RDF 数据浏览的研究综述

吴鸿汉 瞿裕忠
(东南大学计算机科学与工程学院 南京 211189)

摘 要 随着语义网的快速发展,目前 Web 上语义网数据已经达到相当的规模,成为重要的信息和知识来源。因此,RDF 数据浏览的研究开始得到广泛关注。通过对比传统 Web 信息浏览和 RDF 数据浏览两个问题,指出 RDF 数据浏览的 5 个重要问题:确定浏览子图的模式、数据的收集、大规模数据的处理、数据的组织方式以及数据的呈现方式。基于这些挑战,我们调研了多个系统和不同的解决方案。最后,总结了目前的研究现况,讨论存在的挑战,并提出未来的研究方向。

关键词 语义网,RDF 数据浏览,RDF 数据组织,语义网浏览器
中图法分类号 TP399 文献标识码 A

Browsing RDF Data: State of Art Survey

WU Hong-han QU Yu-zhong
(School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

Abstract With the fast growth of Semantic Web data in recent years, the Semantic Web has become an important source of information and knowledge. Hence, the problem of RDF data browsing has attracted much attention in the Semantic Web community. We studied the differences between Web information browsing and RDF data browsing, and then identified some challenges of the latter, namely, graph pattern design, data collection, large scale data processing, data organization and data presentation. According to these challenges, we surveied a number of systems and tools. In the end we summed up the state of art and identified some directions of future work.

Keywords Semantic Web, RDF data browsing, RDF data organization, Semantic Web browser

1 引言

随着语义网的发展,目前 Web 上遵从语义网规范(RDF/RDFS 和 OWL 等)的数据已有相当的规模。据文献[1]估计,目前语义网上大概有 10 亿个语义网文档,而且其数量的增长呈现稳步上升的趋势。如何使得这些语义网数据直接为人们所使用,逐渐成为语义网研究人员关注的问题。在 Web 迅速流行的过程中,传统的 Web 浏览器所扮演的重要角色启发研究人员进行语义网数据浏览机制的研究[2-6]。

作为语义网的基本数据模型,RDF 是通用的资源描述框架,它使得信息和数据的集成更为方便。然而,从人们阅读和理解的角度来看,RDF 模型缺少 HTML 格式所能提供的文档结构:层次和次序[2],使得人们很难对 RDF 数据直接进行阅读和理解。因此,相比 HTML 的浏览,RDF 数据的浏览更加具有挑战。

1.1 比较传统 Web 信息浏览和 RDF 数据浏览

1.1.1 浏览对象——从"文档"到"RDF 数据子图"

传统 Web 是由文档互相链接而构成的,可以通过节点为文档的有向图模型来表示,如图 1(a)所示。该模型中每个节点都是由 URL 支撑的 Web 信息资源,它们就是传统 Web 浏览器处理的对象(如:浏览文档 1,由图 (a)所示)。这些信息资源通常都采用 HTML 进行标记,传统 Web 浏览器所做的就是按照 HTML 标记呈现该信息资源。

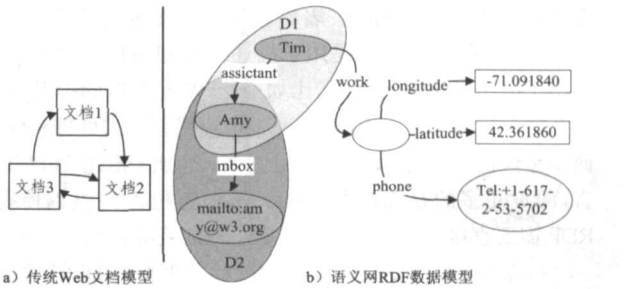

图 1 传统 Web 文档链接模型与语义网数据模型

在语义网环境下,尽管可以通过语义网文档之间的链接关系构建类似于图 1(a)所示的文档互连的图结构[7],进而使得类似于传统 Web 以文档为单位的信息浏览成为可能,但是以语义网文档为单位的方式并不适合 RDF 数据的浏览。因为,一方面,目前 Web 上发布了很多巨大的 RDF 文件,这些文件是数据发布者将所有的 RDF 数据打包在一起提供给使用者下载,比如:DBLP[8] 和 DBpedia[9] 都提供这样的文件。很显然,这种情况下以文档作为浏览的对象就很不适合了。另一方面,RDF 模型使得来自不同应用、不同来源的数据能

到稿日期:2008-03-20 本文受国家自然科学基金项目(语义 Web 本体的搜索方法与技术,No. 60773106)资助。
吴鸿汉 博士研究生,主要研究方向为语义网,E-mail: hhwu@seu.edu.cn;瞿裕忠 教授,博士生导师,主要研究方向为 Web 科学、软件工程。

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

够很方便地集成,以提供更具吸引力的信息服务。而以文档为单位进行 RDF 数据浏览,无法体现这种由 RDF 模型所带来的数据重用和数据互连的优势。

语义网的目标是构建 Web of Data,在 Web 范围内实现数据的集成和重用。这个理想所依赖的数据模型是由图 1(b)所示意的 RDF 模型。在该模型下,信息通过节点以及节点之间的关系来表达。因此,在语义网中的信息单元(也就是被浏览的对象)通常是某个 RDF 图中的子图(如:有关 Amy 的信息,由图 1(b)中的 D1 和 D2 所覆盖的数据)。这种浏览对象的变化带来了一系列需要解决的问题:

