基于语义的Web信息检索

江克勤[1,2]，张玉州[1]，王一宾[1]
（1.安庆师范学院 计算机与信息学院，安徽 安庆 246011；
2.中国科学技术大学 计算机科学技术系，安徽 合肥 230027）

摘 要：语义万维网的研究逐渐引起了知识表示、逻辑编程、信息系统集成和开发等各个领域的广泛关注。文中概述了语义万维网的概念、技术框架，并且对含有自由文本和丰富语义标记的网络文档资源的三种语义检索系统原型进行了深入分析。最后，提出了设计Web语义检索系统应该满足的条件，可以基于它来设计语义检索系统框架。

关键词：语义万维网；语义检索；Web信息检索
中图分类号：TP391.3 **文献标识码**：A **文章编号**：1673-629X(2007)01-0036-04

Semantic-Based Web Information Retrieval

JIANG Ke-qin[1,2], ZHANG Yu-zhou[1], WANG Yi-bin[1]
(1.School of Computer and Information, Anqing Teachers College, Anqing 246011, China;
2.Department of Computer Science, University of Science and Technology of China, Hefei 230027, China)

Abstract: Research on semantic Web has aroused more and more interest in the fields of knowledge representation, logic programming, information integration and Web developing communities. Analyzes the concept and framework of the semantic Web. And analyses three implemented prototype systems that consist of both free text and semantically enriched markup. At last, present the necessary condition which require in the design of semantic retrieval system. Also can build semantic retrieval system framework on the basis of it.

Key words: semantic Web; semantic retrieval; Web information retrieval

0 引 言

语义万维网并不是一个孤立的万维网，而是对当前万维网的扩展。近年来提出的语义Web新标准——可扩展标记语言XML，它的特点就在于用户可根据需要制定能够反映任意数据内容的标签，实现数据内容和数据表现形式的分离。Web页面是一个有大量由机器可以理解的数据所构成的一个分布式体系结构[1]，在这个体系结构中，数据之间的关系通过一些特定的概念表达，这些概念之间又形成一种复杂的网络联系，计算机能够通过这些概念得到数据的含义，并且可以在这种联系上进行逻辑推理。

近两年来，语义万维网的研究逐渐引起了知识表示、逻辑编程、信息系统集成、开发等各个领域的广泛关注。其最大的特点在于用户可以根据需要制订能够反映数据内容的标签，实现数据内容和数据表现方式的分离及其相关的技术[2]，使传统万维网上的信息内容从面向人浏览转为同时面向计算机自动处理迈出了非常重要的一步。

收稿日期：2006-03-23
基金项目：安徽省教育厅自然科学基金资助项目(2001kj161, 2005kj364zc)；安庆师范学院中青年教师科研资助项目(01yj1002)
作者简介：江克勤(1970-)，男，安徽安庆人，讲师，硕士，研究方向为数据挖掘和数据库应用，计算机网络应用和信息安全。

1 语义万维网的组成

1.1 语义万维网层次图

由于语义万维网的知识表示具有创建上的分散性，同时又具有应用上的通用性，所以需要一个统一的框架，这个框架应该能够满足这种分散性以及由这种分散性所带来的安全性，满足这些知识跨应用、跨领域的可互操作性。语义万维网层次图如图1所示。

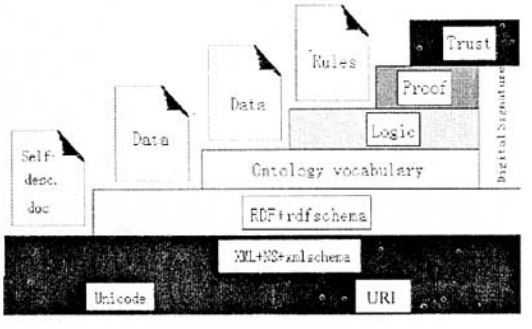

图1 语义万维网层次图

1.2 URI和Unicode

Web环境下的应用之间不可避免地需要相互通信，直接或间接地以机器可读的格式传递发布信息。语义万维网采用统一资源标识符(Uniform Resource Identifiers, URI)来标识资源及其属性。它和万维网常用的统一资源

