互联关系分析[J]. 情报理论与实践, 2022(6): 114-121.
- 26 Snijders T, et al. Manual for SIENA version 3 [J]. Times Literary Supplement Tls, 2005, 14(2): 257-258.
- 27 Feld S L. The focused organization of social ties [J]. American Journal of Sociology, 1981, 86(5): 1015-1035.
- 28 Garlaschelli D, Loffredo M I. Patterns of link reciprocity in directed networks [J]. Physical Review Letters. 2004, 93(26): 268701.
- 29 Tom A. B. Snijders. Stochastic actor-oriented models for network change [J]. Journal of Mathematical Sociology, 1996, 21(1-2): 149-172.
- 30 陈泉. R 软件的知识结构与开发者合作结构及其演化研究[D]. 广州: 华南理工大学, 2013.
- 31 Snijders T, et al. Modeling the co-evolution of networks and behavior [EB/OL]. [2023-03-10]. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315091655-3/modeling-co-evolution-networks-behavior-tom-snijders-christian-steglich-michael-schweinberger>.

(贾君枝 教授 中国人民大学信息资源管理学院 中国人民大学数字人文研究院, 李衍 助理馆员 太原科技大学图书馆)

收稿日期: 2022-09-08