浏览的子图模式。浏览的第一步是确定浏览的对象,从上面的分析我们知道对于 RDF 的浏览,就是从给定的 RDF 数据集中确定所要浏览的 RDF 子图。从 RDF 数据查询的角度来看,也就是确定用于匹配的子图模式问题。因此,RDF 浏览工具第一步需要做的就是,定义其所浏览的子图模式。

数据的收集。在开放的语义网环境下,作为浏览对象的 RDF 子图一般都是由来自不同数据源的 RDF 数据组成的。比如,图 1(b)中所示的有关 Amy 的信息分别来自 D1 和 D2 两个语义网文档。因此,一个 RDF 浏览工具需要考虑如何收集这些分布的 RDF 数据以及收集的调度和收集的策略等。

大规模数据的处理。在语义网的开放世界假设下,很难保证匹配上述子图模式的 RDF 子图的规模都限制在浏览工具或是机器的可处理范围之内。因此,如何限制规模,从这个子图中选取恰当的部分进行浏览也是一个具有挑战性的问题。

1.1.2 描述语言——从"面向呈现"到"面向数据"

传统 Web 信息浏览是基于 HTML 标记语言的。在 Web 范围内发布信息,一个基本要求就是使用一种被广泛接受和理解的语言。HTML 就是这样一种用于信息发布的语言,它面向信息的呈现,提供一系列语法构件用于描述信息该如何进行格式化,比如:标题、表格、列表以及超连接等[10]。与 HTML 不同的是,RDF 是用于描述资源相关信息的语言,主要用于描述资源的元数据,比如一个 Web 文档的标题、作者以及修改时间等[11]。RDF 主要是被应用程序来处理和使用,而不是仅仅用于显示给用户阅读的。因此,相比 HTML 而言,RDF 更多的是面向数据的。这种定位上的差别使得当 RDF 需要直接呈现给用户时带来了新的挑战:

数据的组织方式。从人们阅读的角度看,RDF 模型缺少适合人类理解的信息组织形式。文献[2]指出 RDF 数据模型缺少三元组之间的层次结构以及相互之间的前后顺序。比如,有关 Tim Berners-Lee 的 RDF 数据可能同时包括他的个人信息(FOAF)以及他所进行的研究项目的信息(DOAP),适合人们阅读的组织方式通常会将两者区分开来,在不同的区域进行呈现,而 RDF 模型并不支持这样的组织方式,两者的数据不论是分开还是交错,在 RDF 模型下都是等价的。因此,RDF 数据的呈现的一个重要问题就是如何按照适合人类阅读的方式来组织 RDF 数据。

数据的呈现方式。浏览工具最终需要把 RDF 数据通过可视化的方式呈现给用户。HTML 通过标题、表格以及列表等一系列标记定义信息的呈现方式。作为面向数据的 RDF 模型,并不包含数据呈现的信息。因此,数据的呈现和可视化问题也是 RDF 数据浏览需要解决的问题。

上面的讨论从浏览对象以及描述语言两个方面探讨了传统 Web 的信息浏览和语义网的数据浏览间的主要差别,以及由这种差异所带来的 RDF 数据浏览所需要解决的 5 个主要问题。第 2 节简要介绍相关工具和系统。第 3 节分别从浏览的子图模式、RDF 数据的收集、大规模数据的处理、数据的组织以及呈现方式 5 个方面讨论目前已有的解决方案。最后给出总结以及未来的研究方向。

2 相关工具及系统

从语义网提出至今,如何开发基于语义网数据的工具或系统为人们提供更好的信息服务一直都是语义网研究人员关注的重要问题。特别是国际语义网大会中举办的 Semantic Web Challenge 系列赛事更是促进了很多具有创新性的语义网应用的涌现。其中很多工具或系统都面临着如何处理,组织和呈现 RDF 数据的问题,即 RDF 数据的浏览问题。本节根据这些应用的特性,将它们分为以下 3 类。

面向本体编辑和可视化的工具。这类工具包括编辑工具 RDFAuthor[12], Protégé[1], IsaViz[13]; Protégé 可视化插件 OWLViz[2], OntoViz[3]; W3C 的 RDF 语法验证服务 RDF Validation Service[4]。这类工具对于数据的处理,它们一般都采用一次载入的方式,因此能够处理的数据规模较小。数据组织方面,提供相对简单的组织形式,比如:Protégé 按照实体的类型提供了类、属性和实例三种视图。可视化方面,一般都采用基于图的方式进行呈现。这类工具的用户大都是语义网领域的专家或者开发人员。

面向特定领域的语义网应用系统。这类系统包括计算机学术领域的应用 CS Aktive Space[14];音乐领域的应用 mSpace[15]。由于针对特定的领域,这类系统可以事先对数据的组织和呈现进行定义。虽然这种方式不能直接被通用的 RDF 数据浏览系统所采用,但是由于这些应用和通用的 RDF 数据浏览系统一样都是直接为最终用户服务的,因此,这些应用的成功在 RDF 数据的组织和呈现方面为通用的 RDF 数据浏览提供了宝贵的经验和努力的方向。比如,mSpace 项目将信息空间看作 n-dimensional space 非常接近刻面浏览(Faceted Browsing)的机制,推动了通用的 RDF 刻面浏览系统[4]的出现。

面向通用的 RDF 数据浏览的工具及系统。包括 RDF 浏览器 Tabulator[5], DISCO[16], OpenLink[6];领域无关的 RDF 数据浏览系统 Noadster[2], Haystack[3], Faceted RDF Browser[4];搜寻引擎的相关服务 Swoogle Term Digest[5], Falcons[6,17] 实体摘要服务[18]。这些系统是我们下文讨论和分析的重点。

