

# 语义 Web 链接结构分析之综述

葛唯益 程 龚 瞿裕忠

(东南大学计算机科学与工程学院 南京 210096)

**摘 要** 随着语义 Web 研究的发展,其数据量也不断增长,要实现语义 Web 追求的目标——数据的共享和重用,语义 Web 上的实体搜索和文档搜索必不可少。而面对这样不断增长的数据以及不同于传统 Web 的搜索要求,就需要使用链接结构分析来指导语义 Web 上的搜索。同时,语义 Web 的发展现状也无时无刻不吸引着研究人员的关注,而链接结构分析对于揭示其宏观结构起着关键作用。分别从实体和文档两个粒度对面向语义 Web 链接结构分析的研究进行总结,特别关注链接模型的构建以及链接结构分析方法的应用。

**关键词** 语义 Web, 链接模型, 链接分析

中图法分类号 TP311 文献标识码 A

## Linkage Analysis of the Semantic Web: The State of the Art

GE Wei-yi CHENG Gong QU Yu-zhong

(School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

**Abstract** With the development of Semantic Web research, more and more data are emerging on the Semantic Web. To pursue the goal of the Semantic Web, sharing and reusing data, entity search and document search are two essentials. Linkage analysis on the Semantic Web is a tool which could direct Semantic Web search. Besides, the state of the art of the Semantic Web always attracts researchers, whereas linkage analysis is a key to the mining of the macrostructure of the Semantic Web. This paper surveyed the state of the art of research on linkage analysis of the Semantic Web at two granularities, entity-level and document-level. In particular, we focused on various linkage models as well as the use of various linkage analysis methods.

**Keywords** Semantic Web, Linkage model, Linkage analysis

## 1 引言

语义 Web (Semantic Web, 也称语义网) 提供一个公共框架, 使得数据的共享和重用可以跨越应用系统、企业和社区的边界<sup>1)</sup>。语义 Web 以 RDF (Resource Description Framework, 资源描述框架) 作为数据模型, 以 URI 作为标识机制, 能够将各种不同应用中的数据和服务集成起来。同时, 本体 (ontology) 在语义 Web 中扮演着重要角色, 它使用 RDFS 和 OWL 等 Web 本体语言显式地描述实例 (individual) 的集合——类 (class) 以及实例间的关系——属性 (property) 等。在语义 Web 中, 通常将类、属性和实例统称为实体 (entity), 其中类和属性统称为术语 (term)。

随着语义 Web 研究的深入, 尤其是链接开放数据项目 (Linking Open Data<sup>2)</sup>) 的推动, 在社会网络、生物医学和电子商务等领域的语义 Web 数据不断增加。随着数据量的不断攀升, 要实现语义 Web 追求的目标——RDF 数据的共享和重用, 实体搜索和文档搜索必不可少。但面对如此巨大的数

据量以及不同于传统 Web 的搜索要求, 受到传统 Web 链接图分析对 Web 搜索帮助的启发<sup>[1]</sup>, 需要对语义 Web 上的数据使用链接结构分析的方法进行建模和分析, 以便为语义 Web 搜索引擎的数据搜集、索引和排序的算法设计提供决策依据。

在语义 Web 发展的不同阶段, 研究人员已经从各种角度采用不同的手段去分析它的特征。Cardoso<sup>[2]</sup> 基于调查问卷的形式, 对语义 Web 的应用领域、开发工具、本体语言等方面进行调研, 得出它们的使用现状; 文献 [3-6] 基于统计的方式对文档和实体相关的问题 (比如文档的规模和增长趋势、文档的地区和站点分布等) 进行分析; 文献 [6-8] 等同样使用统计的方式分析了本体中描述逻辑相关的问题, 如使用的语言、描述能力和使用错误等。但这些工作都不能给出明确的语义 Web 的宏观结构。受到传统 Web 链接图模型分析对 Web 结构探索工作<sup>[9]</sup> 的启示, 语义 Web 研究人员正越来越关注语义 Web 图模型的创建以及语义 Web 链接结构的分析, 试图揭示出语义 Web 的宏观结构。

到稿日期: 2009-04-10 返修日期: 2009-07-05 本文受国家自然科学基金项目 (60773106) 资助。

葛唯益 博士生, 主要研究方向为语义 Web 分析, E-mail: wyge@seu.edu.cn; 程 龚 博士生, 主要研究方向为语义搜索、数据集成、数据挖掘; 瞿裕忠 教授, 博士生导师, CCF 会员, 主要研究方向为语义 Web、软件工程。

<sup>1)</sup> <http://www.w3.org/2001/sw/>

<sup>2)</sup> <http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

链接结构分析<sup>[10]</sup>在很多领域都发挥了重要的作用,尤其在 Web 挖掘中<sup>[11]</sup>。而语义 Web 链接结构分析也是语义 Web 挖掘<sup>[12]</sup>中重要的一部分。现有工作从不同粒度对语义 Web 的链接结构进行了分析:一些研究关注细粒度的实体之间的链接结构,并根据实体的不同层次——模式层和实例层,分别进行研究以及综合分析,得出一些重要结论;另一些研究则从更粗的粒度对文档之间的链接结构进行分析,这部分研究主要关注于利用实体间的链接关系挖掘出更多的文档间的链接关系。

