基于 Semantic Web 的校园网知识检索的设计分析

Design And Analysis Of Knowledge Retrieval To Campus Net Based on Semantic Web

杨　枫　（四川南充 西华师范大学计算机科学学院 637002）
张莉华　（河 南 驻 马 店 师 专 计 算 机 科 学 系 463000）
钟乐海　（四川南充 西华师范大学计算机科学学院 637002）

摘要：传统的信息检索已经不能适应校园网知识资源扩大的需要，基于语义的知识检索克服了它的缺陷。本文提出了一种基于 Semantic Web 的校园网知识检索模型，把分散的知识资源整合，建立面向语义理解的数据模式，并对用户的检索请求进行推理，找出概念上的相关性，从而实现有效的智能检索。

关键词：语义网　本体　知识检索　元数据

1　引言

Semantic Web（语义网）是现在 Web 网的下一步发展方向，通过规范定义和组织信息内容，使之具有语义信息，能被计算机“理解”，从而更好的与人沟通。Semantic Web 的主旨是在 XML 的基础上进一步提供语义级互操作，XML 允许用户创建自己的标记从而使信息以一种半结构化的方式来组织。Semantic Web 用 RDF（Resource Description Framework）这种知识表示语言来注解这些半结构化数据所表示的含义，用 Ontology（本体）构建基于内容的知识检索模型，用 Ontology 的形式化描述语言 OWL（Web Ontology Language）来进行逻辑推理，从而实现真正的智能检索。

校园网的信息检索已经成为高校学生学习活动的重要组成部分，如何高效的进行信息检索是目前高校信息化工作的重点，本文将重点分析 Semantic Web 技术和知识检索以及它们在校园网知识资源上的应用，并给出一个基于 Semantic Web 的知识检索模型，分析这个模型的原理，指出它在应用上的现实意义。

2　Semantic Web 技术和知识检索

Semantic Web 作为当前万维网的扩展，于 1999 年由 Tim Berners－Lee 等人提出，其目的是通过结构化和形式化，以表示 Web 上的资源，使得计算机程序能够对网络资源进行分析和推理。Tim Berners－Lee 在 2000 年又提出了语义 Web 的体系结构，如图 1 所示。其中：

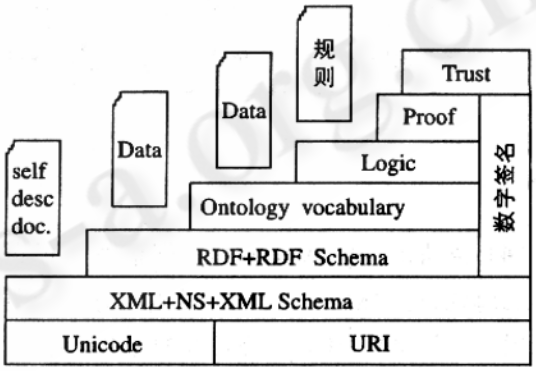

图 1　Semantic Web 的体系结构

（1）逻辑层（logical Layer），即前述逻辑知识体系，实际上，它们多局限于一定应用领域，称为 Ontology（本体）。Semantic Web 在该层次的任务是建立定义和标记 Ontology 的标准方式。

（2）语义层（Semantic Layer），即前述词汇体系，用以表达元数据（Metadata）。实际应用中，不同应用领域会根据不同目的、针对不同概念对象建立多种描述性元数据模式，对具体对象及其属性进行描述，例如 DC、MARC、EAD、GILS、MPEG7 等。Semantic Web 的任

务是建立基于 Ontology 来描述元数据元素、元数据的任务是建立基于 Ontology 来描述元数据元素、元数据关系和约束元数据语义的机制。有时，Ontology 可直接定义元数据，或者将某些元数据模式引入到 Ontology 中。

（3）赋值层（Assertion Layer），即前述赋值机制，以标准方式建立元数据与被描述资源的连接，从而保证计算机能明确地确认元数据、元数据含义及其与资源的关系。目前，RDF 正成为网络环境下的基本赋值机制。

（4）语法层（Syntax Layer），即前述标记语言标准，以计算机可理解和处理的方式对上述三个层次进行标记。目前，XML 作为网络资源的标准标记语言，也正成为语义和知识的基础标记语言。