---

1 http://protege.stanford.edu/

2 http://www.co-ode.org/downloads/owlviz/

3 http://protege.cim3.net/cgi-bin/wiki.pl?OntoViz

4 http://www.w3.org/RDF/Validator/

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

总之,目前有很多工具或系统从其应用的角度出发提供了某种机制来解决 RDF 数据的浏览问题。但是,通用的 RDF 数据浏览问题还存在很多需要解决的挑战,目前尚未达到令人满意的程度。文献[18]募集了 11 位志愿者对目前的主流语义网浏览器进行了一个可用性评估的实验,每位实验者在完成两个任务后都填写了基于文献[19]的系统可用性评估。统计结果显示,目前主流的语义网浏览器可用性(SUS[19])均分在 70 分左右(满分为 100 分);对于其中9 位打分比较接近的实验者的数据进行统计,均分在 50 分左右。大部分实验人员表示目前语义网浏览器的使用需要一定的专业知识,对于普通用户而言使用会比较困难。

3 技术方案比较

本节从解决通用 RDF 数据浏览的问题角度出发,基于该问题需要解决的 5 个挑战,详细讨论上述工具。表 1 简要地给出了上述工具在这 5 个方面的技术特性。据此下面详细分析和比较了各种技术方案的特点。

表 1 各种系统技术方案比较

系统	浏览的子图模式	数据收集	大规模数据的处理	数据的组织	数据的呈现
RDF 编辑及可视化工具	所有 RDF 数据	–	–	无 或者基于实体类型	图
CS Aktive Space	基于边的子图模式	Web Crawler	分页	人工定义刻面分类	HTML
RDF 浏览器 (Tabulator, DISCO, OpenLink)	基于节点的子图模式	303 递归收集	限制收集的数据规模	谓语聚类	HTML
Noadster	基于节点的子图模式	–	–	各种聚类方法综合	HTML
Haystack	基于预先定义的模板	–	–	基于模板的方式	HTML
Facted RDF Browser	基于边的子图模式	–	分页	刻面分类	HTML
Swoogle Term Digest Service	基于节点的子图模式	Web Crawler	摘要(文档排序)	谓语聚类	HTML
Falcons 实体摘要服务	基于节点的子图模式	Web Crawler	摘要(RDF 句子的排序)	概念空间和谓语聚类	HTML

3.1 浏览的子图模式

目前的 RDF 浏览工具所采用的子图模式大致可以分为两种:以 RDF 图模型中的节点为中心的模式,目前大部分的 RDF 浏览器[5,6,16] 通常都采用这种方式;另外一种用于匹配的模式以 RDF 图模型中的边为中心,其中典型代表就是 RDF 数据的刻面浏览机制[4]。

假设给定一个 RDF 数据集,如图 2 中的 RDF 图所示。以节点为中心的子图模式首先需要从数据集中选定一个节点(如图中指称 Tim 的节点),然后选取所有包含该节点的三元组(图中的红色部分)作为浏览的对象。图 3(a)给出了这个例子的 SPARQL 查询语句。可以看出在这种图模式下,只需要输入一个 URI 就可以确定浏览的对象。

文献[4]提出的 RDF 数据的刻面浏览机制所采用的子图模式是以 RDF 图中的边为中心的。这种浏览机制适合同一种类型的实体的浏览,它将该类型的每个属性及其取值都看作一个刻面(facet)。用户通过选择不同的刻面在给定的实体空间中进行浏览。比如,用户浏览类型为 foaf: Person 的人的信息,他可以通过系统给定的刻面,选择浏览工作单位的主页(foaf: workplaceHomepage)为 http://www.w3.org 的所有人。图 3(b)给出了这个例子的 SPARQL 查询语句。可以看出,首先,这种类型的浏览一般需要事先给定所浏览的实体的类型(例中的第一个 FILTER 指定的 foaf: Person);其次,事先需要生成适合浏览的刻面的列表供用户进行选择(例中第二个 FILTER 指定的 Property 和第三个 FILTER 指定的该 Property 的值);最后,作为浏览对象的子图退化成符合某个查询的实体的列表(例中使用 SELECT 而非图 3(a)中的 CONSTRUCT)。

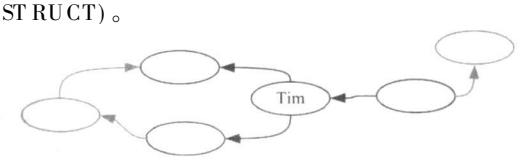

图 2 以节点为中心的查询模式

```
CONSTRUCT{ ?s ?p ?o}
WHERE
{
{?s ?p ?o.
  FILTER ?s =< http : //www.w3.o rg/People/ Berners - Lee/card# i >
  }
UNION
{?s ?p ?o.
  FILTER ?o =< http : //www.w3.o rg/People/ Berners - Lee/card# i >
  }
}
```

(a) 以节点为中心的图模式的SPARQL语句

```
PREFIX  rdf :     < http : //www.w3.o  rg/1999/02   /22 - rdf - syntax  - ns# >
PREFIX  foaf :    < http : //xmlns.co  m/foaf/0.1   / >
SELECT  ?x
WHERE
{
?x  rdf  : type  ?type.
FILTER  (?type  = foaf  : Person)
?x ?facet  ?value.
FILTER  (?facet  = foaf  : workplaceH  omepage)
FILTER  (?value  =< http  : //www.w3.o  rg/  >)
}
```