定位符(Uniform Resource Locator, URL)以及统一资源名称(Uniform Resource Name, URN)的区别在于 URI 泛指所有以字符串标识的网络资源,包含了 URL 和 URN。另外由于语义万维网的最终目的是要构建一个全球信息的网络,在这个网络上应该涵盖各种语言和文字的信息资源,因此采用 Unicode 作为字符的编码方案[3]。这一层是整个语义万维网的基石,它成功地解决了万维网上资源的定位和跨地区字符编码的标准格式的问题。

1.3 XML, NameSpace, XML Schema

在 URI 和 Unicode 之上,是 XML 及相关技术层。XML 允许用户根据需要自定义一些"有意义的"标签对发布的内容进行标记,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。由于 XML 标签可以由用户根据自己的需要来定制,这样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。例如用户可以制定<author>标签:

<K: author xmlns: k = http://foo. bar. com/xml/customer. dtd>

它表明<author>这个标签是在 K 所代表的 NameSpace:http://foo. bar. com/xml/customer. dtd 中详细指明。这样即使其他人也自定义了<author>标签,只要它们的 NameSpace 不同,也不会造成冲突。因此,这一层通过 XML 的特性,实现了文档对自身结构的描述,实现了跨应用的语法互操作层,这是传统的 HTML 语言所无法完成的。但是,XML 是底层的数据交换格式,它只是解决了文档内容的次序、结构的问题,并没有解决文档内容的语义、联系的问题。标签的具体含义的定义和互操作要交给上一层去解决。

1.4 RDF, RDF Schema

XML 层的上一层是数据互操作层——资源描述框架(Resource Description Framework, RDF)和 RDF Schemas。RDF 本身并没有规定语义,但是它为每一个资源描述体系提供一个能够描述其特定需求的语义结构的能力。从这个意义上来说,RDF 是一个开放的元数据框架。这个元数据框架定义了一种描述机器可理解的数据语义的数据模型,主要包含下面的三个对象类型:

(1)资源(Resources):资源可能是整个网页;网页的一部分;或页面的全部集合;或者是不能通过 Web 直接访问的对象。

(2)特性(properties):特性是描述某个资源特定的方面、特征、属性或关系。

(3)声明(statements):一个特定的资源和特性名称加上该特性的值一起构成了一个 RDF 声明。声明中分为三个部分:主体(subject)、谓词(predicate)和对象(object)。从本质上说,RDF 定义 object - property - value 三元组作为基本建模原语并为它们引入了标准的语法。

RDF Schema 机制提供了 RDF 模型中使用的一个基本类型系统。这个类型系统有些类似于面向对象的编程语言。从描述逻辑(Description Logic, DL)的观点来看,RDFS 相当于 Tbox(Terminology Box),而 RDF 相当于 Abox(Assertion Box)。RDF 和 RDFS 共同描述前面事实所用到的结构,如图 2 所示。

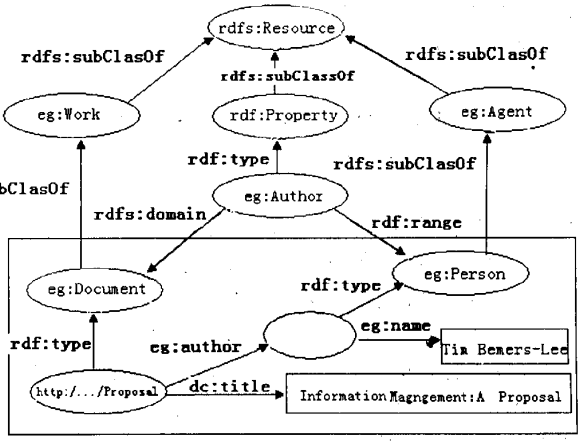

图 2　RDF 与 RDFS 结合描述事实

1.5 Ontology

RDF Schema 可以定义类、子类、超类,并且可以定义特性和子特性,以及它们的约束等。但是对特定应用领域的词汇的描述能力比较弱,需要进行扩展,这个 RDF/RDFS 之上的扩展层称为 ontology 层。ontology 是一种明确的共享概念化的形式说明[3]。概念化是指对现实世界中的一些事物进行抽象建模,所建立的模型确定了该事物的一些相关的概念。共享反映了这样的一种理念:ontology 表达双方都认可的知识,也就是说,ontology 并不会仅仅局限于某些个体,而应该被一个群体所接受。这一层在 RDF Schema 进行基本的类 Q 特性描述之后,更进一步地描述了术语和它们之间的联系,并且可以利用一些 ontology 语言[4],如(Simple Html Ontology Extensions)、OIL(Ontology Inference Language)、DAML(DARPA Agent Markup Language)等,来对领域知识进行建模,定义一些面向领域的共享词汇。