本文第 2 节介绍了与语义 Web 链接结构分析相关的基本概念,尤其是图相关的概念;第 3 节描述了语义 Web 上实体链接结构分析的现状;第 4 节描述了语义 Web 上文档链接结构分析的现状;最后是对语义 Web 链接结构分析的总结,并探讨未来可能的研究方向。

## 2 基本概念

边不带权的无向图  $G_u = (V_u, E_u)$  中两个节点  $i, j \in V_u$  之间的距离  $d_{ij} = |e_n|$ , 其中  $e_n$  是  $i$  到  $j$  的最短路径上的边的集合。 $G_u$  的直径  $d = \max_{i,j \in V_u} (d_{ij})$ 。如果  $G_u$  连通,  $G_u$  的特征路径长度  $l = \text{avg}_{i,j \in V_u} (d_{ij})$ 。度为  $k_i$  的节点  $i$  邻居间实际存在的边数  $n_i$  和最多可能边数的比值称为节点  $i$  的聚类系数  $c_i = 2n_i / (k_i(k_i - 1))$ , 而图的聚类系数  $c = \text{avg}_{i \in V_u} (c_i)$ 。一般将具有小的特征路径长度和大的聚类系数的图称为小世界网络。

若随机变量  $x$  的概率密度函数  $p(x) = Ae^{-\gamma x}$ , 则称  $x$  服从幂律分布, 其中  $A$  和  $\gamma$  都是正数,  $\gamma$  称为幂律指数。对于随机变量  $x$ ,  $p(X \geq x)$  称为互补累积分布函数(Complementary Cumulative Distribution Function, 简称为 CCDF)。而离散随机变量  $x$  中不同值的集合  $D$  按照  $D$  中元素降序排列, 将元素的排名与对应的值分别作为纵横坐标, 这样的分布称为 VR 分布, 即  $VR: [1, |D|] \rightarrow D$ 。图的节点度的分布情况可用概率密度函数  $p(k)$  来表示, 当度数符合幂律分布时, 就说该图具有无标度(scale-free)性质。

语义 Web 文档包含一系列三元组的集合, 每个三元组可以表示成  $\langle s, p, o \rangle$  的形式,  $s, p, o$  分别表示该三元组的主语(subject)、谓语(predicate)和宾语(object)。

词汇表(vocabulary)<sup>[13]</sup>是指拥有共同 URI 命名空间的类或属性的集合。术语  $t$  属于词汇表  $v$ , 当且仅当  $t$  的 URI 命名空间和标识  $v$  的 URI 命名空间相同, 并且  $v$  中三元组能推出  $t$  是类或者属性。

为了简单起见, 本文使用 qualified name<sup>[14]</sup> (QName) 的形式表示 URI, 比如用 foaf:Person 表示 <http://xmlns.com/foaf/0.1/Person>。本文使用的命名空间和它们对应的前缀如表 1 所列。

表 1 URI 命名空间和对应的前缀

前缀(Prefix)	URI 命名空间(namespace)
rdf	<a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a>
rdfs	<a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a>
owl	<a href="http://www.w3.org/2002/07/owl#">http://www.w3.org/2002/07/owl#</a>
daml	<a href="http://www.w3.org/2001/10/daml+oil#">http://www.w3.org/2001/10/daml+oil#</a>
foaf	<a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a>
skos	<a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a>

<sup>3)</sup>DAML 本体库, <http://www.daml.org/ontologies>

## 3 实体链接结构分析

在语义 Web 中, 实体之间的链接关系源于语义 Web 文档所含 RDF 三元组, 因此实体间的链接关系不仅有语义而且有出处(provenance), 这就带来实体间链接关系的语义表达问题以及可追溯问题, 从而对实体链接结构的分析, 尤其是模型的构建带来很大的复杂性。

实体是由类、属性和实例这 3 类不同的元素构成的, 类和属性又通常称为术语。这样语义 Web 上的实体链接结构可以分为如下 3 个部分: 术语之间、实例之间以及实体之间(结合术语和实例)。通常将术语之间的链接结构称为模式层(schema level)实体链接结构, 而实例间的链接结构称为实例层(instance level)实体链接结构, 实例和术语相互之间的链接结构称为混合实体链接结构。

### 3.1 模式层实体链接结构分析

语义 Web 中模式层的实体用来描述类、属性及其之间的关系, 通过定义不同系统间的公共概念, 便于这些系统间的共享和重用。模式层实体链接结构较为复杂, 不仅因为包含不同的语义和出处, 而且术语包含类和属性这两个成分, 这样类和类、属性和属性以及类和属性之间就存在着各种联系。这些不同的联系对应各种链接模型, 并且每个模型所能处理的问题也不相同, 因此链接模型是分析模式层实体链接结构关注的焦点。

