ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.10, October 2009, pp.2834−2843
doi: 10.3724/SP.J.1001.2009.03505

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

语义网站点的发现与排序*

张 祥+, 葛唯益, 瞿裕忠

(东南大学 计算机科学与工程学院,江苏 南京 211189)

Finding and Ranking Semantic Web Sites

ZHANG Xiang+, GE Wei-Yi, QU Yu-Zhong

(School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

+ Corresponding author: E-mail: xzhang@seu.edu.cn, http://iws.seu.edu.cn/

Zhang X, Ge WY, Qu YZ. Finding and ranking semantic Web sites. *Journal of Software*, 2009,20(10): 2834−2843. http://www.jos.org.cn/1000-9825/3505.htm

Abstract: With the rapid growth of online RDF data, emerging semantic search engines facilitate user's searching of RDF (resource description framework) data. It is an open question to all semantic search engines how to find sites containing semantic Web information resources automatically and collect them efficiently. Firstly, the paper introduces a Linking Model of the Semantic Web Sites. The model characterizes the relations among Semantic Web Sites, Semantic Web Information Resources, RDF Models and Semantic Web Entities. This paper discusses the ownerships of Semantic Web Entities based on this model. It also defines a Site Dependency Graph in virtue of the model, and presents a set of ranking algorithms for Semantic Web Sites. Primary tests of these algorithms have been performed in a real-world semantic search engine. Experimental results show that this approach is effective in finding and ranking Semantic Web Sites.

Key words: semantic Web; resource description framework; search engine; link analysis; ranking algorithm

摘 要: 随着语义网中 RDF 数据的大量涌现,语义搜索引擎为用户搜索 RDF 数据带来了便利.但是,如何自动地发现包含语义网信息资源的站点,并高效地在语义网站点中收集语义网信息资源,一直是语义搜索引擎所面临的问题.首先介绍了语义网站点的链接模型.该模型刻画了语义网站点、语义网信息资源、RDF 模型和语义网实体之间的关系.基于该模型讨论了语义网实体的归属问题,并进一步定义了语义网站点的发现规则;另外,从站点链接模型出发,定义了语义网站点依赖图,并给出了对语义网站点进行排序的算法.将相关算法在一个真实的语义搜索引擎中进行了初步测试.实验结果表明,所提出的方法可以有效地发现语义网站点并对站点进行排序.

关键词: 语义网;资源描述框架;搜索引擎;链接分析;排序算法

中图法分类号: TP393 **文献标识码**: A

语义网与传统万维网的一大区别在于:万维网的宏观结构来源于网页之间的超链接;而语义网的宏观结构来源于用户通过使用 RDF 模型(http://www.w3.org/TR/rdf-concepts/)定义的数据之间的链接.随着 RDF 数据的

* Supported by the National Natural Science Foundation of China under Grant No.60773106 (国家自然科学基金); the National Basic Research Program of China under Grant No.2003CB317004 (国家重点基础研究发展计划(973))

Received 2008-03-18; Revised 2008-10-09; Accepted 2008-10-28; Published online 2009-06-09

大量涌现,用户常常感觉在语义网中无法像传统万维网那样容易找到想要的信息.为了解决这个问题,语义搜索引擎渐渐成为语义网的研究热点.Swoogle 是第一个在语义网研究人员中广泛使用的面向本体的搜索引擎[1].在 Swoogle 之后,SWSE[2],Sindice[3],Waston(http://watson.kmi.open.ac.uk/WatsonWUI/),Falcons[4]等语义搜索引擎不断涌现,纷纷提出了自己的语义搜索模型,为语义网带来了新的活力.

但是,语义搜索引擎面临一个最基本的问题:如何自动发现语义网中的 RDF 数据,或者说,如何自动发现包含 RDF 数据的语义网站点.一个语义搜索引擎的数据收集系统需要具有自动发现语义网站点的能力,以捕获最新的 RDF 数据.同时,如果搜索引擎能够掌握 RDF 数据在语义网中的分布状况,就可以避免数据收集的盲目性,极大地提高数据收集的效率.

除了语义网站点的自动发现以外,搜索引擎还需要对已发现的语义网站点进行排序.排序的目标是找到重要的语义网站点.这对于评估语义网资源收集的优先级起到了辅助作用,同时也为搜索引擎的排序结果提供了依据.与传统万维网一样,由于每个用户可以自由地在语义网中发布数据,所以语义网中的信任(trust)问题将决定语义网能否被最终用户接受,而语义网站点的排序也是迈向可信语义网的关键的一步.

本文的贡献在于:通过分析语义网信息资源间的链接提出了语义网的站点链接模型,并以此提出语义网站点的发现与排序算法.基于本文提出的算法,语义搜索引擎可以更加高效地收集语义网中的资源.

本文第 1 节讨论本文的研究背景.第 2 节定义语义网站点的链接模型.语义网站点的发现规则在第 3 节中加以介绍.第 4 节详述语义网站点的排序算法.第 5 节讨论与本文相关的工作.文章最后进行总结和展望.