在逻辑层之上，可能还需要利用 Ontology 和其他描述性元数据进行语义分析的推理机制，基于数字签名方式对 Ontology、元数据、赋值关系、甚至推理机制等进行确认和验证的机制，从而建立可信赖的语义环境和推理环境。另外在语法层下，还需要数据编码和资源标识的标准，如 Unicode 和 URI。

3 校园网知识检索模型及原理

在高校校园网中，知识资源的来源是多样的，这些数据在结构上存在很大差异，因此，信息化的处理过程也就面临着相应的变化。针对目前信息检索存在的资源表示语义贫乏和检索手段语义贫乏、查准率极低等问题，本文分析、研究并利用 Semantic Web 中语义丰富的 RDF 和 OWL（Ontology Web Language）资源表示、基于描述逻辑和知识推理以及知识检索中的最新研究成果和技术，给出一种基于语义的校园网知识检索模型，如图 2 所示。模型能够充分描述信息检索领域的概念空间，建立本体模型，实现对知识资源的语义标注，进行推理，从而在语义网的框架下实现概念检索，在此基础上构建提供多层次信息发现手段的智能检索体系结构。

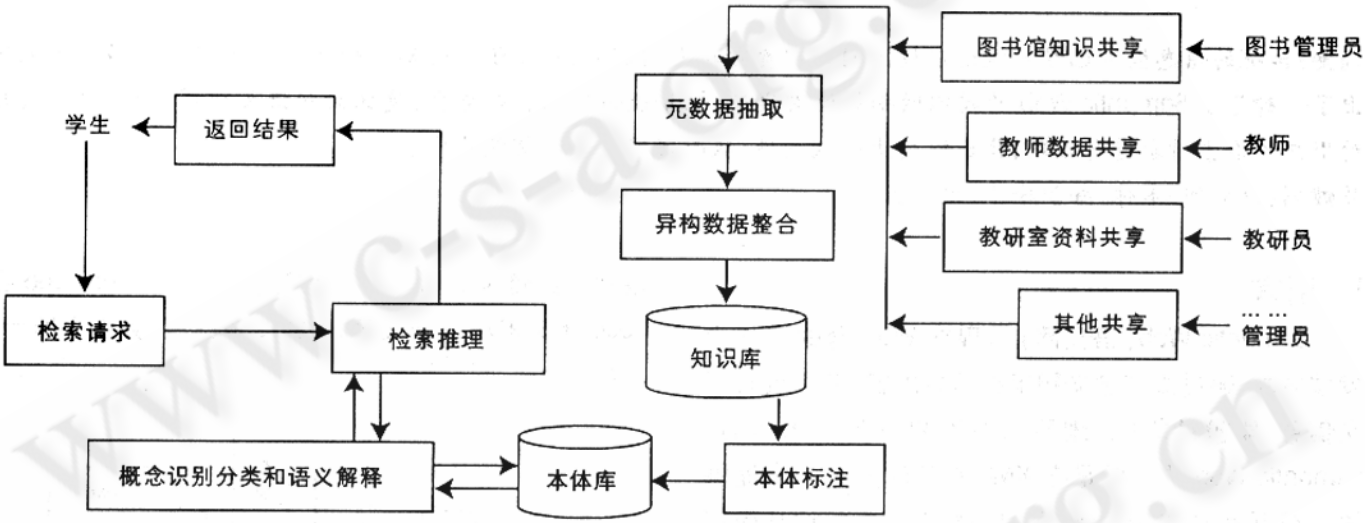

图 2　校园网知识检索模型

3.1 元数据抽取和异构数据整合

传统的信息检索往往只注重信息的字面意思，而忽略了信息的内容，这就使得传统信息检索无法在检索效果和质量方面有大的作为，因此，基于内容的检索技术就成为研究的热点。但是，如何确定信息的内容呢？对此，知识检索技术提出一种获取元数据的方法来解决这个问题。元数据就是那些用来描述信息资源的关键的信息点，它是一种描述资源和服务内容的数据。