#### 3.1.1 模式层 RDF 图

最基本的工作是对模式层 RDF 图的分析, Gil<sup>[15]</sup>收集了 282 个 DAML<sup>3)</sup> 本体文档, 并从中提取出 1365286 个三元组。每个三元组的主语(subject)和宾语(object)分别对应图的两个节点, 而主语和宾语之间的联系对应图的边, 这样生成的图包含了 307231 个点和 588890 条边。文献将该图看成无向图, 对它的度分布、特征路径长度和聚类系数进行计算。结果表明, 图的平均度数为 3.83, 度服从指数为 2.19 的幂律分布。图的特征路径长度和聚类系数分别为 5.07 和 0.092, 具有小世界性质。

#### 3.1.2 类包含图

模式层 RDF 图中的节点和边都有特定含义。当 RDF 图中相邻的两个节点都是类, 边为 rdfs:subClassOf 时, 这样生成的有向子图就表达了类之间的包含关系, 该图是有向图, 称为类包含图(Class Subsumption Graph, 或称为类层次结构)。具体地说, 类包含图  $G_s = (C, P_s)$ , 其中  $C$  是类的集合,  $P_s = \{(c_1, c_2) | c_1 \in C \cap c_2 \in C \cap \text{subClass}(c_1, c_2)\}$ ,  $\text{subClass}(c_1, c_2)$  表示  $c_1$  是  $c_2$  的子类。当满足  $\text{subClass}(c_1, c_2)$  时, 称在  $G_s$  中  $c_2$  位于  $c_1$  的上层。由于包含关系具有传递性, 多数工作都将通过传递关系导出的边增加到  $P_s$  中。也就是如果  $(a, b)$  和  $(b, c)$  是  $P_s$  中两条边, 则  $(a, c)$  也包含在  $P_s$  中。

Tempich 等<sup>[16]</sup>从 DAML 本体库中抽取分类本体(包含大量类, 但包含很少属性)中类之间的包含关系, 生成整体的类包含图。分析表明, DAML 库对应的类包含图中度的分布服从指数为 2.2 的幂律分布。

李增扬等<sup>[17]</sup>通过对 3 个不同领域的本体分析也发现它们各自的类包含图中度分布符合幂律分布。此外, 通过聚类

系数、平均路径长度的计算,他们还发现类包含图并不一定具有小世界性质。

### 3.1.3 类属性图

类之间除了 `rdfs:subClassOf` 形成的包含关系外,另一种常见的是由属性连接的定义域和值域之间的关系。一般称这类图为类属性图(Class Property Graph)。类属性图是一个有向图  $G_p = (C, P_p)$ , 其中  $C$  是类的集合,边  $P_p = \{(c_1, p, c_2) \mid c_1, c_2 \in C \wedge \text{domain}(p, c_1) \wedge \text{range}(p, c_2)\}$ ,  $\text{domain}(p, c_1)$  和  $\text{range}(p, c_2)$  分别表示  $p$  的 domain 和 range 是  $c_1$  和  $c_2$ 。目前都将类包含图  $G_c$  和类属性图  $G_p$  结合起来研究,下面介绍这方面的研究工作。

Theoharis 等<sup>[18]</sup>从 `RDFSuite`<sup>4)</sup>, `SchemaWeb`<sup>5)</sup> 和 `Swoogle`<sup>6)</sup> 本体库中收集并挑选了较大的 250 个本体(包含很多的类和属性)。由于每个本体经常使用其它本体中定义的实体,为了让考虑的本体具有相对的完整性,文献又将该本体中使用到的其它本体加入其中。

类包含图的生成方法与前一节介绍的类似,不过由于类包含图具有自反性,文献将类包含图的每个节点都增加一条自环。为了生成类属性图,文献不仅考虑显式声明的 `rdfs:domain` 和 `rdfs:range` 关系,还将其它能够导出或影响这样关系的构造子也通过某些处理包含在内,这些构造子包括 `owl:unionOf`, `owl:intersectionOf`, `owl:someValuesFrom`, `owl:allValuesFrom` 和 `owl:inverseOf`。

实验对每个本体生成一个类包含图和一个类属性图,然后计算每个图度的 CCDF 分布和 VR 分布,观察它们是否符合幂律分布(判断 ACC(absolute value of the correlation coefficient)是否大于 0.9)。通过观察类包含图可以发现,多数类包含图中入度的 CCDF 和 VR 分布都满足幂律分布,并且随着类的增加这种趋势更为明显。通过观察类属性图可以发现,很多类属性图中入度和出度的 VR 分布都符合幂律分布,而 CCDF 符合幂律分布的情况较少。随着类属性图中边(属性)的增加,度的 CCDF 和 VR 幂律分布现象更为明显。