1 研究背景

本文的工作是 Falcons 语义搜索引擎(http://iws.seu.edu.cn/services/falcons/)相关研究的一部分.Falcons 系统的原始数据集来源于 Google 和 Swoogle 及其他著名的 RDF 数据源.Falcons 系统在 Google 中检索以.rdf 和.owl 为后缀名的文档,在解析并确认了这些文档的确为 RDF 文档后,将其纳入 Falcons 的原始数据集.在 Swoogle 中,我们通过列举可能的关键字组合来发现 Swoogle 索引的 RDF 文档.此外,Falcons 系统还收集了例如 dbpedia.org 等站点中提供的 RDF 文档.尹导在文献[5]中详细介绍了 Falcons 原始数据集的来源.截至到 2008 年 3 月,Falcons 的原始数据集中已包括了 1 098 万个语义网信息资源.

Falcons 现有的数据收集方式的优势在于:通过已有的搜索引擎和著名的 RDF 数据源,能够在较短的时间内收集到大量的 RDF 文档,为 Falcons 系统的启动提供了强有力的数据支持.然而,这种数据收集方法的缺点也很明显:首先,该方法并不是一种全面的数据收集方法.对已有搜索引擎和已知数据源的依赖,将使得 Falcons 系统数据集的来源非常具有局限性;其次,该方法不能保证数据集的更新速度.随着语义网的快速发展,大量新开发的 RDF 文档不断涌现,Falcons 现有的数据收集方法无法快速地发现这些新出现的 RDF 文档.因此,为了达到可持续性的数据收集,Falcons 系统需要一套完整的数据收集机制,在语义网中高效而自动地收集 RDF 文档.

本文的目的就是研究如何提高 Falcons 系统数据收集的效率,而提高数据收集效率最重要的因素是提高数据收集的目的性.本文的工作将为 Falcons 系统提供明确的数据收集目标,我们以 Falcons 的原始数据集作为“种子”集合,通过对种子集合的分析,我们找出包含语义信息的文档,将其称为语义网站点.语义网站点的发现为数据收集提供了必要的线索.之后,我们借助链接分析方法对这些站点按重要性进行排序.站点的重要性将决定 Falcons 系统数据收集的优先级,从而作为制定收集策略的依据.

2 语义网站点的链接模型

在详细阐述语义网站点的发现机制之前,首先需要定义语义网站点的链接模型,以语义网站点为出发点,刻画语义网信息资源之间的链接形成的站点间的链接.传统的 Web 爬虫系统通过网页之间的链接发现新的 Web 站点.与此类似,对语义网数据收集系统而言,通过对已知的语义网站点间的链接分析发现新的语义网站点是最为直接而有效的方式.我们首先定义如下术语:

SWIR(semantic Web information resource):语义网信息资源.SWIR 是指在语义网站点内,包含 RDF 数据的

信息资源.目前,大量的 RDF 数据保存于 RDF/XML 文档或者嵌入 RDF 数据的 HTML 文档(基于 RDFa 规范(http://www.w3.org/TR/xhtml-rdfa-primer/));某些 Web Service 提供基于 RDF 的数据服务,我们也将其称为 SWIR.另外,SWIR 还包括了嵌入 RDF 数据的图像、视频等多媒体文件.每个 SWIR 使用其 URL 作为唯一标识.例如,Tim Berners-Lee 将其个人信息保存在 SWIR http://www.w3.org/People/Berners-Lee/card.rdf 中.一个 SWIR 可以通过超链接引用到另一个 SWIR.

SWS(semantic Web site):语义网站点.由于 W3C 并未约定语义网站点必须遵循的规范,所以简单来说,目前任何一个提供 SWIR 的站点均可看作是一个 SWS.每个 SWS 以其站点名(host name)作为唯一标识.例如,www.w3.org 就是一个重要的 SWS.

RDF Model:SWIR 所代表(represents)的 RDF 模型.例如 Tim Berners-Lee 在 www.w3.org 上的 FOAF 文档代表了一个描述其个人信息的 RDF 图,或者说,代表了一组 RDF 句子(RDF sentence)[6].W3C 曾经讨论过使用 Named Graph[7]作为语义网的数据模型,每个 RDF 图均由一个 URI 来标识.但是目前,Named Graph 并未在语义网中得到广泛使用.因此,我们依然使用 SWIR 的 URL 作为其 RDF Model 的标识.

SWE(semantic Web entity):语义网实体.一个 RDF Model 刻画了一组语义网中的对象和/或概念以及这些对象和/或概念之间的关联.我们将对象和概念统称为 SWE.每个 SWE 均由唯一的 URI 来标识.一个 RDF 句子可以通过 URI 引用(URI reference)来描述一个 SWE 的属性或多个 SWE 间的语义关联.根据 Web 体系结构(http://www.w3.org/TR/webarch/)的建议,每一个 URI 的拥有者(URI owner)应当为其 URI 提供相应的信息资源,所以我们假设每个 SWE 的拥有者是一个或多个 SWS,且这个(些)SWS 提供了定义该实体的 SWIR.在下一节,我们会重点讨论 SWE 的拥有者问题.

图 1 显示了语义网的站点链接模型.模型以 SWS 为出发点,以 SWS,SWIR,RDF Model 和 SWE 为基本元素,刻画了语义网 4 个基本元素之间可能存在的关系.一个 SWS 提供了一组 SWIR 的访问服务(serve);一个 SWIR 代表了(represent)一个相应的 RDF Model;一个 RDF Model 对一组 SWE 进行了建模(model);而每一个 SWE 的拥有者为一个或多个 SWS(owned_by).

我们将不同的 SWIR 之间可能存在的关联统称为 SWIR2SWIR;RDF Model 之间可能存在的关联统称为 model2model;SWE 之间可能存在的关联统称为 SWE2SWE.不同的 SWS 之间不存在直接的关联,而是通过如图 1 所示的各种路径间接地关联.