1.6 Logic, Proof, Trust

到目前为止,利用 RDF/RDFS 以及对 RDFS 进行扩展的一些 ontology 语言可以对 Web 上的资源内容做出描述。仅有这些描述还远远不够,基于语义的 Web 应用还需要根据特定的规则从这些描述性的知识中进行推理。逻辑层的目标就是提供一种方法来描述规则。描述逻辑标记语言(Description Logic Markup Language, DLML)就是这样的一种方法,它用 DTD 封装了描述逻辑中的逻辑连接词,可将基于描述逻辑的形式化知识嵌入到被描述的文档之中。

语义万维网环境下的应用在事实的基础上,通过应用逻辑推理,得出某种结论。首先应该可以信任所见的数据,并且可以信任所做的推理过程,只有在这个基础上才

能最终信任得到的结论。然而,就人们所见的数据而言,前面介绍的 RDF 模型允许任何人对任何资源进行任何描述,不同观点的人对同样的资源可能会做出完全相反的描述。正是因为 RDF 模型这种强大的描述能力,才有必要对这种描述进行身份认证并确保声明没有被篡改。因此在 SW 的层次体系结构中,从 RDF 层以上所有对资源的描述都贯穿了数字签名技术。数字签名技术就是对资源描述者的身份进行认证的关键性技术,是建立可信任网页的基石。

在 XML,RDF/RDFS, Ontology 以及 Logic 层和 Proof 层之上,人们就可以建立一些可以信任的应用了。

2 现有的基于语义的 Web 信息检索

通过分析现有三种语义检索原型系统,可以对其进行改进使其提供更好的服务。第一种系统框架是 OWLIR,它接受自由文本和结构化属性通过定制的查询接口进行查询。第二种原型系统是 Swangler,它是对传统的 RDF 文档加注语义标记,然后由像 Google 这样一般的传统搜索引擎索引。第三种原型系统是 Swoogle,它是对 RDF 文档基于爬虫索引的检索系统。它利用网络机器人程序搜索 RDF 文档同时把其中元数据加入到本地数据库中,它也可以把这些元数据加载到特殊版本的 HAIRCUT 检索引擎中,然后进行检索。

2.1 OWLIR 原型系统

OWLIR 是对含有自由文本和用 RDF 或 DAML 本体语言描述的语义标记的文档进行检索的系统[5](如图 3 所示)。

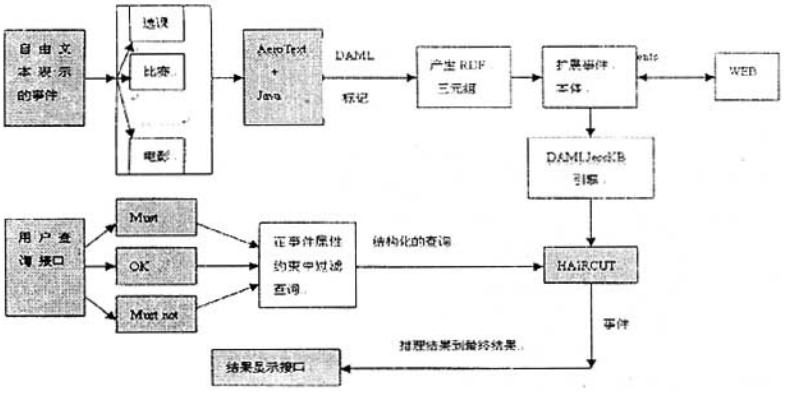

图 3 OWLIR 原型系统的数据流图

文本抽取事件通知是用自由文本表示的文档,但其包含有语义标记。我们利用 AeroText 系统对其抽取关键短语和本体元素。这些短语和元素对识别事件类型和增加语义标记起到一个非常重要的作用。AeroText 有一个 Java API 提供对抽取结果在系统内部形式的存取,使用 DAML 产生器部件存取这内部形式,然后把其翻译成相应的 RDF 三元组,这可以通过在抽取过程中直接绑定事件本体和语言知识库来完成。