文献进一步将类包含图从上层到下层平均地分为 A, B, C, D 4 个层次,通过统计类包含图中各层次类的数量来分析类包含图的形态。实验结果表明,类包含图中类数量的分布呈现出花瓶的形状。也就是,更多的类位于中层和底层之间(C层),而高层和底层之间(A和B层之间)包含的类最少。文献分析认为,中层和底层包含的类最多,说明类的层次结构并不平衡,也就是有些分支很深(位于D层)而更多的较浅(位于C层)。另外,文献对高层包含的类并不是最少也作出了解释,这是由于 XML datatype 位于最高层,并且很多本体引入的公共本体(比如 Dublin core 和 FOAF)中的类都位于类包含图的高层。

文献进一步考虑属性在类包含图中的层次分布,其中将属性的层次定义为它在类属性图中的两个端点(也就是类)在类包含图中的层次取中间值。文献发现,类包含图中属性的

分布与类的分布相反,呈现出倒置花瓶的形状。也就是属性更多位于类包含图中的中上层(A和B之间)而不是中下层(C层)。结合类包含图中类的分布,文献得出结论,类的层次结构更多地用于分类而不是通过进一步地增加属性来精细化类(即并没有定义新的属性来描述子类的实例之间关系)。

Huang 等<sup>[19]</sup>也研究了类包含图和类属性图,不过他们将两个图进行了结合。首先对类属性图给出了一个公式,用于评估类层次结构的平衡性。然后将类属性图加入到类包含图中,合成一张图,分析该图的连通性和中心概念,从而利用这些指标来评估本体的质量。

### 3.1.4 术语关系图

类包含图和类属性图对于了解本体的结构以及评估本体质量带来很大的帮助,但这类图更多地考虑类之间的关系,而忽视了属性的研究。有些工作试图兼顾类和属性,也就是术语。将术语作为图的节点,术语之间的各种关联作为边,生成的这类图称为术语关系图。

Cheng 等<sup>[20]</sup>分析了术语之间的依赖关系。文献将这种关系描述成有向的术语依赖图(Term Dependence Graph)  $G_d = (\langle C \cup P \rangle, D)$ , 其中节点为类或属性的集合,  $D = \{(t_1, t_2) \mid t_1 \in \text{Subj}(s) \wedge t_2 \in (\text{Pred}(s) \vee \text{Obj}(s))\}$ 。若  $t_1, t_2$  满足上面关系,则称  $t_1$  依赖于(dependence)  $t_2$ , 或  $t_2$  影响(influence)  $t_1$ 。  $D$  定义中用到的  $s$  表示  $t_1$  所在的 sentence<sup>7)</sup>, `Subj`, `Pred` 和 `Obj` 分别表示 sentence 的主语、谓语和宾语部分。

文献以 `Falcons`<sup>8)</sup> 截止 2008 年 4 月的数据为分析对象,生成的术语依赖图包含 1278233 个点、7312657 条边。术语依赖图的出入度分别代表术语的直接依赖和直接影响程度,也就是较大的出度表示术语定义中直接使用了其它术语,较大入度则表示该术语在很多其它的术语定义中被直接使用。分析表明,术语依赖图中入度服从指数为 1.82 的幂律分布。其中有 7 个术语的入度大于 10 万,它们都是语言层词汇表中的术语,并且从 `rdfs:label` 和 `rdfs:comment` 的大量使用可以看出很多的开发者都会在术语的定义中增加人们可读的信息。此外,类的包含关系(层次结构)也是本体中使用最广的结构(`rdfs:subClassOf`)。作者还发现有 64.6% 的术语并没有被其它的术语所依赖。术语依赖图的出度分布曲线头部不满足幂律分布,出度为 5 的节点最多,达到 40.9%。

为了分析术语总的依赖情况,文献引入了可达性分析。通过图的遍历,计算每个点的正向和反向可达点数,即依赖度和影响度。实验表明,每个术语平均要依赖 1105 个其它的术语,并且分析发现有一个由 13 个术语构成的强连通分支被几乎所有的术语所依赖。这些术语都在 RDF 或 RDFS 中,这就意味着它们的改变将完全地改变整个语义 Web 的含义。

每个术语除了考虑总的依赖个数外,还要分析它们之间的依赖深度,即为了得到一个术语的完整定义 BFS 算法需要处理的层数。实验表明,平均的依赖深度为 10.05, 51.4% 的依赖深度小于 6,然而却有 11.5% 的依赖深度大于 25。此

<sup>4)</sup> <http://athena.ics.forth.gr:9090/RDF/VRP/Examples/>

<sup>5)</sup> <http://www.schemaweb.info/>

<sup>6)</sup> <http://swoogle.umbc.edu/>

<sup>7)</sup> sentence 是为了处理空白节点。简单来说就是将由公共空白节点连接的三元组看成一个集合。具体参照文献[36]。

<sup>8)</sup> <http://iws.seu.edu.cn/services/falcons/>

外,从强连通分支的分布可以看出,93.4%的术语都是平凡的强连通分支(只包含一个节点),而最大的强连通分支也只有14883个术语,并且绝大多数的强连通分支中术语只来自一个词汇表。通过逐步删除图中最高度数的点,发现图的弱连通性急剧下降。此外,删除语言层词汇表(RDF, RDFS, OWL, DAML)中的术语,图的平均出入度由原来的5.72下降到1.92。

