●刘 瑛（南京大学信息管理系 江苏 210093）
黄 奇（国家信息资源管理南京研究基地 江苏 210093）

基于语义的网络信息资源组织

摘 要：基于语义的网络信息资源组织采用 XML、RDF 和本体技术，实现互联网上信息资源的语义描述和语义关联，完成信息资源组织从形式组织到内容组织、知识组织的转变，从而提高网络信息资源存取和利用的效率。

关键词：网络；信息资源；信息组织/本体

Abstract: The semantics-based Internet information resources organization uses the technologies of XML、RDF and Ontology to describe the Internet information and the relationship between interrelated concepts, completing the transformation from form organization to content and knowledge organization so as to improve the efficiency of storage and utilization of Internet information resources.

Keywords: network; information resources; information organizing/Ontology

20 世纪 80 年代，T. Berners-Lee 将超文本技术应用于计算机网络，从而极大地促进了互联网的发展。现在的互联网已经成为人们进行信息交流的重要渠道之一，任何人都可以通过互联网发布自己的信息，也可以通过互联网查找自己需要的信息。互联网上的信息每天都以指数级的速度增长，如何对互联网上的海量信息进行合理有效的组织以满足用户需求，已经成为国内外研究者共同关注的焦点。

1 现有网络信息资源组织方式的不足

网络信息资源组织即为网络信息资源提供有序化的结构，使之形成一个有机化的整体，以便于对网络信息资源的存取和利用[1]。根据网络信息资源的特征和构成及人们对网络信息开发利用的需要，很多研究者将网络信息划分为一次信息、二次信息和三次信息来对网络信息资源组织方式进行研究[2]。然而不管是哪一种组织方式，它的效率都将取决于一次信息的描述方式。下面就以搜索引擎为例进行说明。

作为目前网络上应用最为广泛的一种信息组织工具，搜索引擎的使用在一定程度上避免了用户在互联网上进行信息浏览的盲目性，给用户的信息搜索带来了方便。但是，随着网络信息资源的不断增长，现有搜索引擎在返回大量不相关结果的同时却又漏检了一些相关页面，远远无法满足用户需求。为此，很多搜索引擎采取了不同的技术进行了技术改良，如 Google 采用先进的 PageRank 排序技术保证重要的搜索结果排列在结果列表的前面；Goto 使用超链接分析和根据用户点击行为分析与重排序的方法，以提高检索结果的相关性；Askjeeves 采用逼近式方法让用户选择问题和答案来提高查准率，但效果并不理想[3]。

究其原因，可以发现，目前互联网上的信息主要是用 HTML 标记语言书写的，HTML 标记语言的简单性和易用性促进了互联网的快速发展，但是它的标签集只是对内容的显示格式做了标记，数据的表现格式和数据揉合在一起，缺乏对数据内容的标识。例如，在现在的互联网上，< H1 > orange < /H1 > 虽然有其特定的表现，但是 HTML 并没有明确地指出它到底是什么，是指水果还是指颜色，计算机根本无从判断。因而，作为对网络信息资源进行组织的工具，搜索引擎只能基于简单的形式匹配，无法对知识进行理解和处理，也就不可能真正理解用户的查询意图，不可避免地会出现一词多义和同义词现象，无法达到较高的查准率和查全率。

另外，由于网络信息之间缺乏良好的语义关联，现在的搜索引擎也不能将显示在不同网页上的相关信息整合在一起提供给用户。

因此，要提高网络信息资源组织的效率，就必须改变目前互联网上一次网络信息资源的描述方式，在网络信息创建之初就加入语义信息，实现基于语义的网络信息资源组织，从而有效提高网络信息资源组织的效率。

2 基于语义的网络信息资源组织

要实现基于语义的网络信息资源组织，必须依赖于两个重要的技术，它们是可扩展标记语言（Extensible Markup

Language，XML）和资源描述框架（Resource Description Framework，RDF）。

2.1 利用 XML 实现 Web 信息资源的深度标识

XML 是由 W3C 于 1998 年 2 月发布的一种描述任意文本结构的标准[4]，它的目的在于标识数据以便机器处理。与 HTML 不同，XML 将数据的内容与显示格式分开，允许使用者创建适合自己需要的标记，对信息进行确切描述，并使用文档类型定义（Document Type Definition，DTD）或 XML Schema 来约束这些标签的结构。这样，对于上文提到的“orange”，用户可以根据自己的需要创建 < fruit > orange < /fruit > 或者 < color > orange < /color > 对不同情况加以区分。