(b) 以边为中心的图模式的SPARQL语句

图 3 两种子图模式的 SPARQL 查询示例

通过上面两个例子的比较可以看出,基于边的子图浏览模式需要提供属性/值对的列表(即所谓的刻面)供用户选择。而从可用性方面来看,一方面这些刻面总数不宜太多,另一方面用户更希望看到经过组织的刻面列表,比如事先按照数据集的特征进行排序[4]。因此,事先对刻面的处理是必要的,这个要求使得基于边的模式并不适合开放环境下的浏览,比如在整个语义网中进行浏览。相比之下,基于节点的子图模式,用户只需要输入感兴趣的实体的 URI 就可以确定浏览的数据,更加适合通用的 RDF 数据的浏览。

3.2 数据的收集

根据浏览工具处理的数据集是集中的还是分布的可以将 RDF 数据的浏览分为两类。早期 RDF 数据的可视化和浏览工作一般都是面向单个文件或是集中式的数据源,无需考虑数据的收集问题。比如,面向本体编辑的工具[12,13] 以及 On-

[5] http://swoogle.umbc.edu/index.php?option=com_frontpage&service=digest&queryType=digest_swt

[6] http://iws.seu.edu.cn/services/falcons/

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

toViz 和 OwlViz 处理的都是单个本体文档。而 CS Aktive Space 和 BrowseRDF[4] 项目,处理的都是集中式的数据源。CS Aktive Space 在构建自己的数据源时,采用传统 Web 爬虫抓取的机制,从相关的信息源中抓取信息,以事先给定的本体进行描述,存放在集中的 RDF 数据源中。虽然也涉及了数据的收集,但是最终浏览的对象还是集中管理的 RDF 数据。

以 Tabulator[5] 为代表的 RDF 浏览器目标是使得分布的语义网数据能够直接为最终用户使用,它们的视角是将整个语义网看作一个大的可浏览的图(Browsable Graph)[20]。通过第 3 节的讨论,我们知道,RDF 浏览器的浏览模式是基于节点(URI)的。在语义网这个庞大的 RDF 图中,匹配这个模式的子图的数据可能来自不同的数据源。因此,要使得这种浏览机制能够实现,需要提供以 URI 为中心的 RDF 数据的收集机制。

3.2.1 303 重新定向机制

万维网是一个由资源组成信息空间[21]。Web 体系结构[22] 将资源分为信息资源和非信息资源。传统 Web 上的文档、图片以及其他媒体文件都是信息资源。Web 体系结构提供一种称作 303 重新定向的机制为非信息资源发布数据:当一个 HTTP 的客户端 dereference 一个非信息资源的 URI– u1 时,服务器端通过 HTTP 响应代码 303 将其重新定向到某个信息资源的 URI– u2,该信息资源包含一组有关 u1 所指称的资源的三元组集合[20]。

303 重新定向的机制使得语义网浏览器[5,6,16] 可以收集给定 URI 所指称的语义网实体的 RDF 数据。文献[21] 将通过 303 机制直接获取的数据称为该实体的权威描述。然而,语义网提倡的数据互连和 URI 重用使得描述同一个实体的数据可能来自不同的数据发布者。因此,303 机制并不能直接解决分布数据的收集问题。为了收集到权威描述之外的数据,目前的语义网浏览器通常采用递归收集的方式,扩大收集的覆盖性。比如,DISCO[16] 浏览器首先 dereference 用户给定的 URI,获得一个 RDF 图 G,然后再通过 303 的机制收集 G 中新发现的 URI 的数据,形成一个递归收集的过程。该过程结束条件是 1) RDF 三元组总数达到 15,000 个;或者 2) 所有 URI 都已经收集过或者连接超时(每个 URI 的超时设定为 2 秒)。然而,这种递归的 303 收集机制也很难保证数据收集的覆盖性。因为,递归进行 303 搜集到的数据只是那些从 URI 的权威文档出发有路径可达的。比如,在图 1(b)中,假设图中的 D2 为 Amy 的权威描述,采用 303 递归收集机制无法收集到 D1 中有关 Amy 的 RDF 描述,因为从 D2 并没有路径可以到达 D1。

3.2.2 爬虫收集

由于 303 收集机制并不能覆盖一个资源的所有的 RDF 数据,Sindice[23] 提出了一个第三方的解决方案。Sindice.com 是一个语义网的查询索引。给定一个资源的 URI,它可以返回哪些文档或是数据源包含该资源的语义网数据。Sindice 通过语义网爬虫在 Web 上收集语义网文档和数据源,并通过建立 URI 到数据源的索引来提供上述服务。值得一提的是,他们的爬虫通过一个爬虫本体[7] 进行指导,可以优化效率并避免给被抓取的站点带来太多的负担。比如:某个站点如果提供一个包含所有三元组的大文档供下载的话,该爬虫就不会通过 303 机制去抓取每个 URI 的 RDF 数据。

此外,语义网搜索引擎 Swoogle 和 Falcons 同样通过爬虫抓取的方式收集一个语义网实体的所有 RDF 数据,为每个实体提供 RDF 数据浏览服务(Swoogle Term Digest Service 和 Falcons 的信息摘要服务[18])。

3.3 大规模数据的处理

目前的语义网浏览器一般都通过限制 RDF 数据的收集规模来控制浏览的数据的规模。比如,DISCO 限制最大的 RDF 三元组个数为 1,5000 个。而文献[4] 通过分页显示的方式来处理大规模的数据。这两种方法都是一种初步的解决方案。限制规模的方法,丢失了部分 RDF 数据,有可能用户关心的信息就在这些数据之中。而文献[4] 采用的分页显示的方法,并没有提出结果排序的方法,因此,用户需要通过浏览大量的数据才能发现其所需要的信息。