Hoser等<sup>[21]</sup>将类和属性都作为图的节点,边则涵盖了类与类(rdfs:subClassOf)、属性与属性(rdfs:subPropertyOf)、类与属性(rdfs:domain)以及属性和类(rdfs:range)之间的关系。文献分别应用出度、入度、介数(betweenness)、特征向量等方法寻找图的中心,从而分析本体中的核心内容和结构。

### 3.2 实例层实体链接结构分析

由于实例层只是描述实例之间的关系,对应的链接结构模型也只包含实例之间的链接,因此除了链接类型(属性)不同之外,其链接结构较为清晰。下面从不同应用领域介绍其研究现状。

#### 3.2.1 社会网络

社会网络的分析可用于研究群体的演化,标记在线社区,甚至发现潜在社区。随着语义Web的发展,出现了很多用RDF描述的社会网络数据源,从而为社会网络分析提供了素材;同时很多成熟的社会网络分析方法也应用到语义Web的分析中。在社会网络中,节点就是各种社会实体,但是社会实体之间的关系却是多种多样的。有些关系在数据中显式声明,比如FOAF<sup>9)</sup>的foaf:knows说明某个人认识的人,但更多关系则隐含在特定事件或活动之中,比如合作的文章、共同的研究兴趣等。下面主要从不同社会实体关系的角度简述语义Web实例分析在社会网络中的应用。

Ding等<sup>[22-24]</sup>提出一系列方法用于判别、发现、抽取和融合FOAF数据,并对处理后的数据进行分析。文献将来自于非blog站点的5000个文档作为分析目标,利用实例的反函数属性等方法将描述同一实体的不同URI归为一组。实验结果表明,这5000多个文档共包含50559个实例,其中56%的文档只包含一个实例,也就是没有他们认识的人的描述。而通过将不同的URI归为一组,得到的图包含42504个组、35299个foaf:knows关系。将该图看成无向图,度的分析显示,只有7%的组有出度和入度,并且97.7%的组只有一个入度,因此该图的连通性也不好。实际上,最大的连通分支只包含24559个组,而多数的连通分支包含的组都不超过5个。

Paolillo等<sup>[25]</sup>以LiveJournal<sup>10)</sup>站点中的FOAF文档为分析目标,对2004、2005年的两个数据集中foaf:knows和foaf:interest两个关系进行研究。他们将前500个最被关注的人和前500最被关注的兴趣分别作为两个矩阵的列,而矩阵的行则分别对应有边指向这500个人和这500个兴趣的用户。这样就得到2004、2005年中两种关系对应的4个矩阵,最后

将这4个矩阵合并成一个大矩阵(2000列)。他们对这个矩阵进行主成分分析(Principal Components Analysis)后发现,选择朋友的方式和拥有的兴趣之间并无多大联系。接着作者根据这些人的兴趣关系对矩阵进行层次聚类,得出5个不同的类别,并发现两年的数据聚类结果差别较大。

Mika<sup>[26,27]</sup>将在几个特定会议上发表论文的人作为研究对象,通过分析他们的主页、FOAF文档、公共邮件以及合著的文献,得到这些人之间的关系,接着使用不同方法识别出网络的中心,最后对他们的研究兴趣进行了分析。

Aleman-Meza等<sup>[28]</sup>关注利益冲突的检测(Conflict of Interest Detection),他们从FOAF和DBLP<sup>11)</sup>中收集了一些研究者的信息,分析其间的利害关系,得出的结果可以避免因论文分配不当而带来的有偏评价。

#### 3.2.2 生物医学

Zhang等<sup>[29]</sup>对传统中医药的实例数据进行研究。他们将药物、症状、诊断看成节点,他们之间的关系作为边。对这个图使用基于介数中心的测度方法,寻找结构上重要的点,从而判断病原。

Tari等<sup>[30]</sup>通过网络观察基因间交互的全局行为。该网络以基因作为节点,在参与同一个生化过程的两个基因间增加一条边。文献分别对人、水果、蠕虫和酵母的基因网络进行研究,发现这些网络都展现出小世界性质。但是除了蠕虫基因网络外,其它网络都没有无标度的性质。

#### 3.2.3 Linking Open Data

为了让语义Web朝着更为实用的方向发展,Web的创始人Tim Berners-Lee提出链接数据(linked data)的思想<sup>12)</sup>,在他的倡议下,Linking Open Data(简称LOD)项目发展起来。LOD的目标是通过在Web上发布RDF形式的数据集,并创建大量这些数据集间的链接,带动整个社区链接数据的发布,从而使得语义Web实现自我促进与发展<sup>[31]</sup>。到2009年4月,这些数据集共包含了45亿的RDF三元组,这些数据集之间有多达1800万的RDF链接,同时覆盖了书籍、电影、地理、音乐等众多领域。

Hausenblas等<sup>[32]</sup>研究了LOD中位于中心位置的DBPedia<sup>13)</sup>数据集,对可以解引用<sup>14)</sup>的属性使用情况进行研究,分别得到它们用于内部链接和用于外部链接的情况。最后对FOAF实体的属性使用情况进行分析。