对于 SWIR 之间的关联而言,关联的类型可能有多种:例如一个 SWIR 可以在其注释中使用超链接指向另一个 SWIR(注意到 SWIR 中也可以包含 HTML 的语言结构);或者通过声明 rdfs:seeAlso 来表明与另一个 SWIR 之间的关联.同样地,RDF Model 之间和 SWE 之间也存在不同的关联类型.我们需要对这些基本元素之间的关联作进一步细化.

SWIR 之间与 RDF Model 之间可能存在的关联类型较为简单.其中,SWIR2SWIR 可以进一步划分为 4 种不同类型的关联:

1) r_1 hyperlinks_to r_2:r_1 通过超链接的方式指向 r_2(r_1 与 r_2 均为 SWIR,下同);
2) r_1 re-direct_to r_2:提供 r_1 的站点通过 HTTP 重定向的方式将对 r_1 的访问重定向到 r_2(r_1 与 r_2 具有不同的 URL);
3) r_1 rdfs:seeAlso r_2:r_1 中显式地声明了与 r_2 的 rdfs:seeAlso 关系,表明 r_2 提供了与 r_1 在语义上相关的内容;
4) r_1 rdfs:isDefinedBy r_2:r_1 中显式地声明了与 r_2 的 rdfs:isDefinedBy 关系,表明 r_1 中某个 SWE 的定义来源于 r_2.

以上 4 种不同类型的关联均为有向的关联.hyperlinks_to,rdfs:seeAlso 和 rdfs:isDefinedBy 这 3 种关联均可以通过对 SWIR 的分析得到;而 re-direct_to 较为特殊,需要通过访问 SWIR 对应的 SWS 得到.SWIR 的重定向问题往往涉及到 SWE 的归属问题,我们会在下文中对重定向问题作详细的讨论.

model2model 也可以进一步划分为 4 种不同类型的关联:

1) m_1 owl:imports m_2:m_1 认可并导入了 m_2 中的模型;
2) m_1 owl:priorVersion m_2:m_1 有一个较早的版本 m_2;
3) m_1 owl:backwardCompatibleWith m_2:m_1 向后兼容 m_2 中的模型;
4) m_1 owl:incompatible m_2:m_1 不兼容 m_2 中的模型.

SWE 之间的关联类型比起 SWIR 与 RDF Model 之间的关联类型要丰富得多.两个 SWE 之间可以通过一个 owl:ObjectProperty 进行关联.我们将 SWE2SWE 细化为两种类型:language-level relation(语言层关联)及 user-defined relation(用户自定义的关联).当 SWE 之间使用语言层内建的 ObjectProperty 进行关联时,该关联为语言层关联.一般来说,两个表示概念的 SWE 之间的关联为语言层关联,例如(foaf:Person rdfs:subClassOf foaf:Agent)等等;当 SWE 之间使用用户自定义的 ObjectProperty 进行关联时,该关联为用户自定义的关联.一般来说,两个表示实例的 SWE 之间的关联为用户自定义关联,例如(“Chris Bizer”(http://www.bizer.de#chris) foaf:knows “Tim Berners-Lee”(http://www.w3.org/People/Berners-Lee/card#i)).

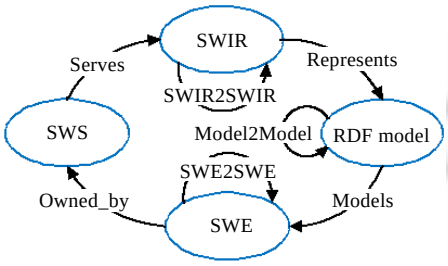


Fig.1 Linking model of the semantic Web sites
图 1 语义网站点的链接模型

3 语义网站点的发现

从 Falcons 系统已收集到的 SWIR 中我们可以得到当前语义网站点的链接模型.与传统的 Web 站点的发现类似,发现语义网站点的基本思路是从语义网站点链接模型出发,通过已知的 SWIR,RDF Model,SWE 和它们之间的链接找到可能的新的站点,并通过爬虫软件验证此类站点是否的确提供 SWIR.我们将 SWS 发现算法归结为一系列规则.

在详细解释规则之前,首先需要解释 SWE 的归属问题,即如何判断一个 SWE 是否属于某个 SWS.对 SWE 的建模是否权威等等.只有明确了这些问题,才能明确规则的含义,并且在发现的过程中选择适用的规则.

3.1 语义网实体的归属

由于 URI 使用的随意性,任何机构与个人都可以在自己的 SWIR 中对任意的 SWE 进行建模.如何定义 SWE 的归属,并找到 SWE 的权威定义成为语义网面临的一大问题.

例如,FOAF 是语义网中被广泛使用的用于描述个人信息和朋友关系的词汇表.该词汇表包含了 foaf:Person,foaf:knows 等 SWE.目前公认的这些 SWE 的权威定义来源于 http://xmlns.com/foaf/0.1/index.rdf,该 SWIR 属于 xmlns.com 这个语义网站点.

一般来说,SWE 的标识采用 HTTP scheme 的 URI(http://www.w3.org/TR/swbp-vocab-pub/),按照 Web 体系结构的建议,每一个 SWE 都应当是“可访问的”:在浏览器中输入一个 SWE 的 URI,应当得到一个表示该实体相应模型的 SWIR.大多数情况下,这个资源属于 URI 中的主机名部分所对应的 SWS.FOAF 词汇表就属于这种情况,其中的每个 SWE 都可以在浏览器中访问,且返回的资源来源于标识中指定的站点.我们可以简单地将任何一个 SWE 的 URI 中的站点视作该实体的拥有者.

