《智能信息处理》课程考试

基于本体的语义查询研究

王威

考核	到课 [10]	作业 [20]	考试 [70]	课程成绩 [100]
得分				

2016 年 11 月 13 日

文章编号：1001-0920(0000)00-0000-00

基于本体的语义查询

王 威

（大连海事大学，信息科学技术学院，大连 116026）

摘 要：随着互联网技术的迅猛发展和网络信息资源的不断增加，基于关键字匹配的传统信息检索方法不能满足智能城市建设过程中海量数据处理的要求。基于本体的语义查询扩展通过本体语义信息和扩展推理使查询条件更符合用户需求。 意图，可以提高召回率和精度，并优化搜索结果。 基于本体语义查询扩展技术的研究，主流本体编辑工具Protégé被用于创建“计算机”域的本体，并且根据实际需要修改规则。 最后，应用于智能城市远程教育资源的个性化搜索，可以取得更好的效果。

关键词:本体，语义查询，智能信息检索

中图分类号：TP391　　**文献标识码：**A

Research on Semantic Query Based on Ontology

Wang Wei

(Department of Information Science and Engineering, Dalian Maritime University, Dalian 116026)

Abstract： With the rapid development of Internet technology and the continuous increase of network information resources, traditional information retrieval methods based on keyword matching can not meet the requirements of mass data processing in the process of intelligent city construction. Ontology-based semantic query expansion makes query conditions more consistent with user requirements through ontology semantic information and extended reasoning. Intent can improve recall and accuracy, and optimize search results. Based on the research of ontology semantic query extension technology, Protégé, the main ontology editing tool, is used to create the ontology of "computer" domain, and modify the rules according to actual needs. Finally, it can be applied to the individualized search of intelligent urban distance education resources, which can achieve better results.

Key Words: Ontology, semantic query, intelligent information retrieval

信息科学中的本体论首先被定义为“构成相关领域的词汇的基本术语和关系，以及利用这些术语和关系来定义词汇外延的规则”[1]。Borst“本体论是对规范概念的清晰定义”[2]解释了本体本质的本质。基于关键字匹配的传统搜索只能完成简单的查询，不能满足用户的多样化需求[3,4]。本体可以描述概念本身的概念，并且可以显示概念与概念之间的关系，具有良好的概念层次和逻辑推理支持[5]。基于本体的搜索引擎可以基于模糊关键词本体，通过输入搜索信息和知识结构以及与准确的基于知识的语义搜索相关的规则。语义查询扩展使用本体知识来扩展用户输入查询关键词，并且可以扩展和推理用户的检索要求，使得查询条件更符合用户的意图，在一定程度上提高系统的回忆和精度，并优化搜索结果。

基于本体语义查询扩展技术，主流本体编辑工具Protégé用于创建“计算机”领域本体，最终将应用于智能城市远程教育资源。 对于个性化搜索，希望得到一个很好的查询结果

1　本体与查询扩展

1.1 本体组成与语义关系

本体通常包括以下内容：（1）在域中对象

类的层次结构中的不同对象类之间存在一种，一部分，种类的关系，并且层次关系构成整个域（2）语义关系系统，即对象类之间的逻辑关系，例如，造成，使用，交互，协作类型，对象类系统；（3）对象类属性和属性值限制；有，监督，写作等；（4）对象类和推理规则之间的语义关系。

在领域本体中，存在四个主要语义关系[6]：（1）同义词（同义词），所述相似数据源在等价关系之间对称，即两个术语的不同本体具有相同的语义。例如，在教育领域，信息科学，例如，模式识别和模式分类可以被称为同义词。（2）主机关系（Hypernym），意味着一个本体中的术语的语义比另一个本体中的另一个术语的语义更一般更抽象。（3）下位词（hyponym），说一个本体在语义上的术语在另一个本体语义上更专业或更特殊。例如，同样是信息科学的情况，自然语言处理是子词技术之间的关系，对应的词分词技术是自然语言处理的下一个关系。（4）是一种关系（正关联），一种事物属于另一种事物，如关系的一部分。信息科学，例如信息收集，信息整理和用户查询和搜索具有“属”关系。