### 3.3 混合实体链接结构分析

模式层的分析结果有利于研究实例层的实体关系。相反,利用实例层的分析结果也有利于模式层关系的研究,因此一些工作将模式层和实例层综合考虑。

较为基本的工作就是不加区分地分析术语和实例的链接结构。Ma等<sup>[33]</sup>从传统中医药本体(TCMLS)中选取了两个子本体。将这两个子本体分别表示成RDF图的形式,将RDF的节点看成图的节点。若两个节点在同一三元组中出

<sup>9)</sup> <http://www.foaf-project.org/>。FOAF本体包含12个类和51个属性,人们可以利用这些类和属性描述他们的个人信息以及和别人的联系。

<sup>10)</sup> <http://www.livejournal.com>

<sup>11)</sup> <http://dblp.uni-trier.de>

<sup>12)</sup> <http://www.w3.org/DesignIssues/LinkedData.html>

<sup>13)</sup> <http://dbpedia.org/>

<sup>14)</sup> dereference,是指通过HTTP GET操作得到URI的表示,这里指得到URI对应的RDF文档。

现的,则存在一条边,这样就生成一个无向无权图。分别对这两个图使用复杂网络分析技术来计算度的分布、平均最短路径长度和聚类系统,结果显示这两个图都符合小世界和无标度性质。Zhang<sup>[34]</sup>对7个生物学方面的本体分别进行分析得知都具有无标度性质。

不区分术语和实例使得分析结果失去某些实际含义。因此将类、属性和实例区分开来,并较为详细地阐述它们之间的联系,是更多工作的关注点。

Qu等<sup>[35]</sup>通过实例层的结构研究模式层的特征。简单来说就是考虑具有类型的对象(一般称为实例)对应的对象链接图(Object Link Graph,简称OLG),通过对象对应的类型,导出类关联图(Class Association Graph,简称CAG)。本文从Falcons搜索引擎截止2008年8月的数据集中抽取出30852370个带有类型的对象。这些对象共对应到56631个类,每个类平均含有598个实例,而rdf:Statement所含实例最多(5378695个)。文献将CAG看成无向图,通过对象之间的链接归纳出类之间的281141个关联。这样CAG包含了56631个点和281141条边。通过CAG度分析发现,CAG的平均度数为9.91。度数最大的3个类分别为foaf:Person,skos:Concept和foaf:Document,说明这些类的实例与很多其他类的实例间存在关联。连通性的分析发现,超过一半的类不与其他类关联,这说明很多类的实例都没有与其它类的实例存在链接关系。但是也发现了一个很大的连通分支(称为LCC),该连通分支包含27.17%的类,90.56%的类关联(边),并且这些类包含了96.16%的对象,所以接下来的工作都关注于LCC的研究。实验发现,LCC特征路径长度和聚类系数分别为3.8和0.46,具有小世界性质。最后,通过节点重要性图摘要方法对LCC进行可视化呈现,通过人工观察进一步加深对LCC的理解。

#### 4 文档链接结构分析

语义Web文档是语义Web信息的载体,分析语义Web文档之间的链接结构对了解语义Web宏观结构很有必要,尤其对于语义Web搜索引擎的文档获取、文档排序起着指导作用。但由于语义Web文档之间比较缺乏显式链接关系(owl:imports和rdfs:seeAlso等),而大量的链接来自语义Web文档所含RDF三元组实体的链接中。标识实体的这些URI,既可能标识术语和对象,也可能标识Web信息资源(例如Web页面)。通过dereference这些URI得到语义Web文档或者超文本Web页面,从而获取语义Web文档之间的隐式链接。所以目前的主要工作都关注于如何挖掘更多的隐式链接。

Swoogle<sup>[36]</sup>提出一个浏览模型,用于发现语义Web文档之间的隐式链接,并用于语义Web文档排序。该浏览模型包含3个基本元素:语义Web文档(SWD)、语义Web术语(SWT)、语义Web本体(SWO)。其中SWO是一种包含术语定义的特殊文档。文献通过计算SWT之间的连接关系,以及归纳SWT和SWD之间关联,给出SWD之间的关系,最后通过PageRank算法给出文档重要性排序。

OntoKhoj<sup>[37]</sup>利用语义Web本体链接结构分析评估本体重要性。OntoKhoj的排序算法中规定两个类C1,C2为引用(reference)关系,当且仅当两个类间存在{rdf:type,rdfs:subclass,daml:subclass,rdfs:domain,rdfs:range,rdf:seeAlso,

rdf:about}中的一个关系。引用关系是有向的且可以传递,关系的强度随着传递的步数的增加而削弱。两个本体的连接强度取决于本体内部互相引用的类的个数和引用的强度。通过这样的模型,OntoKhoj同样采用了类似于PageRank的算法来计算本体的重要性。