传统 Web 的搜索引擎在处理海量搜索结果的浏览问题上采取排序算法。实践证明一个好的排序算法,比如:PageRank[24] 或 HITS[25] 算法,在大部分情形下使得用户需要的信息能得到较高的排名。因此,设计一种好的排序算法是解决大规模 RDF 数据的浏览问题的可行方案之一。Falcons 的实体摘要服务就是这种方案的一个典型应用。给定一个 URI,该服务为该 URI 所指称的语义网实体自动生成一个语义网信息的摘要,使得用户能够通过阅读少量的数据,快速获得对该实体的理解[26]。其中,RDF 数据的重要性通过综合基于链接分析的重要性[27]、用户偏好以及来源文档的重要性进行评估。通过链接分析的重要性评估,选取出 RDF 图结构中重要的数据;用户偏好的引入使得用户的需求直接加入重要性评估;来源文档的重要性是考虑数据源的质量。最后,实体的摘要通过选取前 K 个重要的 RDF 数据单元组成。

3.4 数据的组织

如果我们将给定的 RDF 数据看作一个信息空间,那么 RDF 数据的组织的目的就是通过某种方法划分和组织这个信息空间,以便于用户阅读。现有对 RDF 数据进行组织的方法大致可以划分为自动聚类、刻面分类和基于模板三大类。

3.4.1 自动聚类的数据组织

当人们在一个大的信息空间中进行浏览、阅读的时候,通常希望通过某种方式将内容相近的信息组织在一起,将整个信息空间组织成有限的、易于理解的若干侧面。比如,图书的目录和章节。在信息检索领域,在搜索结果较多时,通常采用自动聚类的方法方便用户在结果集中进行浏览,以定位到需要的结果。RDF 数据的自动聚类就是通过定义数据的特征,然后将具有相同特征的放在一个类别中的信息组织的方法。

RDF 浏览器[5,6,16] 对数据的组织就是采取自动聚类的方法。它们将三元组的谓语看作特征,将具有相同谓语的三元组聚类在一起。这种方式在一定程度上可以帮助人们浏览和定位到自己感兴趣的信息。但是,这种简单聚类的粒度太细,并不符合人们的思维特征。比如:对于描述某个人(如 FOAF:Person)的数据,人们通常希望将其不同侧面的信息区分开来进行组织,例如其个人信息(如 FOAF)和所参加的研究项目的信息(如 DOAP)。而基于谓语的聚类方法无法区分这

[7] http://sindice.com/srobotsfile

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

些信息,两者的信息可能交织在一起,给人们的浏览带来困难。

基于这个原因,Falcons 实体摘要服务提出了基于概念空间的数据组织方法。在语义网中,人们使用一组词汇来描述一个特定领域的概念及其关系,这组词汇被称作领域本体。而语义网实践中,一个领域本体中的术语的 URI 通常包含相同的前缀,这些被共享的前缀称作词汇 URI。比如: FOAF 词汇 URI 为 http://xmlns.com/foaf/0.1/。Falcons 的实体摘要服务将一个 RDF 句子(RDF 句子的概念可参见文献[27])的主谓语的词汇 URI 看作该句子的特征。然后,根据该特征对 RDF 句子进行聚类[18],如图 4 所示,对于 Tim Berners-Lee 的摘要,其 RDF 数据被聚成 FOAF, DOAP, DC 和 CONTACT 等若干类,分别描述不同概念空间下的信息。对于每个概念空间,其下的 RDF 数据再按照谓语的方式进行聚类。通过实际用户参与的实验表明,采用基于概念空间的组织方法比简单的谓语聚类能够提高浏览效率 4%~17%,对于更加熟悉 RDF 的使用者,能够提升 20% 左右。但是,基于概念空间的聚类的名称(图 4 中的标签页所示的 FOAF 和 DOAP 等)对于一般用户来说比较难于理解。这也正是这个原因,浏览效率的提升对于一般用户而言不太显著。

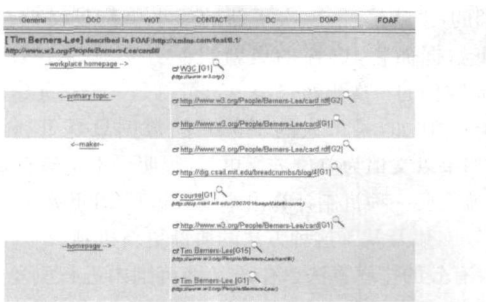

图 4 基于概念空间的数据组织

3.4.2 刻面分类

起源于图书管理领域的刻面分类理论(facet analysis theory)[28]是对信息空间的划分最有影响力的工作之一。该理论应用在信息搜索和浏览领域较早的工作是 Flamenco[8][29]。刻面搜索或浏览是采用刻面分类法对信息空间进行划分,使得用户可以在自己感兴趣的类别中搜索或浏览。被处理的信息对象(digital objects)通常都有共有的属性。分类法中的每个刻面(Facet)都对应到某个共有属性的取值范围。

根据 Facets 的生成方法,这些系统可以分为手工构建 Facets 系统和自动构建 Facets 系统。手工构建 Facets 是人工对数据集进行分析提取特征,并进行分类。手工构建 Facets 的优点是这些 Facets 对于用户而言含义清晰,便于理解和使用。南安普顿大学的 CS AKTive Space[14]就属于这样的系统。但是,手工构建 Facets 主要的问题就是需要大量的人工参与,适合应用于特定的领域,比如上述的 CS AKTive Space 就是针对计算机领域的应用。