另一种情况更为复杂:一个 SWS 虽然声明了一个 SWE,但是由于某些原因,无法将建模该实体的资源存放在自己的站点中.这种情况在语义网中并不鲜见.例如,某些站点受限于存储和流量限制,将属于自己的语义网

信息资源交给其他站点“托管”.这种情况往往会使用到 HTTP 303 重定向功能,当用户在浏览器中输入一个 URI 时,该 URI 的拥有者会自动地将请求重定向到另一个站点,后者向用户返回相关的 SWIR.一个 URI 的访问请求可能会在 Web 中进行多次重定向,用户通过浏览器输入一个 SWE 的 URI 后,该请求进行了 n–1 次重定向,最终由第 n 个站点返回了相关资源.重定向的使用增加了 SWE 归属问题的复杂性,也为解决语义网信任问题增加了难度.由于重定向反映了站点之间的某种“信任”关系,所以我们将 SWE 的拥有者定义为其重定向路径中的所有站点构成的集合.例如,Music 本体对应的 SWIR(http://purl.org/ontology/mo/)中定义了实体 music:MusicArtist (http://purl.org/ontology/mo/MusicArtist),在浏览器中访问该实体的 URI 时,将会重定向到另一个由站点 musicontology.com 提供的 SWIR(http://musicontology.com/musicontology.rdfs).所以,该实体的拥有者包括且仅包括 purl.org 以及 musicontology.com.

另外,从 SWIR 的角度来看,每一个 SWIR 背后的 RDF Model 均对一组 SWE 进行了建模,但却可能存在这样的情况:一个 RDF Model 建模的 SWE 并非全部归属于该模型背后的 SWS,此时,我们称这种建模是非权威的;反之,称为权威的建模.例如上述的 Music 本体中对 foaf:Person 进行了建模,该实体归属于 xmlns.com 这个站点.所以,这种建模是非权威的.非权威的建模可以简称为 SWE 的“重定义”.假如一个 SWE 的重定义与权威定义不同,就会对语义网起到不良的影响,因为这会给语义网带来不一致性.

3.2 语义网站点的发现规则

传统的万维网中,网页通过超链接形成简单的、链接类型单一的网络.传统的网页收集系统通过分析 HTML 网页,解析出其中的超链接来发现新的网页.寻找新网页的过程是一个迭代的过程,每一次迭代将新发现的网页纳入到下一轮网页分析的范畴.网页收集系统并不关心 HTML 语法中与超链接无关的语法结构,而 HTML 语法中描述超链接的标签是固定的,所以从网页中解析出超链接的规则是简单的.

与网页不同,语义网信息资源的结构更为复杂.如前文所述,一个语义网信息资源往往描述了一组语义网实体间构成的复杂网络.这种复杂网络涉及大量本体语言层的语法结构.从一个语义网信息资源中解析出语义网实体,进而发现语义网站点的过程将更为复杂.给定一组“种子”语义网信息资源,我们在发现语义网站点的过程中,无法通过与网页解析类似的方法来处理语义网信息资源.但是注意到,我们可以将复杂的分析过程归纳为一些简单的启发式规则,这些规则告诉我们,在语义网信息资源的分析过程中,在什么样的条件下,某个新发现的站点可能包含语义网信息资源.

我们通过以下 5 条规则来发现新的语义网站点:

规则 1. $serves(s_1,r_1)$ & $SWIR(r_1)\rightarrow SWS(s_1)$.

该规则是发现 SWS 的最基本规则.每一个 SWIR 所属的站点都是被确认的 SWS.这里,$serves(s_1,r_1)$ & $SWIR(r_1)$表示站点 s_1 提供了资源 r_1,并且 r_1 被确认为一个 SWIR;$SWS(s_1)$表示站点 s_1 被确认为一个 SWS.

规则 2. $represents(r_1,m_1)$ & $model2model(m_1,m_2)\rightarrow queue(host(m_2))$.

规则 2 中 $represents(r_1,m_1)$的含义是 *SWIR* r_1 表示了 RDF Model m_1;$model2model(m_1,m_2)$表示 m_1 中声明了与另一个 RDF Model m_2 存在关联;$queue(host(m_2))$表示系统猜测 RDF Model m_2 所对应的站点很有可能是一个 SWS,并将其加入到数据收集的队列中.也就是说,通过这条规则,我们为数据收集系统提供了一条线索:假如我们分析到的一个 RDF Model 导入了另一个 Model,并且后者从标识上来看属于另一个站点,那么这个站点很有可能是一个新的 SWS,需要通过爬虫软件去分析站点内容,然后找出其中的 SWIR.

规则 3. $models(m_1,e_1)$ & $SWE2SWE(e_1,e_2)$ & $owned\_by(e_2,s_2)\rightarrow queue(s_2)$.

该规则中 $models(m_1,e_1)$表示 RDF Model m_1 对 SWE e_1 进行了建模.通过对模型的分析,我们得到了与 e_1 有关联的 e_2.无论 e_1 与 e_2 之间的关联是何种类型,假如 e_2 从标识上来看属于另一个站点 s_2,我们就将 s_2 也纳入到站点收集的队列中.