1.2 查询扩展

查询扩展属于自然语言处理和信息检索的类别。 它是重组初始查询输入以改进信息检索过程中的查询结果的过程。 在搜索引擎领域，查询扩展涉及分析用户输入，扩展查询术语以匹配更多资源。 其中用户输入不仅是输入搜索框的关键字，而且还包含其他类型的数据，如检索方法等。

查询扩展涉及四种技术[7]：（1）搜索关键字的同义词并检索它们；（2）查询词为每个词根处理，找到所有不同形式的条目；（3）正确的拼写错误，作为建议列表或直接搜索其正确的形式；（4）重新安排初始返回结果的重量。

基于文档空间向量距离的查询扩展，基于上下文平均互信息的查询扩展，基于特定问题类别的查询扩展，基于词间关联规则挖掘的查询扩展，基于用户日志扩展的统计模型，本文使用 基于本体的语义查询扩展

1.3 语义网的层次模型

语义网的基础是数据表示、数据查询、数据应用规则的一组标准，核心技术包括用于表示的RDF(Resource Description Framework，资源描述框架)用于查询的 SPARQL(Simple Protocol and RDF Query Language)、用于构造的 RDFS(RDF Schema)以及用于构造和推理的 web 本体语(Web Ontology Language,OWL)。

RDF 是语义网建立的基础，是将数据表示成有向带标记图(Directed Labeled Graph，DLG)的一种标准。资源是用全局唯一的、可解析的资源标识符(Uniform Resource Identifier，URL)标记的实体。图的结点是资源和字面量，结点之沈阳工业大学硕士学位论文间用又向边连接，边用谓词标记。图中的每一个边被称为一个语句描述(statement)，都有一个主词(subject)(边的起点)、一个谓词(边的标记)和一个宾词(object)(边的终点)。由于每个语句都有主词、谓词和宾词，所以两个结点以及连接它们的又向边也被称为三元组(triple)。每个语句的主词必须是一个资源，谓词也必须是资源，宾词也可以是资源或字面量。

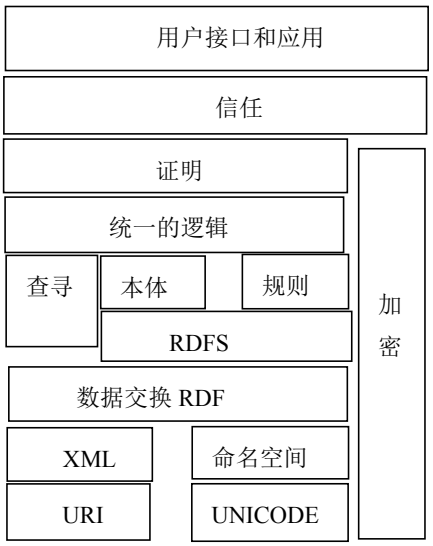

图 1.1 语义网的层次模型

在语义网的层次模型中可以看出，语义网是当前认 Web 的延伸，继承了 Web 的诸多优点。语义网的研究涉及到的领域是非常广泛的，并且各领域间是相互交叉的。另外一些层次和模块尚未充分研究，例如证明层、信任层和逻辑模块。

2 语义网的技术及其应用

语义网是建立在资源描述框架(RDF)的基础上的，而 RDF 借助 XML 的语法形式进行描述，所

以 XML 和 RDF 成为开发语义网的两项主要技术。语义网将要借助 XML 去定义用户化标记策略，要利用 RDF 灵活的方法去表示数据。然而，语义网的第三项关键技术是网络本体语言(OWL)，利用它来实现对网络文档中使用的类别和属性进行语义上的形式化描述。

2.1 XML

扩展标记语言 XML 是一种简单的数据存储语言，使用一系列简单的标记描述数据，而这些标记可以用方便的方式建立，虽然 XML 占用的空间比二进制数据要占用更多的空间，但 XML 极其简单易于掌握和使用。

XML 与 Access、Oracle 和 SQL Server 等数据库不同，数据库提供了更强有力的数据存储和分析能力，例如：数据索引、排序、查找、相关一致性等，XML 仅仅是展示数据。事实上 XML 与其他数据表现形式最大的不同是：他极其简单。这是一个看上去有点琐细的优点，但正是这点使 XML 与众不同。