文献[2,4,15,30]采用 Facets 自动生成的方案。文献[2,30]对生成的 Facets 采用聚类的方法生成层次化的 Facets。这种自动生成 Facets 的方法能够处理不同的数据集,但是主要的问题是自动生成的 Facets 是否合理,是否能够很好地帮助用户进行检索或浏览。比如,文献[4,30]以 RDF 三元组的谓语以及其取值作为 RDF 数据集的 Facets。这种方案在处理大数据集时,由于粒度过细,会导致产生过多的 Facets。因此,文献[4]中提出了 Facets 的度量标准,对 Facets 进行排序。该度量标准的基本思想是将用户的浏览过程看作构建和遍历一颗决策树的过程,这颗树的枝是 Predicates,节点是限制值(Restriction Values)。评价的标准主要考虑的是如何在数据集上保持该树的平衡。这种评价标准主要是从数据浏览的效率方面进行考虑,忽略了信息本身聚合性,比如,同样属于 FOAF 的数据可能被划分到不同的分支中去。

3.4.3 基于模板的数据组织

对于一个语义网资源,不同的用户可能会有不同的兴趣或信息需求,因此会从不同的角度去了解。基于这个原因,文献[3]将这种不同的视角抽象成"视图(Views)"的概念。一个资源可以在不同的视图进行浏览,其中默认的视图是全信息视图(All Information View)。该视图显示所有定义在给定资源的类型上的"透镜(Lenses)"。所谓透镜是指一组属性的集合,根据给定的领域的特征,这组属性组合在一起描述给定资源的同一个侧面。比如,对于一个航班,可以将其基本信息定义为一个透镜,包括以下属性:航班名称、航空公司、航班号、出发地、出发时间、目的地和到达时间。通过视图和透镜的组合,一个语义网资源的数据就能够以符合领域特征和适合人类阅读的方式进行组织。视图和透镜是人工定义在给定类型上的数据组织的模板。在文献[3]的基础上,W3C 定义了一组用于描述 RDF 数据显示的词汇[31]。开发人员可以使用该词汇来为给定的类型描述上述的模板。虽然标准化的词汇使得这种基于模板的数据组织可以实现跨越应用和组织的边界进行重用和共享,但是,这种模板的定义需要大量的人工介入,从 2005 年提出至今还没有得到广泛的使用。

3.5 数据的呈现

目前,RDF 数据的浏览/可视化主要包含两种方式:以图的方式展示和转化为 HTML 页面的方式。

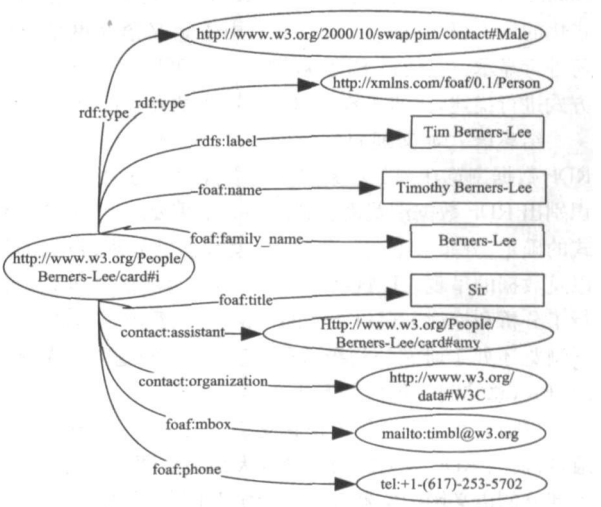

图 5 RDF 图模型

使用图的方式的工作包括:IsaViz, OntoViz, OWLViz 等,它们直接将 RDF 数据按照 RDF 图模型的方式展示出来,如图 5 所示,其特点是非常直观,适合展示数据集的整体形状和

[8] http://flamenco.berkeley.edu/index.html

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

稠密度[32]。但是其缺点也是明显的,文献[32]从用户交互的角度提出了图方式的6种不足,包括:结构平坦,不易编辑以及布局单一等。其中最大的问题是,当数据集逐渐增大时,基于图的界面很快就变得不可控,因而限制了采用图模式的浏览系统的可扩展性。

一种常见的RDF数据的可视化方式是将RDF数据按照给定的模板转化为HTML页面。采取这种方式最典型的系统是以Tabulator为代表的各种RDF浏览器。图6是Tabulator浏览Tim Berners-Lee的URI的截图。相比于图的方式,HTML方式在用户交互方面有着很大的优势。首先,传统Web的流行,使得大部分用户比较习惯于这种显示和交互方式;其次,HTML显示的灵活性,丰富的组件以及多种媒体格式的支持使得这种方式能够提供更为强大的浏览机制。比如,Tabulator可以用更加直观的方式显示RDF数据,图6中,对于Tim Berners-Lee的照片,HTML可以以直观的图片形式展示出来。

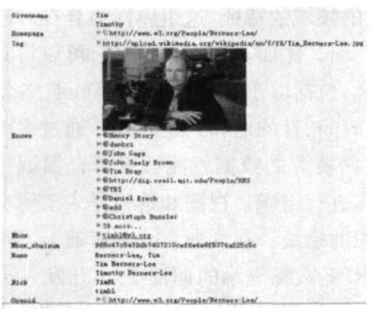

图6 Tabulator浏览器

除了确定RDF数据的可视化方式之外,RDF数据呈现还需要通过某种机制定义RDF数据该如何进行格式化,比如:通过HTML方式显示一篇文章的信息时,其标题应该通过粗体居中显示;通过图的方式显示某个人的数据时,其email地址应该加上下划线等。Fresnel[31]定义了一组称作格式化的词汇(Format Vocabulary)用于定义语义网信息该如何进行显示。它使得RDF数据的显示可以通过样式语言的方式进行描述,比如:CSS[33]和SVG中的样式部分。