规则 4. $models(m_1,e_2)$ & $owned\_by(e_2,s_2)\rightarrow queue(s_2)$.

这条规则是通过对 RDF Model 的分析找出模型所建模的每一个 SWE 所属的站点,从而发现新的 SWS.但是如果一个 RDF Model 对该模型所属的 SWS 之外的实体进行了建模,按照上一节的论述,这种建模是非权威

的.在这种情况下,得到的站点被确认为 SWS 的概率将小于以上 3 种规则.

规则 5. *serves*(s_1,r_1) & *SWIR2SWIR*(r_1,r_2) &→*queue*(r_2).

*SWIR*2*SWIR*(r_1,r_2)表示 r_1 中声明了与 r_2 之间存在的关联,本规则考虑了通过 SWIR 之间的关联来发现新的 SWS.在实际使用中我们发现,通过这条规则发现 SWS 的效率较低,因为一个 SWIR 中显式声明的与其相关的资源并不需要一定也是 SWIR,它可能是网页或者图片,甚至也可能是无法访问的链接.

3.3 分析结果

我们截取了 Falcons 数据收集系统截止至 2007 年 12 月的数据镜像,分析了超过 2 000 000 个种子 SWIR.从这些种子中,Falcons 发现了 271 178 个可能包含 SWIR 的站点.我们正在逐步验证这些站点是否确为 SWS,同时在这些站点中收集新的 SWIR.从 SWS 的发现、验证到 SWIR 的收集、分析,这是一个长期并且循环的过程.目前,我们对这 271 178 个站点在地理位置上的分布状况作了初步的分析,并将该分布与传统万维网中已注册域名的地理位置分布进行了比较.

271 178 个站点中有 3%左右的站点无法在 DNS 中得到对应的 IP 地址.通过 hostip.info 站点提供的 IP 地址与物理位置的映射数据库,我们得到了剩余的 263 051 个站点所属的国家或地区.如表 1 所示,我们发现大部分可能的语义网站点来源于美国,占据了所有站点的 55%左右;英国和加拿大分列第 2 位、第 3 位,共有 15%的站点来源于这两个国家.分析结果中约 85%的站点来源于表 1 所包含的排名前 10 的国家或地区,仅有 15%来源于其他国家或地区.我国在该排名中仅位列第 19 位,占据所有站点的 0.5%左右.这也说明,我国虽然在语义网的研究上取得了一定的成就,但在语义网的参与程度和语义网信息资源的丰富程度上还不够.

为了比较传统万维网域名与语义网站点的地理位置分布,我们从 webhosting.info 站点提供的数据中得到了传统万维网域名在地理位置上的分布概况(http://www.webhosting.info/domains/country_stats/).截止至 2008 年 10 月 27 日,万维网中注册的域名总数约为 102 781 301 个左右.如表 1 所示,与可能包含 SWIR 的站点的地理位置分布相似,万维网中绝大多数已注册域名集中在少数几个国家或地区上,且大部分语义网参与度较高的国家在传统的万维网中已相当成熟.

虽然以上的分析从很大程度上依赖于第三方数据,从而无法保证绝对的精确性,但是我们依然能够得出结论:通过前文所述的发现规则得到的 263 051 个站点在地理位置上基本分布均匀,合乎我们对传统万维网的认知.然而,仅仅发现这些站点并对地理位置进行分析是远远不够的,更重要的是如何确认这些站点是否为 SWS 并在站点中高效地挖掘 SWIR.对 SWS 的排序将对 SWIR 的挖掘起到重要作用.

Table 1 Statistics of location distribution of semantic Web sites and country-wise rankings of total domains

表 1 语义网站点地理位置分布及各国域名总数排名统计

No.	Country or region	Number of SWS	Ratio (%)	Number of domains	Ranking by number of domains
1	United States	144 537	54.9	66 080 363	1
2	United Kingdom	24 768	9.4	3 670 367	3
3	Canada	14 627	5.6	3 230 991	4
4	Germany	13 677	5.2	5 797 469	2
5	Australia	5 742	2.2	2 240 395	7
6	France	5 480	2.1	2 380 595	6
7	Japan	5 100	1.9	1 353 451	9
8	Italy	3 996	1.5	1 007 390	11
9	Netherlands	3 111	1.2	897 787	12
10	Spain	1 913	0.7	1 186 629	10

4 语义网站点的排序

优秀的语义网搜索引擎应当能够及时地反映语义网的变化,所以高效地收集 SWIR 一直都是 Falcons 的目标之一.这要求 Falcons 不仅知道语义网中哪些站点包含 SWIR,同时还需要掌握这些站点的重要性,对站点进行排序.数据收集系统在访问这些站点时需要按照站点的重要性安排访问的频率和优先级.

本文所述的重要性是指语义网站点在语义网中的流行度.重要的语义网站点往往定义了一些被语义网广

泛接受并流行起来的语义网实体.我们借助传统的链接分析方法来分析这些站点的重要性.通过分析语义网站点的链接模型,我们可以得到站点间的依赖度.简单来说,假若某个语义网站点被众多其他站点所依赖,就意味着该站点在语义网中具有重要的地位.

在传统的万维网中,网页的重要性就引起了研究者的重视.传统的网页链接分析的方法已经非常成熟,例如,通过 PageRank[8]和 HITS(hypertext-induced topic search)算法[9]计算网页的重要性均得到了较为理想的效果.其中,PageRank 作为 Google 搜索引擎的网页排序算法,已经得到了用户的认可.HITS 算法不仅可以通过计算网页在图中的 authority 来评估网页的重要性,还可以通过计算网页的 hubness 来评估一个网页链接到其他重要网页的程度.