Jung<sup>[38]</sup>在利用实体链接计算文档链接的同时,又将文档链接的关系应用到实例的链接结构分析中。本文对一个由人编写本体的系统进行研究,从计算本体中实体之间的相似性出发,利用实体所属本体的关系,计算本体之间的距离,接着又通过本体的作者关系计算人与人之间的亲近性(affinity)。

**结束语** 随着语义Web的发展,语义Web的链接结构分析逐步得到重视。特别地,模式层的链接结构分析方面已经有较多研究。一些工作已经很好地揭示了模式层实体的链接结构特征,如文献[18,20],这些工作对于本体的构建和评估起到重要作用;一些工作较好地利用语义Web文档链接结构辅助文档的排序,如文献[36,37];文献[35]则独创性地利用实例层实体链接结构辅助分析模式层实体的链接结构。但总体来看,语义Web链接结构的分析研究尚处于初始阶段,体现在如下两个方面:

1)尚未对语义Web的大规模实例层实体进行全局的链接分析,没有工作能很好地结合模式层和实例层并给出它们之间的联系。

2)很少工作涉及结合文档和实体的研究,也没有形成公共模型来刻画它们之间的联系。

因此,语义Web链接结构分析仍有大量的工作有待开展,包括:

##### 1)语义Web的实体链接模型

在语义Web中,实体之间链接关系源自语义Web文档所含的RDF三元组,因此实体间的链接关系不仅有含义而且有出处,需要建立模型来综合这些因素。另外,如何处理模式层与实例层实体之间的联系,也是很难的问题。

##### 2)语义Web文档的链接模型

语义Web文档之间链接关系比较复杂,除了显式链接之外,大量的链接关系隐含在实体间的链接中,因此需要对这种隐式和复杂性进行建模并做相应的分析。

##### 3)语义Web复杂网络模型

语义Web复杂网络模型要综合语义Web文档的链接模型和语义Web的实体链接模型,而且要做必要的简化和抽象,以便获得语义Web的宏观结构。另外,该结构的演化模型也是一个值得关注的研究方向。

#### 参考文献

- [1] 王晓宇,周傲英.万维网的链接结构分析及其综述[J].软件学报,2003,14(10):1768-1780
- [2] Cardoso J. The semantic web vision where are we[J]. IEEE Intelligent Systems, 2007, 22(5): 1541-1672
- [3] Ding L, Finin T. Characterizing the semantic web on the web[C]// Proc. of the 5th International Semantic Web Conference (ISWC). LNCS 4273. 2006:242-257
- [4] Lee J, Goodwin R. The semantic webscape: a view of the semantic web[C]// Proc. of the 14th International Conference on World Wide Web (WWW). 2005:1154-1155 (poster)

(下转第45页)

网络中传输子网的选择。对于输入图为赋权图的情形,本文给出算法输出的支撑子图略微突破度数限制。如何设计总是可以输出满足度数限制的支撑子图的近似算法,是一个有趣的问题。