XML 最大的特点是存储数据格式不受显示格式限制。文档通常包括数据、结构和显示方式三部分。由于 XML 描述了文档的每一成分和利用自身的嵌套结构定义了文档成分之间的关系，所以计算机更加容易理解 XML 文档，而不是 HTML 文档。XML 的一个优点是可以允许定义值约束，允许计算机可解读的信息表示方式。XML 的格式和内容是分开的，相同的信息可以用不同的方式来显示，而不需要同一内容的许多副本，在显示信息的同时还可以用来做别的有价值的事情。目前，XML 已经逐步成为数据和文档交换的标准机制之一。

2.2 RDF

RDF(Resource Description Framework)，即资源描述框架，是一种用于描述 Web 资源的标记语言。

RDF 是一个处理元数据的 XML 应用，所谓元数据，就是“描述数据的数据”或者“描述信息的信息”。也许这样解释元数据有些令人难以理解，举个简单的例子，书的内容是书的数据，而作者的名字、出版社的地址或版权信息就是书的元数据。数据和元数据的划分不是绝对的，有些数据既可以作为数据处理，也可以作为元数据处理，例如可以将作者的名字作为数据而不是元数据处理。

通过 RDF，人们可以使用自己的词汇表描述任何资源，但人们更乐意将它用于描述 Web 站点和页面，由于使用的是结构化的 XML 数据，搜索引擎可以理解元数据的精确含义，使得搜索变得更为智能和准确，完全可以避免当前搜索引擎经常返回无关数据的情况。当然前提是 RDF 和标准化的 RDF 词汇表在 Web 上广泛使用，而且搜索引擎需要能够理解使用的词汇表。

RDF 的主要概念包括:数据模型、基于 URI 的词汇、数据类型、字面量、简单示例表达式、推导。

RDF 表达式的结构是三元组集合，每一个三元组包括一个主词(subject)、一个谓词(predicate)和一个宾词(object)。三元组的集合称为 RDF 图。RDF 图可以通过结点和弧表示，其中每一个三元组表示为弧一结点一弧的链(link)。如图所示。每个三元组表示一个声明，它是关于结点所称的事物之间的联系。弧方向很重要，它总是指向宾词。一个 RDF 图的结点就是它的所有主词和宾词。

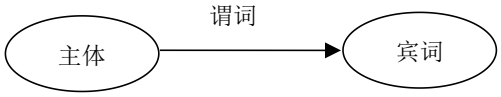

图 2.1 一个简单的 RDF 三元组

2.3 OWL

OWL 是 Web 本体语言，它是和语义网相关的 W3C 推荐标准栈的一部分。OWL 提供了三种表达能力递增的子语言，以分别用于特定的实现者和用户团体。分别包括：OWL Lite 用于提供给那些只需要一个分类层次和简单约束的用户。OWL DL 支持那些需要最强表达能力的推理系统的用户，且这个推理系统能够保证计算的完全性和可判定性。OWL Full 支持那些需要尽管没有可计算性保证，但有最强的表达能力和完全自由的 RDF 语法的用户。

OWL 是语义网活动的一个组成部分。这项工作的目的是通过对增加关于那些描述或提供网络内容的资源的信息，从而使网络资源能够更容易地被那些自动进程访问。由于语义网络固有的分布性，OWL 必须允许信息能够从分布的信息源收集起来。其中，允许本体间相互联系，包括明确导入其他本体的信息，能够部分实现这样的功能。

另外，OWL 提出了一个开放世界的假设。也就是说，对资源的描述并不局限于在一个简单的文件或范围内。如类 C1 本来是由本体定义出来的。然而，它也可以是由其他的本体扩展出来的。对 C1 进行这样的假设的结果是单调的。新的信息不能否定之前的信息。新的信息可以是和旧的信息矛盾的，但是事实和推导只能被增加而不能被删减。

2.4 语义网的应用

语义网并非像人们期待的那样出现的那么快，万维网协会已经促使其标准有所进展，但实际的语义网的应用技术还很少。部分原因是那些重要模式的复杂性，而我们应用的那些方面还无法有效地隐藏这种复杂性。另外一部分原因是语义网本身也缺乏用户友好的应用界面。