我们首先定义了语义网站点依赖图,接着分别使用基于 PageRank 和 HITS 的算法对该图进行分析.

4.1 语义网站点依赖图

为了便于分析语义网站点的重要性,首先需要量化语义网站点之间链接的权重.我们称其为站点间的依赖度,用 $dep(s_i,s_j)$表示.语义网站点依赖图 $G=(V,E,W)$是带权有向图,其中,V 为节点集合,每一个节点表示一个已知的语义网站点;E 为边集,对于任意的 $s_i,s_j\in V,(s_i,s_j)\in E$ 当且仅当语义网站点 s_i 对 s_j 存在依赖性.W 是边的权重的集合,每一条边的权重按照以下公式计算:

$$dep(s_1,s_2)=\frac{\sum_{\{e|owned\_by(e,s_1)\}} dep(e,s_2)}{|\{e\,|\,owned\_by(e,s_1)\}|} \tag{1}$$

$$dep(e,s_2)=\frac{\sum_{\{e'|owned\_by(e',s_2)\}} dep(e,e')}{|\{e'\,|\,owned\_by(e',s_2)\}|} \tag{2}$$

$$dep(e,e')=\begin{cases}1, & SWE2SWE(e,e')\\ 0, & \text{else}\end{cases} \tag{3}$$

公式的直观含义是:一个语义网站点 s_1 对另一个语义网站点 s_2 的依赖度,是归属于 s_1 的每一个语义网实体对归属于 s_2 的每一个语义网实体的依赖度的平均值.当一个站点所拥有的大部分语义网实体都通过某种类型的关联与另一个站点所拥有的语义网实体发生联系时,前者就对后者有较高的依赖度.

语义网站点间的依赖度刻画了语义网站点间通过语义网实体形成的间接关联的强弱,体现了语义网站点相互之间相对的重要性.对语义网站点间依赖度的定义是语义网站点链接分析的基础.我们可以通过依赖度的定义将语义网站点构成的复杂网络转化为一个关联类型单一的简单带权网络,从而对语义网站点在整个语义网中的重要性进行定量的分析.

我们在实际建立语义网站点依赖图时对重定义进行了过滤.也就是说,当分析每一个收集到的种子 SWIR 时,该资源中非权威的部分将不会对站点间的依赖度产生影响.

4.2 语义网站点依赖图的分析

我们在语义网站点依赖图的分析中分别采用了带权 PageRank 算法与 HITS 算法.由于语义网站点依赖图是带权图,所以我们对原始的 PageRank 和 HITS 算法作了相应的变形,分别称为 Weighted PageRank 和 Weighted HITS.对于语义网站点依赖图中的每一个节点 s_i,我们对其计算 3 个指标,分别是其 PageRank 值 $PR(s_i)$、HITS Authority 值 $authority(s\neg_i)$以及 HITS Hubness 值 $hubness(s_i)$.为了方便陈述,我们使用迭代的方式计算以上指标.迭代计算方法与基于特征向量的计算方法是等效的.

我们使用 Weighted PageRank 对语义网站点依赖图中的每个节点按照公式(4)计算出 PR 值.语义网站点依赖图中,语义网站点的 PR 值越高,说明语义网中其他站点对该站点的依赖度越高,该站点在语义网中的重要性也越高.这种依赖度可以是直接的,也可以是间接的;公式(5)和公式(6)为 Weighted HITS 计算公式.对于图中的每个节点,Weighted HITS 公式不仅计算了节点的 authority 值,用于评价语义网站点在语义网中的重要性,还计算了 hubness 值这个指标.该指标体现了语义网站点在语义网中对其他站点的依赖总体依赖度.语义网站点的

authority 值与 hubness 值是一对互增量(reinforcement).

$$PR(s_i)=\frac{1-d}{|V|}+d\cdot\sum_{(s_j,s_i)\in E}\frac{dep(s_j,s_i)\cdot PR(s_j)}{\sum_{(s_j,s_k)\in E}dep(s_j,s_k)} \tag{4}$$

$$authority(s_i)=\sum_{(s_j,s_i)\in E}dep(s_j,s_i)\cdot hubness(s_j) \tag{5}$$

$$hubness(s_i)=\sum_{(s_i,s_j)\in E}dep(s_i,s_j)authority(s_j) \tag{6}$$

公式(4)中的 d 是原始 PageRank 算法中的阻尼因子(damping factor).阻尼因子的作用是保证迭代公式能够用合理的随机模型进行解释,从而能够在有限次迭代后收敛.通常,阻尼因子被设为 0.8~0.9 之间的值,我们将其设为 0.85.我们反复地对语义网站点依赖图的每一个节点计算以上 3 项指标,直到计算结果达到一定的收敛效果为止.每次迭代后均需对每个节点的各项指标进行正规化:按照 PageRank 算法的要求,所有节点的 PR 值之和需正规化为 1;按照 HITS 算法的要求,所有节点的 authority 值的平方和需要正规化为 1,hubness 值类似处理.各项指标正规化的作用同样是为了保证迭代公式能够用合理的随机模型进行解释,从而收敛.

4.3 分析结果

我们分别使用第 4.2 节中所述的 Weighted PageRank 和 Weighted HITS 算法对 Falcons 已发现的 271 178 个站点进行了排序.表 2 中包含了使用不同指标得到的前 10 个结果.