结束语 本文通过对比分析传统Web的信息浏览和RDF数据浏览在浏览对象和描述语言两个方面的主要区别,识别出RDF数据浏览需要解决的5个主要问题:浏览子图模式的确定、分布数据的收集、大规模数据的处理、数据的组织以及数据的呈现。以这些问题为线索,对当前的研究现状进行了分析和探讨。通过上文的讨论,可以看出目前的RDF数据浏览还处于研究的初级阶段,还有许多问题需要做进一步的研究,总结起来包括以下几点:

更加灵活的浏览子图模式支持。通过上文的讨论我们知道,目前的RDF数据的浏览系统大都固定浏览的模式,比如,大部分的语义网浏览器都通过给定URI,浏览相关的RDF数据。然而,对于一个特定应用而言,通常需要执行一些常用的查询,来浏览相关的数据。Fresnel[31]定义了一种RDF上的数据选择语言(FSL),用于定义RDF图上的路径查询。领域专家或者应用开发者可以定义领域或是应用中的常用查询。因此,语义网浏览器需要能够执行类似于FSL这样的数据选择语言,提供更加灵活的浏览模式。

推动定制的数据组织和数据呈现。从上文中的讨论我们知道,RDF数据的浏览需要经过合适的组织才能直接为人们所浏览。除了上文中提到的自动组织的机制之外,应该允许RDF数据的发布者、语义网应用开发人员或是领域专家对数据的组织和呈现进行定义,因为通常这三类人员都比较了解数据该如何进行使用。Fresnel[31]提出一系列的词汇帮助人们定义这种数据的组织和呈现。然而,这种机制如何能够真正被广泛接受,涉及整个语义网研究领域的很多方面,包括:数据发布的机制是否能够支持发布者定义默认的组织和呈现方式,Web体系结构是否应该考虑将数据和其呈现通过某种机制结合起来,浏览器端如何支持这种机制以及允许用户在不同的组织模板中进行切换等等。

稳定高效的数据收集、更新机制。我们知道,一次语义网浏览通常需要从多个数据源收集数据。然而,Web上的应用的稳定性受到网络拥堵状况、服务器可访问性等诸多因素的影响。因此,当浏览需要访问多个服务器时,其稳定性和可靠性则相应下降。另外一方面,在服务器端的数据没有更新的前提下,重复的数据访问,不论是对服务器还是网络资源的占用都是一种浪费。因此,建立类似于传统浏览器端的缓存机制对于RDF数据的浏览同样重要。然而,传统的缓存是基于文档的,也就是说,一次更新的最小单位是文档。但是对于RDF数据而言,这样的更新机制并不是很有效率。因为,RDF数据的一次更新最小单位可以是一个三元组(或者RDF句子)。比如,某个人的FOAF的数据只是更新了email地址,如果以文档为单位进行更新,很明显不是最有效率的。因此,就需要一种机制计算出本地和远程RDF数据的差异的方法[34,35],以及基于这种方法的协商更新机制。

有效的排序算法。在语义网范围内进行浏览,符合条件的RDF数据可能会是一个庞大的三元组集合。一个有效的排序算法能够帮助人们快速获得高质量的、有用的信息。由于语义网上的数据重用和提取的粒度更细,比如,可以分别以实体、RDF三元组和语义网文档为单位,因此,语义网上的排序算法需要处理的类型的多样性以及数据的规模都远远大于仅以文档为单位的传统Web上的算法。

语义网应用的支持。传统Web浏览器除了可以帮助人们在Web上浏览互连的Web文档之外,还在目前广泛存在的Web应用中扮演着客户端的作用。语义网的发展将会促进语义网应用的发展,语义网浏览器同样需要在语义网应用中担当用户接口的作用,并且能够利用数据的语义提供更好的服务。比如,假设某位研究人员需要定机票去某个城市参加国际会议,语义网浏览器可以根据会议的日程自动筛选出符合条件的航班供选择。

总之,随着语义网的发展,RDF数据的浏览问题已经得到广泛的关注,包括:各种语义网浏览器的涌现、W3C相关草案的提出、Web体系结构的支持,以及专门的Workshop(Semantic Web User Interaction Workshop)的举办。到目前为止,这个领域还有许多挑战需要解决。因此,如何提供更好的RDF数据的浏览服务是具有潜在研究价值的重要问题。

参考文献

[1] Ding L, Finin T. Characterizing the Semantic Web on the Web// Proceedings of the 5th International Semantic Web Conference. 2006

(下转第41页)

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

works‖ Proceedings of the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom_00). Boston, MA, 2000

[11] Braginsky D, Estrin D. Rumor routing algorithm for sensor networks‖ Proceedings of the First Workshop on Sensor Networks and Applications (WSNA). Atlanta, GA, 2002

[12] Karp B, Kung H T. GPSR: greedy perimeter stateless routing for wireless sensor networks‖ Proceedings of the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom_00). Boston, MA, 2000

[13] Yu Y, Estrin D, Govindan R. Geographical and energyaware routing: a recursive data dissemination protocol for wireless sensor networks. Technical Report UCLA-CSD TR-01-0023. UCLA Computer Science Department, 2001