语义网的一般框架包括：从一个或多个数据源收集、合并和存储数据。对数据进行处理。查寻并使用数据，包括浏览、可视化。分析、集成等。

语义网典型应拥有很多，主要包括语义信息管理、语义信息检索、Linking Open Data。

3 智能信息的检索

3.1 信息检索的概念

网络信息检索，就是将描述特定用户所需网络信息的提问特征，与信息储存的检索标识进行异同比较，从中找出与提问一致或基本一致的网络信息的过程。信息检索技术是将因特网上的海量数据，通过软件系统的检索查询，根据用户提供的需求，把用户所要的信息提取出来，经十多年发展已经取得了不少突破性的进展，发展了一大批有效、不同的信息检索软件，如文本信息检索、动态网页 Web 信息检索、检索引擎等一些重要的检索方法和技术，特别是检索引擎，成了 Web 上查找信息不可缺少的工具。

3.2 信息检索的特点

(1)检索速度快：手工检索需要数日甚至数周的课题，计算机检索只需要数小时甚至数分钟。

(2)检索的途径多：除手工检索工具提供的分类、主题、著者等检索途径外，还能提供更多的检索途径，如题名途径等。

(3)更新快：尤其是国外的计算机检索工具，光盘多为月更新、周更新，网络信息甚至为日更新。

(4)资源共享：通过网络，用户可以不受时空限制，共享服务器上的检索数据库。

(5)检索更方便灵活：可以用逻辑组配符将多个检索词组配起来进行检索，也可以用通配符、截词符等进行模糊检索。

(6)检索结果可以直接输出：可以选择性的打印、存盘或者 E-mail 检索结果，有我还可以在线直接订购原文。有的计算机检索工具甚至可以直接检索出全文。

3.3 传统信息检所存在的问题

一是网络信息良莠不齐。网络信息的发布具有很大的自由性和随意性，缺乏规范，无用信息掺杂其间，垃圾信息、虚假信息、冗余过时信息的存在，增加了信息的不确定性和用户的不安全感，使信息质量和精度降低，其可靠性、权威性和利用价值受到质疑，令网络用户无所适从。

二是动态信息多，信息资源的分类比较混乱。由于网上动态信息多，而从事网络信息的工作人员大多不是专业的分类人员，缺少专业知识，同时又受到工作量的制约，不可能有时间去仔细考虑对信息资源的精心组织，使得网上信息资源的类目设置不够合理，划分类目的标准也比较混乱，即是同一类目的划分标准亦不统一，使得网上信息资源的组织没有规律、没有层次、没有逻辑性，类目容易出现重复，或遗漏现象。这就给用户的检索带来了不便。

三是用户使用网络检索的局限性。网络信息资源在数量、结构、类型、分布及传递手段等方面，都与传统的文献信息资源有着显著的差异，习惯于传统信息检索的用户，对网络信息资源的了解及利用还有一个认识、接受和熟练使用的过程；另外，用户的受教育程度、知识结构等原因，也造成用户利用网络资源的局限性。主要表现在：用户对检索需求的理解程度和检索策略制定得恰当与否，直接关系到检索的质量；用户的网络知识和计算机操作能力影响着信息检索的效率；用户对网络检索工具运用的熟练程度影响着检索的效果；用户的外语水平影响着信息检索的广度与深度。

3.4 提出解决检索问题的方案

3.4.1 建立特色数据库

特色数据库是指 Web 上提供的特殊数据库，

主要包括专业特色数据库、科研成果数据库、学位论文数据库、地语义网的智能信息检索系统模型方资源数据库等，信息服务机构应根据本地区历史发展和社会发展的需求结合本地人口、经济及文献资源现状建立特色数据库，努力做到“人无我有，人有我优，人优我特”。另外，可根据重点学科和文献资源特色等优势建立专题数据库，以特色服务来提高自己的社会地位，实现自身的价值。