推理系统 OWLIR 使用文本抽取过程中的元数据信息来推理语义关系,这些关系用来确定搜索范围。OWLIR 是基于 DAMLJessKB 的推理,DAMLJessKB 部件读取和翻译 DAML 文档,然后推理;DAMLJessKB 提供基本的事实和规则进行关系推理像子类和子属性等关系。

2.2 Swangler 原型系统

目前对 HTML 文档嵌入 RDF 或 OWL 等语义标记仍然没有一个统一的标准[6]。像 Google 这样的信息检索引擎本来就可以发现和索引 RDF 文档,但是 Google 只把这些带有语义标记的文档当作简单的文本文档来处理,其主要因素有:一是 XML 命名空间机制对搜索引擎是不透明的;二是用于处理自然语言的符号规则并不总是能很好地处理 XML 文档;三是不能直接利用这些语义标记。

我们已经运用 Swangler 技术于 SWDs 系统(如图 4 所示)中,与 OWLIR 系统类似,此系统中每个术语也用一个三元组表示。像 Google 的"机器人"程序可以搜集这些带

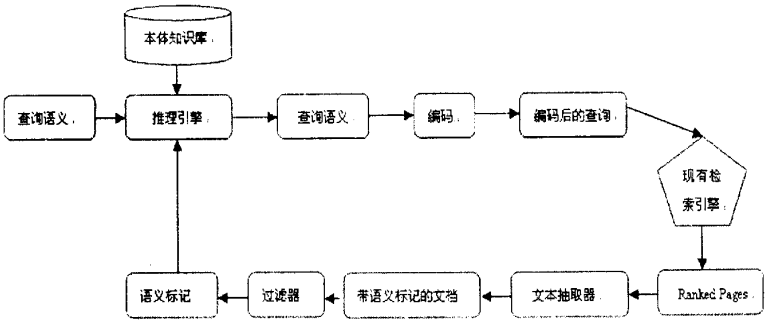

图 4 Swangler 原型系统框架图

语义标记的 RDF 文档,然后它可以索引这些内容表示成语义三元组的形式。这些三元组可以通过一个简单的接口来进行检索。

2.3 Swoogle 原型系统

由于当前语义 Web 中包含有用 RDF 编码的文档,所以可以考虑用专门的检索引擎对这些语义文档进行处理,使这些文档可以被机器所理解和处理。由于自然语言处理不是传统的搜索引擎任务,所以它不能翻译文档所代表的意思,即使可以,所付出的代价也是相当大的[7]。Swoogle 是一个针对用 RDF 或 OWL 这样的语义 Web 文档进行索引和检索的专用语义检索引擎。本系统由多个部件组成,包括:用来存储 SWDs 元数据的数据库,负责 RDF 网络文档搜集的机器人程序,计算有用文档元数据部件,计算 SWDs 中文档语义关系的部件,RDF 本体编辑器,索引器和一个用户查询接口。类似 Page Rank 概

念,此系统也有一个计算网页重要程度的算法 SWD Rank。该系统具有和当前基于关键字搜索引擎的所有相关技术,被 Swoogle 计算的元数据将提供关于语义 Web 的结构化信息等等。

2.4 三种原型系统的分析

先来分析一下以上三种系统异同点:①希望处理的是什么样类型的文档,三种系统要处理的都是用 XML 编码的 RDF 文档或者是带有语义标记的自由文档。②语义标记怎样被处理,是作为具有数据、知识模型结构化的信息还是作为与模型无关的自由文本信息。OWLIR 和 Swangler 把语义标记作为结构化的信息同时在其上进行推理。而 Swoogle 系统以文档内容意义存储这些 RDF 文档在本地数据库中,这就允许基于类、属性集作为检索文档的依据。③最终检索系统使用传统的检索引擎还是专用的语义检索引擎。Swangler 设计的目标是使像 Google 这样当前的检索引擎来检索语义 Web,而 OWLIR 和 Swoogle 采用的是专用的语义检索引擎。所以如果设计语义检索系统,就必须满足以下几个要求:

(1)此框架必须同时支持检索驱动的和推理驱动的处理过程;

(2)检索必须可以使用术语,语义标记及两者结合起来的术语索引;

(3)搜索是以文本为基础的现有搜索引擎或元搜索引擎;

(4)推理机制和检索机制应该紧密结合,检索性能的提高应该能够提高推理的准确性,同时推理性能的提高也将促进检索的准确率的提高。

3 结束语

语义 Web 的出发点是改变现有互联网依靠文字信息来共享资源的模式,通过本体来描述资源的语义信息,达到语义级的共享,从而提高网络服务的智能化、自动化。文中根据一些资料分析了三种语义检索系统原型,这对集成搜索和推理功能的语义检索系统框架的架设有很大的参考意义。

参考文献:

[1] Kopena J, Regli W. DAMLJessKB:A tool for reasoning with the Semantic Web[J]. IEEE Intelligent Systems, 2003, 18(3):74-77.

[2] 曹志松,曹文君.语义 Web 实现有效 Web 信息检索的研究[J].复旦大学学报:自然科学版,2004,43(3):422-427.

[3] An efficient algorithm for Web usage mining[R]. Montpelier, France:[s.n.],2003.

[4] Hotho A, Staab S, Stumme G. Explaining Text Clustering Results using Semantic Structures[R]. Karlsruhe, Germany: University of Karlsruhe,2002.

[5] Quan D. How to Make a Semantic Web Browser[R]. Cambridge, MA,USA:[s.n.],2003.

[6] Kevin, Chen Chuan Chang. Mining Semantics for Large Scale Integration on the Web:Evidences, Insights, and Challenges [R]. Urbana-Champaign:University of Illinois,2003.

[7] Laboratoire PriSM Université de Versailles Avenue des Etats-Unis Versailles Cedex. Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure [R]. France:[s.n.],2002.

(上接第 35 页)

5 结　论

文中提出的先验模型方法可以根据具体图像定义目标模型模板,在经典 Greedy 算法的基础上,增加外部约束力,弥补了原算法缺少外部约束力的不足。通过实验,对深度凹陷的区域本方法均可以取得较好的收敛效果,这也弥补了 GVF 算法的不足。但是,本算法在描述子的选择上还有待于改进,只是用模比值的归一化描述子,丢失了图像的方向性,蛇点轮廓在和先验模型的比较中,可能会出现偏差,有待于进一步研究。

参考文献:

[1] Kass M. Snake:active contour models[J]. International Journal of Computer Vision,1988(1):321-331.

[2] Cohen L D,Cohen I. Finite element methods for active contour models and balloons for 2D and 3D images[J]. IEEE Trans PAMI,1993,15(11):1131-1147.

[3] Xu C, Prince J L. Snakes, Shapes and gradient vector flow [J]. IEEE Trans Image Processing,1998,7(3):359-369.

[4] Williams D J,Shah M. A fast algorithm for active contour and curvation estimation[J]. CVGIP Image Understanding,1992, 55(1):14-26.

[5] Kauppien H,Sepanen H. An experiment comparison of autoregressive and Fourier-based descriptors in 2D shape classification[J]. IEEE Trans on PAMI, 1995(2):201-207.

[6] Persoon E, Fu L S. Shape discrimination using Fourier descriptors[J]. IEEE Trans on PAMI,1986(8):388-397.

[7] Zahn C T, Roskies R Z. Fourier descriptors for plane closed curves[J]. IEEE Trans on Computers,1972,21: 269-288.

[8] Chen H H,Su J S. A syntactic approach to shape recognition [C]//In: Proc Computer Symp. Tainan, Taiwan:[s.n.], 1986:103-122.

[9] Ge Yuan,Guo Xing-wei,Wang Li-quan. The Application of Fourier Descriptors to The Recognition of Alphabet Gesture [J]. Computer Applications and Software,2005,22(6):12-13.

[10] Jain A K. Fundamentals of Digital Image Processing[M].[s.l.]:Prentice Hall Press,1989:370-371.