元数据抽取就是要抽取知识资源的关键信息点，对于概念建模来说抽取能够揭示资源主题的索引词元数据是其重点。抽取出的元数据具有巨大价值，它在以下几个方面发挥重要作用：

（1）对资源的描述；

（2）增强各种资源之间的可交换性；

(3) 提高资源的可访问性;

(4) 为不同的数据格式架起沟通的桥梁。元数据是定义和组织知识资源的基础,它提供了一种精确描述数据内容和语义的机制,同时它还可以提供对服务的描述。但是由于知识资源类型多种多样,单一元数据标准不能满足描述各种数字资源的需要,从而出现适用于不同资源或不同组织的元数据标准。元数据的多样性和灵活性带来了元数据的集成和互操作问题。这就需要在异构数据之间进行整合。

随着数据结构的不同,信息化处理的过程也会发生变化,主要有以下几种:

① 处理的对象由以前的简单数据转化为现在的多模式知识,如声音、图形、图像、文字等。

② 处理的形式和方法由以前的数据结构表示方法转化为现代的知识表示,其中包括声音、图像、动画的综合表示。

③ 处理的过程由以前的算法指导数据的处理转化为采用推理方式指导问题的求解。

④ 处理的结果由以前的静态文本的方式输出转化为动态画面、多模式链接输出等方式。

3.2 本体标注

经过异构数据整合后,需要为检索系统构建 Ontology(本体),本体是用户提问语义和计算机检索语义取得一致的核心,因此构建本体并用本体对知识库进行标注是检索系统模型的关键。系统采用 OWL 描述本体。OWL 主要从以下几个方面表示本体术语及其间的关系:

(1) 类及类层次机构(子类和超类的关系)。类的定义除了直接定义以外,还可以通过描述逻辑(如析取、合取、否定等)或属性约束进行定义。

(2) 个体。表示本体中最具体的对象。

(3) 属性及其值。类属性的定义实际是对类个体的特征描述,其类型分为对象属性和数据类型属性,由具体的类确定。

(4) 类和类、类和个体、属性间的关系。它们的关系可通过约束和描述逻辑来表示。

提取元数据之后需要对所得到的信息集进行本体标注,生成 RDF 描述。主要的流程如下:

① 对知识库进行预处理,主要是文本提取,去除对 RDF 元数据抽取无用的文本等;

② 进行基于本体的文档分类;

③ 针对相应类别,进行属性提取,提取过程将使用相应的提取规则,这些规则主要是通过对样本集的分析和生成;

④ 把生成的数据转化为 RDF 描述。

3.3 概念识别分类和语义解释

Ontology 层次化的概念体系可以看作是一组分类,不同的概念形成不同的类别,而概念中的属性描述则对应于类别的属性,概念之间的语义关系也可映射成类别之间的关系。下面介绍如何进行 Ontology 概念识别。

识别概念的过程也即成为一个分类的过程,基于 Ontology 的分类有其特殊的地方:第一,分类的主要依据不是数据的主题内容,而是它的其他信息,这些信息将决定应该使用 Ontology 中的什么概念去描述数据;第二,分类还需综合考虑以后的 Ontology 属性和关系的识别。

针对以上的特殊性,需要采取符合实际的数据的预处理和特征表示及相关的分类方法。在这里数据即为表达相关主题的一组网页。通过对网页的分析,可以认为,网页的以下几个部分能很好的反映 Ontology 的类别:网页的 URL、网页的标题、网页的 Meta 信息、网页正文和网页中的超链接。对样本网页进行统计分析后,对以上每个部分都进行了词汇的选择。通过选择,除去一些无意义的单词以及其他无用单词,保留了能够反映特征的单词。

我们用模型论的方法来描述一个形式语言的语义,并对其作出解释。模型论假设这种语言是针对一个世界,而且描述了这个世界必须满足的最小条件集,从而给语言中的表达式指定相应的语义。形式语义理论所提供的技术手段可以用来确定推理的过程是否有效,即判断推理是否为真。语义定义基于 RDF 抽象语法。为此规定如下的 RDF 术语:URI 引用、常量、简单常量、带类型常量、匿名节点、三元组。