## 参 考 文 献

- [1] Kodialam M, Nandagopal T. Characterizing achievable rates in multi-hop wireless networks; the joint routing and scheduling problem[C]//MobiCom. 2003;42-54
- [2] Alicherry M, Bhatia R, Li L E. Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks[C]//MobiCom. 2005;58-72
- [3] Wang Weizhao, Wang Yu, Li Xiang-yang, et al. Efficient interference-aware TDMA link scheduling for static wireless networks[C]//MobiCom. 2006;262-273
- [4] Gao Jie, Guibas L J, Hershberger J, et al. Geometric spanners for routing in mobile networks[J]. IEEE Journal on Selected Areas in Communications, 2005, 23(1): 174-185
- [5] Cormen T, Leiserson C, Rivest R. Introduction to algorithms [M]. The MIT Press, 2002
- [6] Geomans M X. Minimum bounded degree spanning trees[C]//FOCS. 2006;273-282
- [7] Singh M, Lau L C. Approximating minimum bounded degree spanning trees to within one of optimal[C]//STOC. 2007;661-670
- [8] Gabow H N. An efficient reduction technique for degree-constrained subgraph and bidirected network flow problems[C]//STOC. 1983;448-456
- [9] Fürer M, Raghavachari B. Approximating the minimum degree spanning tree to within one from the optimal degree[C]//SO-DA. 1992;317-324
- (上接第 21 页)
- [5] 叶俊, 翟裕忠. 语义网数据分析之初探[J]. 东南大学学报: 自然科学版, 2008, 38 (Sup(1)): 301-307
- [6] D' Aquin M, Baldassarre C, Gridinoc L, et al. Characterizing Knowledge on the Semantic Web[C]//Workshop on Evaluation of Ontologies and Ontology-based Tools. 2007
- [7] Wang T D. Gauging ontologies and schemas by numbers[C]//4th Workshop on Evaluation of Ontologies for the Web. 2006
- [8] Wang T D, Parsia B, Hendler J. A survey of the web ontology landscape[C]//Proc. of the 5th International Semantic Web Conference (ISWC). LNCS 4273. 2006;682-694
- [9] Broder A, Kumar R, Maghoul F, et al. Graph structure in the web[J]. Computer Networks, 2000, 33(1-6): 309-320
- [10] Getoor L, Diehl C P. Link mining; a survey[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12
- [11] Chakrabarti S. Mining the web; discovering knowledge from hypertext data[M]. Morgan Kaufmann, 2002
- [12] Stumme G, Hotho A, Berendt B. Semantic web mining: state of the art and future directions[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2006, 4(2): 124-143
- [13] Berrueta D, Phipps J. Best practice recipes for publishing RDF vocabularies[S]. W3C Working Draft. 2008
- [14] Bray T, Hollander D, Layman A, et al. Namespaces in XML 1.0 [S]. Second edition. W3C Recommendation. 2006
- [15] Gil R. Measuring the semantic web[J]. AIS SIGSEMIS Bulletin, 2004, 1(2): 69-72
- [16] Tempich C, Volz R. Towards a benchmark for semantic web reasoners-an analysis of the DAML ontology library[C]//Proc. of the 2th International Semantic Web Conference (ISWC). 2003
- [17] 李增扬, 李兵, 何克清, 等. 本体中的复杂网络特性研究[J]. 微电子学与计算机, 2006, 23(9): 23-25
- [18] Theoharis Y, Tzitzikas Y, Kotzinos D, et al. On graph features of semantic web schemas[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 20(5): 692-702
- [19] Huang N, Diao S. Structure - based ontology evaluation [C] // Proc. of IEEE International Conference on e-Business Engineering (ICEBE). 2006; 132-137
- [20] Cheng G, Qu Y. Term dependence on the semantic web [C] // Proc. of the 7th International Semantic Web Conference (ISWC). LNCS 5318. 2008;665-680
- [21] Hoser B, Hotho A, Jaschke R, et al. Semantic network analysis of ontologies [C] // Proc. of the 3rd European Semantic Web Conference (ESWC). LNCS 4011. 2006;514-529
- [22] Ding L, Finin T, Joshi A. Analyzing social networks on the semantic web[J]. IEEE Intelligent System, 2004
- [23] Finin T, Ding L, Zhou L. Social networking on the semantic web [J]. The Learning Organization, 2005, 12(5): 418-435
- [24] Ding L, Zhou L, Finin T, et al. How the semantic web is being used; an analysis of FOAF documents[C]//Proc. of the 38th Hawaii International Conference on System Sciences. 2005
- [25] Paolillo J C, Mercure S, Wright E. The social semantic of live-journal FOAF; structure and change from 2004 to 2005[C]//Proc. of the 1st Workshop on Semantic Network Analysis at the ISWC 2005 Conference. 2005;69-80
- [26] Mika P. Flink; Semantic web technology for the extraction and analysis of social networks[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2005, 3(2/3): 211-223
- [27] Mika P. Social networks and the semantic web[C]//Proc. of International Conference on Web Intelligence. 2004;285-291
- [28] Aleman-Meza B, Nagarajan M, Ramakrishnan C, et al. Semantic analytics on social networks; experiences in addressing the problem of conflict of interest detection[C]//Proc. of the 15th International Conference on World Wide Web (WWW). 2006; 407-416
- [29] Zhang D, Gao L, Zhang H, et al. Centrality research on the traditional Chinese medicine network [C] // 2008 Workshop on Knowledge Discovery and Data Mining. 2008;59-62
- [30] Tari L, Baral C, Dasgupta P. Understanding the global properties of functionally-related gene networks using the gene ontology[C]//Pacific Symposium on Biocomputing. 2005;209-220
- [31] Bizer C, Heath T, Ayers D, et al. Interlinking open data on the web[C]//Proc. of 4th European Semantic Web Conference (ESWC). 2007;802-815 (poster)
- [32] Hausenblas M, Halb W, Raimond Y, et al. What is the size of the semantic web[C]//Proc. of the International Conference on Semantic Systems (I-Semantics2008). 2008
- [33] Ma J, Chen H. Complex network analysis on TCMLS sub-ontologies[C]//Proc. of the 3rd International Conference on Semantics, Knowledge and Grid. 2007;551-553
- [34] Zhang H. The scale-free nature of semantic web ontology[C]//Proc. of the 17th International Conference on World Wide Web (WWW). 2008;1047-1048 (poster)
- [35] Qu Y, Ge W, Gheng G, et al. Class association structure derived from linked objects[C]//Proceedings of the WebSci'09; Society On-Line. 2009
- [36] Ding L, Pan R, Finin T, et al. Finding and ranking knowledge on the semantic web[C]//Proc. of the 4th International Semantic Web Conference (ISWC). LNCS 3729. 2005;156-170
- [37] Patel C, Supekar K, Lee Y, et al. OntoKhoj; a semantic web portal for ontology searching, ranking and classification[C]//Proc. of the 5th ACM International Workshop on Web Information and Data Management. 2003;58-61
- [38] Jung J J, Euzenat J. Towards Semantic Social Networks[C]//Proc. of 4th European Semantic Web Conference (ESWC). LNCS 4519. 2007;267-280