Table 2 Top 10 sites ranked by different metrics

表 2 以不同的指标得出的排名前 10 的站点

No.	PageRank	HITS-Authority	HITS-Hubness
1	www.w3.org	www.w3.org	dbpedia.org
2	purl.org	purl.org	bio2rdf.org
3	xmlns.com	xmlns.com	chatlogs.planetrdf.com
4	creativecommons.org	www.typepad.com	data.semanticweb.org
5	www.daml.org	www.movabletype.org	ajft.org
6	esw.w3.org	www.ldodds.com	blogiem.lv
7	www.ics.forth.gr	en.wikipedia.org	danbri.org
8	www-sop.inria.fr	del.icio.us	blog.drry.jp
9	simile.mit.edu	creativecommons.org	crschmidt.net
10	dublincore.org	web.resource.org	dblp.l3s.de

从排序的结果来看,使用 Weighted PageRank 排序的结果相对来说最符合人们对这些站点重要性的判断:排在第 1 位的 www.w3.org 定义了语义网中最重要的元本体(meta-ontology),例如 RDF 和 OWL 等,绝大多数站点中的语义网信息资源均需要使用到这些本体中的词汇;排在第 2 位、第 3 位的 purl.org 和 xmlns.com 中定义了相当多的常用本体,前者包括 Dublin Core 和 RSS 1.0;后者包括了 FOAF 和 WOT.根据 Swoogle 的分析结果[10],除 www.w3.org 所定义的本体外,这 4 个本体在语义网中使用得最为广泛.

虽然 HITS-Authority 得到的前 3 个结果与 PageRank 相同,但是从第 4 名开始却包含了很多意料之外的结果,其中有相当数量的 Blog 和 Wiki 站点.我们对这些站点进行了分析发现,这些站点都被 dbpedia.org 所依赖.dbpedia.org 是一个数据密集型站点,它从 wikipedia.org 中抽取结构化信息并提供结构化查询,同时它也将这些结构化信息以 RDF/XML 的格式进行封装并提供下载.作为目前 Falcons 系统最大的数据源,dbpedia.org 拥有最高的 HITS-Hubness,而该站点所依赖的站点也就相应地拥有了极高的 HITS-Authority.这被称为 HITS 算法中的 TKC 现象(tightly knit community effect)[11].简单来说,如果图中的少部分节点形成了一个内聚度很高的社区,则 HITS 算法会导致这个社区中的节点在排名中占据极大的优势,而社区外的节点无法与其抗衡.PageRank 算法相对来说较好地避免了 TKC 现象.

在 HITS-Hubness 中排名较高的站点往往是一些提供海量 SWIR 的站点.这些站点也许在语义网中并不流行,但是由于资源数量众多,Falcons 数据收集系统仍可以通过这些站点来发现其他 SWS.从排序结果来看,与 HITS-Authority 类似,除了 dbpedia.org 之外,大多数排名较高的站点是一些 Blog 和 Wiki 站点,这些站点中包含

了大量的 RSS 数据.而 RSS 数据并不是 Falcons 系统关注的重点.这也是 Falcons 数据收集系统下一步需要考虑的问题.

尽管 HITS 算法的排序结果不如 PageRank 算法,但是根据我们的统计,使用 PageRank 和 HITS-Authority 分别得到的前 1 000 个站点中有 85%是相同的(不考虑排名的先后).也就是说,HITS-Authority 的排序结果也具有一定的参考价值.我们综合考虑了以上 3 种指标的排序,并以此来设计并调整 Falcons 的数据收集系统的收集策略.由于收集策略超出了本文的讨论范围,所以这里不再赘述.

5 相关工作

到目前为止,还未有相关文献详细讨论语义网站点的概念及对其进行大规模分析.与本文相关的工作主要分为两个方面:RDF 文档的收集问题和语义网中的排序问题.SHOE-Expose Crawler(http://www.cs.umd.edu/projects/plus/SHOE/Expose.html)是最早的面向半结构化文档的爬虫系统,它利用超链接在万维网中搜索包含 SHOE 标记(SHOE mark-up)的文档,并从中抽取出基于 SHOE 规范的元数据.DAML Crawler (http://www.daml.org/crawler/)是一个面向 DAML 文档的爬虫系统.给定一个站点,DAML Crawler 仅仅通过超链接遍历 Web 站点并找出其中的 DAML 文档.RDF Crawler(http://ontobroker.semanticweb.org/rdfcrawl/)是一个面向 RDF 文档的爬虫程序.它从已知的 RDF 文档中通过 rdfs:seeAlso 和 rdfs:isDefinedBy 抽取出 URL,并试图访问这些 URL 来发现新的 RDF 文档.但是,RDF Crawler 没有从站点的角度去规划 RDF 文档的收集过程.Harth 等人在文献[12]中提出了一套完整的分析方法,用于从已有的 RDF 文档中抽取出 URI,以此为线索找到新的 RDF 文档并对其进行索引.Swoogle 作为最早被研究人员接受的本体搜索引擎,不仅通过 RDF 文档的分析找出新的 RDF 文档,也使用到类似于语义网站点的概念.Swoogle 的用户可以向 Swoogle 提交他们认为包含 RDF 文档的站点,Swoogle 会自动地在这些站点中进行搜寻.但是,Swoogle 并未考虑通过 RDF 文档的分析进行语义网站点的自动发现,更未考虑对语义网站点进行排序.