[14] Sohrabi K, Gao J, Ailawadhi V, et al. Protocols for self-organization of a wireless sensor network. IEEE Personal Communications, 2000, 7(5): 16-27

[15] He T, Stankovic J A, lu C Y, et al. SPEED: a stateless protocol for real-time communication in sensor networks‖ Proceedings of International Conference on Distributed Computing Systems. Providence, RI, 2003

[16] Sabbah E, Majeed A, KangK D, et al. An Application Driven Perspective on Wireless Sensor Network Security‖ Proceedings of the 2nd ACM International Workshop on Quality of Service & Security for Wireless and Mobile Networks. Terromolinos, Spain, 2006: 1-8

(上接第 10 页)

[2] Rutledge L, Hardman L. Making RDF presentable: integrated global and local semantic Web browsing‖ Proceedings of the 14th International Conference on World Wide Web. 2005: 199-206

[3] Quan D A, Karger R. How to make a semantic web browser‖ Proceedings of the 13th International Conference on World Wide Web. 2004: 255-265

[4] Oren E, Delbru R, Decker S. Extending faceted navigation for RDF data‖ ISWC 2006, 5th International Semantic Web Conference. Athens, Georgia, USA, 2006, Proceedings, November 2006

[5] Berners-Lee T, et al. Tabulator: Exploring and Analyzing linked data on the Semantic Web‖ Proceedings of the 3rd International Semantic Web User Interaction Workshop. 2006

[6] OpenLink. [cited; Available at: http://demo.openlinksw.com/DAV/JS/rdfbrowser/index.html

[7] Ding L, et al. Swoogle: A Search and Metadata Engine for the Semantic Web‖ Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management. 2005

[8] DBL PXML Dump. 2006 [cited; Available at: http://dblp.uni-trier.de/xml/

[9] DBpedia. DBpedia 3.0 Download. 2008 [cited; Available at: http://wiki.dbpedia.org/Downloads

[10] W3C. Introduction to HTML 4. 1999 [cited; Available at: http://www.w3.org/TR/html401/intro/intro.html

[11] W3C. RDF Primer. 2004 [cited; Available at: http://www.w3.org/TR/rdf-primer/

[12] Steer D. RDFAuthor. 2003 [cited; Available at: http://rdfweb.org/people/damian/RDFAuthor/

[13] Pietriga E. IsaViz. [cited; Available at: http://www.w3.org/2001/11/IsaViz/

[14] Schraefel M C, et al. CS AKTive space: representing computer science in the semantic Web‖ Proceedings of the 13th International Conference on World Wide Web. 2004: 384-392

[15] Gibbins N, Harris S. Applying mSpace interfaces to the Semantic Web. Techinical Report. 2003

[16] DISCO. [cited; Available at: http://sites.wiwiss.fuberlin.de/suhl/bizer/ng4j/disco/

[17] Cheng G, et al. Searching Semantic Web Objects Based on Class Hierarchies. in Linked Data on the Web (LDOW2008). 2008

[18] 吴鸿汉, 瞿裕忠. 理解语义网实体: 基于概念空间的摘要方法. 电子学报(审稿中), 2007

[19] Brooke J. SUS: a"quick and dirty"usability scale. Usability Evaluation in Industry, 1996

[20] Berners-Lee T. Linked Data. 2006 [cited; Available at: http://www.w3.org/DesignIssues/LinkedData.html

[21] W3C. Architecture of the World Wide Web. 2004 [cited; Available at: http://www.w3.org/TR/webarch/

[22] W3C. Dereferencing HTTP URIs. 2007 [cited; Available at: http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14

[23] Tummarello G, Delbru R, Oren E. Sindice.com: Weaving the Open Linked Data‖ Proceedings of the International Semantic Web Conference (ISWC). Nov. 2007

[24] Page L, et al. The pagerank citation ranking: Bringing order to the web. 1998, Technical report. Stanford Digital Library Technologies Project, 1998

[25] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 1999, 46(5): 604-632

[26] Wu H, et al. Comprehensive Summarization of URIs on the Semantic Web 2007. School of Computer Science and Engineering, Southeast University

[27] Zhang X, Cheng G, Qu Y. Ontology summarization based on RDF sentence graph‖ Proc. of WWW. 2007: 707-715

[28] Ranganathan S R. Elements of library classification. Asia Publishing House New York, 1962

[29] Yee P, et al. Faceted Metadata for Image Search and Browsing‖ Proceedings of ACM CHI 2003. 2003

[30] Rutledge L, et al. Finding the story: broader applicability of semantics and discourse for hypermedia generation‖ Proceedings of the fourteenth ACM Conference on Hypertext and Hypermedia. 2003: 67-76

[31] Bizer C, et al. Fresnel- Display Vocabulary for RDF. 2005-2007 [cited; Available at: http://www.w3.org/2005/04/fresnel-info/

[32] Karger D, Schraefel M. The Pathetic Fallacy of RDF‖ SWUI06 Workshop at ISWC06. Athens, Georgia, 2006

[33] W3C. Cascading Style Sheets. [cited; Available at: http://www.w3.org/Style/CSS/

[34] GiovanniTummarello C M, Bachmann-Gmr R, Erling O. RDFSync: Efficient Remote Synchronization of RDF Models‖ International Semantic Web Conference. 2007

[35] Zeginis D, Tzitzikas Y, Christophides V. On the Foundations of Computing Deltas Between RDF Models‖ International Semantic Web Conference. 2007

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net