由于 XML 标签可以由用户根据自己的需要来定制，因此不可避免地会造成标签同名的情况，为了避免这样的冲突，W3C 采用了 NameSpace 机制，通过统一资源标识符（Uniform Resource Identifier，URI）限定元素名字，从而避免冲突[4]。

XML 为基于语义的网络信息资源组织提供了网络信息描述的语法标准，用户可以创建 XML 标签实现对文档信息的深度标识，但是这些标签对人来说很容易理解，对计算机来说却很困难。因此，XML 只是提供了 Web 信息资源的语法描述标准，并没有实现 Web 信息的语义描述和语义关联。

2.2 利用 RDF 实现信息之间的语义关联

互联网上信息之间的语义关联可以通过 RDF 来实现。RDF 是一种描述和使用数据的方法[5]，它提出了一个简单的数据模型，通过属性（Property）和值（Value）来描述资源以及资源与资源之间的关系。RDF 的数据模型实质上是一种二元关系的表达，由于任何复杂的关系都可以分解为多个简单的二元关系，因此 RDF 的数据模型可以作为其他任何复杂关系模型的基础模型。同时，RDF 提供了一种基于 XML 的语法（称为 XML/RDF）用于保存和交换 RDF 图。例如对于“国家信息资源管理南京研究基地是资源 http: //irm. nju. edu. cn 的创建者”这样一个句子用 RDF 数据模型表示见图 1。

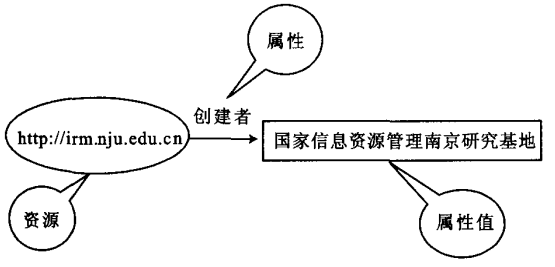

图 1 RDF 数据模型示例

用 XML/RDF 表示为：

```
< ? Xml version = “1.0” >
< rdf: RDF xmlns: rdf = http: //www. w3. org/1999/02/22-rdf-syntax-ns # ”
    xmlns: s = http: //description. org/schema/ >
  < rdf: Description about = “http: //irm. nju. edu. cn” >
    < s: Creator > 国家信息资源管理南京研究基地 < /s: Creator >
  < /rdf: Description >
< /rdf: RDF >
```

其中，xmlns: rdf 和 xmlns: s 指特定的命名空间。只有在命名空间中事先进行明确的定义，才能确保后面使用的 rdf: Description 和 s: Creator 有意义，并且避免产生理解上的混乱。

这样，利用 RDF 数据模型就可以实现信息之间的语义关联，解析器在阅读 XML 的同时，不仅可以获得 XML 所要表达的主题和对象，还可以根据它们的关系进行推理，从而做出基于语义的判断。

2.3 利用本体实现语义规范和约束

RDF 只是定义了一个描述资源关系的模型，但它还不具备解决同义词和一词多义的能力，基于 RDF 的数据语义描述存在着语义冲突。

为了消除语义冲突，实现真正基于语义的网络信息资源组织，必须引入本体的概念。本体描述的是一个特定研究领域的一个形式化的、共享的概念化模型。它通过对概念的严格定义和概念与概念之间的关系来确定概念的精确含义，为某个领域提供了一个共享的通用的理解，从而帮助人和机器进行明确的交流，支持语义级的交换，而不仅仅是语法级的[6]。

Web 上的本体通常包括分类和一套推理规则[7]。分类定义对象的类别及其之间的关系，使用户能够表达实体之间的大量关系，而根据推理规则，程序可以进行自动推理。简单地说，本体就是在不同的系统间定义一本字典或者度量表，使它们对实体及其之间的关系达成共识，以便交流和共享。

本体通常用基于逻辑的语言来表示，因此可在类、属性和关系之间做出详细、准确、一致且完备的区别。本体描述语言有 RDFS、OIL、DAML + OIL、OWL 等，其中 OWL 是 W3C 最新推荐的本体描述语言的标准[8]，它在 RDFS 的基础上引入了大量描述逻辑的建模原语，增强了表述能力。

OWL 有 3 个表达能力递增的子语言：OWL Lite，OWL DL 和 OWL Full，可以根据需要分别用于特定的实现者和

用户团体。

有了本体以后，一个 Web 页面上使用的术语或 XML 代码的意义可以通过该网页指向一个本体的指针来定义，从而解决网络信息资源组织中由同义词和一词多义所造成的语义冲突，进一步实现网络信息之间的逻辑推理。