在万维网中,面向网页的排序问题是搜索引擎最关心的问题,并诞生了一系列成熟的链接分析方法.传统的 PageRank 和 HITS 算法属于基于特征向量的分析方法,Diligenti 在文献[13]中提出一个随机模型用于解释和模拟这些方法和它们的变种.在语义网环境下,排序问题的对象发生了很大的变化.Swoogle[10]和 AkTiveRank[14]对本体的排序作了研究,而文献[6]探讨了本体中的 RDF 句子的排序方法,并以此方法对本体进行摘要.实际上,语义网中的排序问题大多借用了传统的链接分析方法.文献[15]对这些排序问题进行了综述.但到目前为止,并未有文献讨论语义网站点的排序问题.

6 总结与展望

本文详尽地讨论了语义网站点的相关概念,借助传统的链接分析方法提出了语义网站点的发现与排序算法,并在实际的语义搜索引擎 Falcons 中初步测试了这些算法.对于语义搜索引擎的数据收集系统而言,这些研究问题起到了重要的辅助作用.

我们目前通过语义网中各元素之间的关联来发现新的语义网站点.在未来的工作中,我们将首先考虑关联的度量问题.对关联强弱的量化分析将有助于判断语义网站点的权威性和可靠性.我们还将研究语义网站点内部的结构信息和在语义网站点内部高效收集语义网信息资源的相关算法.此外,我们将重点研究如何利用已有的分析结果来调整收集策略.对数据收集结果的评估也是我们正在考虑的问题.

致谢 感谢吴鸿汉博士及尹导同学在算法实现和测试中给予我们的帮助.

References:

[1] Ding L, Finin TW, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi V, Sachs J. Swoogle: A semantic web search and metadata engine. In: Grossman D, ed. Proc. of the 13th ACM Conf. on Information and Knowledge Management. Washington: ACM, 2004. 652−659.

[2] Harth A, Decker S. Optimized index structures for querying RDF from the Web. In: Proc. of the 3rd Latin American Web Congress. Buenos Aires: IEEE Computer Society, 2005. 71−80.

[3] Tummarello G, Delbru R, Oren E. Sindice.com: Weaving the open linked data. In: Aberer K, ed. Proc. of the 6th Int'l Semantic Web Conf. Busan: Springer-Verlag, 2007. 552−565.

[4] Cheng G, Ge WY, Qu YZ. Falcons: Searching and browsing entities on the semantic Web. In: Huai JP, ed. Proc. of the 17th Int'l Conf. on World Wide Web. Beijing: ACM, 2008. 1101−1102.

[5] Dao Y. Design and implementation of a collection system of ontologies [MS. Thesis]. Changsha: Southeast University, 2008 (in Chinese with English abstract).

[6] Zhang X, Cheng G, Qu YZ. Ontology summarization based on RDF sentence graph. In: Williamson CL, ed. Proc. of the 16th Int'l Conf. on World Wide Web. Banff: ACM, 2007. 707−716.

[7] Carroll JJ, Bizer C, Hayes P, Stickler P. Named graphs, provenance and trust. In: Ellis A, ed. Proc. of the 14th Int'l Conf. on World Wide Web. Chiba: ACM, 2005. 613−622.

[8] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford University, 1998. http://ilpubs.stanford.edu:8090/422/

[9] Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,45(5):604−632.

[10] Ding L, Pan R, Finin TW, Joshi A, Peng Y, Kolari P. Finding and ranking knowledge on the semantic Web. In: Gil Y, Motta E, Benjamins VR, Musen MA, eds. Proc. of the 4th Int'l Semantic Web Conf. LNCS 3729, Springer-Verlag, 2005. 156−170.

[11] Lempel R, Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. The Int'l Journal of Computer and Telecommunications Networking (Computer Networks), 2000,33(1-6):387−401.

[12] Harth A, Umbrich J, Decker S. MultiCrawler: A pipelined architecture for crawling and indexing semantic Web data. In: Cruz IF, Decker S, Allemang D, Preist C, Schwabe D, Mika P, Uschold M, Aroyo L, eds. Proc. of the 5th Int'l Semantic Web Conf. LNCS 4273, Springer-Verlag, 2006. 258−271.

[13] Diligenti M, Gori M, Maggini M. A unified probabilistic framework for Web page scoring systems. IEEE Trans. on Knowledge Data Engineering, 2004,16(1):4−16.

[14] Alani H, Brewster C. Ontology ranking based on the analysis of concept structures. In: Clark P, Schreiber G, eds. Proc. of the 3rd Int'l Conf. on Knowledge Capture. Banff: ACM, 2005. 51−58.

[15] Zhang X, Qu YZ. Ranking problems in the semantic Web. Journal of Computer Science, 2008,35(2):196−200 (in Chinese with English abstract).

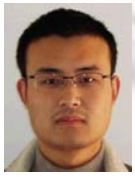
附中文参考文献:

[5] 尹导.一个本体采集系统的设计与实现[硕士学位论文].长沙:中南大学,2008.

[15] 张祥,瞿裕忠.语义网中的排序问题.计算机科学,2008,35(2):196−200.

张祥(1979－),男,江苏南京人,博士生,主要研究领域为语义网,网络分析,信息检索.

瞿裕忠(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为软件工程,语义网,Internet 计算.



葛唯益(1985－),男,博士生,主要研究领域为语义网,信息检索,本体匹配.