一个三元组是形如 <S,P,O> 的结构。可以对一个三元组集合与 RDF 图作如下约定:匿名节点用一个椭圆形表示。椭圆形和长方形的符号就是 URI 引用中的地址和常量。匿名节点所对应的图形没有符号。对每个三元组 <S,P,O> 作一条从 S 到 O 的有向边,并标以 P。从此约定可以看出,模型论的语义就是判断

一个句子,即对相应的世界作出声明。

3.4 检索推理

为了实现检索的智能化,对基于语义本体的检索必须有一种机制能够完成概念以及概念之间关系的推理,只有这样知识检索才有意义。这种机制就是要突破机械式字面匹配局限于表面形式的缺陷,从词所表示的概念意义层次上来认识和处理用户的检索请求。它必须实现语义蕴涵扩展、语义外延扩展、语义相关扩展,从而实现同义扩展检索和相关概念联想。

知识是语言和推理机制的结合,推理则建立在逻辑的基础之上,逻辑可分为命题逻辑、一阶逻辑、描述逻辑和框架逻辑。对于命题逻辑,其原子公式(Atomic Propositions)仅仅是真或假的陈述,而一阶逻辑的原子公式是对对象之间关系的陈述,因此,一阶逻辑使用谓词,并以常量或者变量作为参数。一阶逻辑可以进行推理,例如:Male(x) V Female(x)。其中,x 表示"任何一个人",V 表示"或"。故本例表示:对于任何人,要么 Male(x)是成立的,要么 Female(x)是成立的,也即一个人要么是男的,要么是女的。描述逻辑研究概念知识的表示问题,它提供了定义好的语义和推理机制。框架逻辑的主要作用是将概念建模集成到一致的逻辑框架中。

Jena 是一个 Java 开发工具包,它允许应用系统解析、创建和查询 RDF 模型。对 OWL 处理而言,语义逻辑的处理才是推理机制的实现。Jena 支持 OWL 的语义逻辑处理,它提供的 OWL 支持包括:方便的访问标准 OWL 的类及属性;支持多种版本的 OWL 规范;在基本查询中通过 subClassOf 这样的关系来实现类的层级访问和使用;可以注册用来映射 XML Schema 数据类型和 java 对象的转换器;支持基本的对 list 的处理;自动处理本体中 import 的 statement;识别传递(Transitive)属性和互斥(Inverse)属性。

4 系统实现

本系统的实现过程是:首先建立校园网知识资源的共享,它以各种数据方式存放在不同的数据空间中。在一定的权限下,这些数据能够方便获得和使用;其次,利用数据获取工具取得这些分布的知识资源,并对它们进行语义基础之上的元数据获取和整合,然后存放到知识库中;第三,构建本体并对知识库进行本体标注,然后把数据按照一定的格式存储在本体库中;第四,建立信息检索系统,系统对用户输入的检索词进行推理和概念识别分类及语义解释,并在本体库中进行检索,检索的结果进行识别分类和解释过滤之后返回给用户。

5 结论

基于 Semantic Web 的知识检索系统把校园网中的各种知识资源统一了起来,针对传统信息检索机械式匹配的缺陷,提出了基于语义的检索。对知识资源进行元数据抽取,并整合了异构数据;构建了知识领域本体并对知识库加以本体标注;对概念进行识别分类和语义解释,并利用概念和概念之间的关系构建了模型,从而实现对用户检索要求的推理。

参考文献

1 万捷、滕至阳,本体论在基于内容信息检索中的应用[J],计算机工程,2003,3。

2 张晓林, Semantic Web 与基于语义的网络信息检索[J],情报学报,2002,8。

3 楼向英,Ontology:概念及其在数字图书馆中的应用[J],图书馆杂志,2002,11。

4 Tim Berners - Lee. James Hendler. Ora Lassila. The SemanticWeb. http : / / www. sciam. com/ article. cfm ? articleID =00048144 - 10D2 - 1C70 - 84A9809EC588EF21.

5 张维明,语义信息模型及应用[M],北京 电子工业出版社,2002。

6 杜津媛、张立厚,在元数据标准框架基础上用 XML 构建语义网[J],图书馆论坛,2004,4。