3 基于语义的网络信息资源组织对互联网应用的影响

基于语义的网络信息资源组织其本质是利用元数据对信息资源进行描述，实现了从数据描述到概念描述的转变。可以预见，这种组织方式的实现将会对互联网应用产生重大影响。

3.1 实现网络信息的智能搜索

正如本文第一部分提到的，现有的搜索引擎基于简单的形式匹配，已经无法满足用户检索的需求。实现基于语义的网络信息资源组织，智能搜索引擎就可以在本体向导的指引下获取知识，检索到那些涉及某一确切概念的网页，而不是把那些使用含混不清的关键词的网页统统检索出来，从而提高检索的查全率和查准率。例如，对“orange”进行查询，现有的搜索引擎将含有关键词“orange”的所有网页都返回给用户，而不能判断 orange 指的是水果还是颜色。

但是，基于语义的搜索引擎就可以根据食品本体向导返回与桔子（orange）相关的页面，根据色彩本体向导，返回橙色（orange）的页面。

同时，现有的搜索引擎所面临的第二个问题也可以通过基于语义的网络信息资源组织来解决。因为 RDF 对信息之间的语义关系进行了描述，所以智能搜索引擎可以通过分析网页内容之间的相互联系，将不同网页中的相关信息综合在一起提供给用户。

3.2 实现 Web 信息的自动过滤

RDF 最初是为了配合 W3C 提出的因特网内容选择平台（Platform for Internet Content Selection，PICS）规范而提出来的[9]。因此，使用 RDF 也可以对不同的 Web 内容进行分级描述，以方便不同的用户进行选择。这样，实现基于语义的网络信息资源组织就可以保证 Web 信息的自动化过滤。

3.3 提供自动化服务

当前网上已经有许多自动的网上服务，但是因为现有的大量信息并不是以有意义的方式互相链接的，因此造成了信息之间存在着意义上的断层，而无法紧密地结合在一起，从而导致其他一些程序，例如代理，就无法对执行某个特定功能的服务进行定位。

如果能够实现基于语义的网络信息资源组织，信息之间具有意义上的良好关联，这样，软件代理就可以有效地自动化处理用户的需求，从一个服务登记转移到另一个，按用户的要求查找到所需要的 Web 服务，从而节省用户宝贵的时间。

3.4 提供个性化服务

基于语义的网络信息资源组织可以通过 RDF 来描述用户和他们的行为，这一点在电子商务中可以帮助商家对用户行为进行综合分析，然后根据每个用户的特点提供符合他们能力和偏好的个性化服务，从而实现良好的客户关系管理。

3.5 实现有效的知识管理

基于语义的网络信息资源组织在信息之间建立了良好的语义关联，在此基础上建立的基于本体的信息导航和询问系统，可以帮助用户快速方便地找到所需要的知识，并且将最恰当的知识在最恰当的时间传递给最合适的人，从而实现有效的知识管理。

4 结语

基于语义的网络信息资源组织采用 XML、RDF 和本体技术，实现互联网上信息资源的语义描述和语义关联，完成信息资源组织从形式组织到内容组织、知识组织的转变，从而有效提高互联网上海量信息资源存取和利用的效率。它的实现必将进一步促进互联网在电子商务、电子政务、知识管理等各方面的应用。□

参考文献

1 刘嘉．网络信息资源的组织——从信息组织到知识组织．北京：北京图书馆出版社，2002

2 高丹．网络信息组织方法研究综述．图书馆杂志，2004（10）

3 高凡．网络信息检索的发展方向．情报理论与实践，2004（2）

4 Bray T，et al．XML 1.0 Recommendation．http：//www．w3．org/TR/2000/REC-xml-20001006

5 Manola F，Miller E．RDF Primer．http：//www．w3．org/TR/rdf-primer

6 白同强，刘磊．语义 Web 的研究与展望．吉林大学学报（信息科学版），2004，22（2）

7 郭韦钰，丁连红．语义 Web 和语义网格概述．http：//bbs．w3china．org/dispbbs．asp? boardid = 57&ID = 14070

8 W3C Recommendation．OWL Web Ontology Language Guide．http：//www．w3．org/TR/owl-guide/

9 罗威．RDF——Web 数据集成的元数据集成方案．情报学报，2003（2）

作者简介：刘琰，女，1982 年生，硕士生。研究方向：电子商务，网络信息资源管理。

黄奇，男，1961 年生，教授。

收稿日期：2005－07－25