3.4.2 开发多语种网络信息检索工具。

随着世界各地上网人数的增多，语言障碍越来越明显，网络信息检索工具应该能够减轻用户因语言不同而带来的障碍，提供多语种的检索功能。

3.4.3 加强基于内容的检索技术的研究。

当前的网络信息检索是根据 URL 进行定位搜索，常常返回死链接，这是因为 Web 信息更新太快，信息重组、移动、删除司空见惯，是由于索引库中 URL 不能同步更新所致。

3.5 基于语义网的智能信息检索系统模型

在对语义网技术与系统进行充分研究的前提下，利用语义网的核心技术和支撑工具，开发一个语义网知识组织系统，目标是实现知识的语义化组织与服务。此系统参考目前已有的叙词表和元数据资源，并在此基础上增加本体，构造资源之间的关系，使原来缺乏语义关系的数据得到更好的组织。其框架模型如图 3.1。

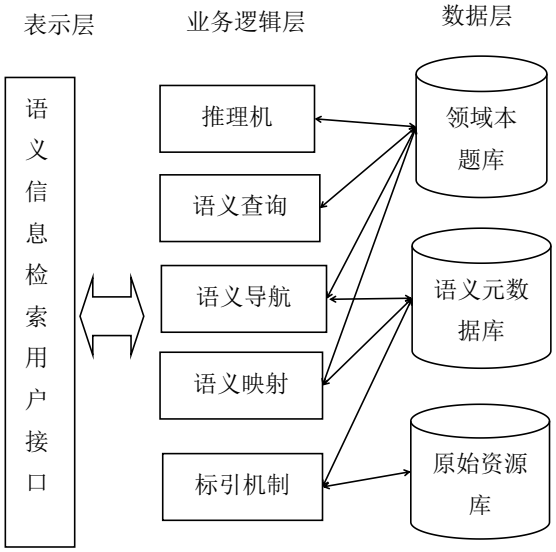

图 3.1 语义网知识服务系统层次

3.5.1 表示层

表示层是语义网的入口，负责与外部实体(如人或其他系统)通信。表示层向用户显示信息，为用户通过浏览器访问系统提供了一个可视化的接口。它把用户的请求发送到业务逻辑层，业务逻辑层进行处理后将语义信息返回给表示层。用户对表示层展现的语义信息进行评估，如果不满意，可以在此基础上优化其服务请求。也就是说，表示层和业务逻辑层有一个用户交互的过程，这样能让语义处理过程更好地满足用户的需求。

3.5.2 业务逻辑层

业务逻辑层接受表示层发送的请求，并负责在结果被返回到表示层之前对数据进行处理。所有的数据处理都在业务逻辑层完成。它包括 5 个部分：标引机制、语义映射、语义导航、语义查询以及推理机。

3.5.3 资源管理层

资源管理层包括所有语义网的数据资源，它能管理数据库、知识库、文件系统等重要的资源。本系统的资源管理层包括本体库、语义元数据库以及各种原始数据库等。

3.6 未来网络信息检索技术的发展方向

3.6.1 智能化

现有的检索引擎存在着查全率和查准率低的问题，未来的搜索引擎技术必须具有能及时挖掘新信息和及时能链接新增的信息，多途径检索功能，用户可以交互式检索，搜索出满意的信息。提高网络信息检索技术水平并实现智能检索，智能化是网络信息检索未来主要的发展方向。智能检索是基于自然语言的检索形式，机器根据用户所提供的以自然语言表述的检索要求进行分析，而后形成检索策略进行搜索，智能检索技术就是采用人工智能进行信息检索的技术，它可以模拟人脑的思维方式，分析用户以自然语言表达的检索请求，自动形成检索策略进行智能、快速、高效的信息检索。最近几年，智能信息检索作为人工智能的一个独立研究分支得到了迅速发展，而且目前已有一些搜索引擎支持智能检索，但智能化程度还不高，这方面还有待进一步的发展。

3.6.2 标准化

现在的网站信息瞬息万变，杂乱纷繁，很是需要进行分类整理。目前虽然有信息检索结课论文 大量的搜索引擎，但还没有一个统一严格的

分类方法来管理，网络信息资源在组织分类上需要制定一个统一的分类标准。还要规范网络术语，提高资源共享的程度，这样可以有效保证用户的检索效率。

3.6.3 个性化

科技的发展要以人为本，随着科学技术的发展，个性化服务也将成为网络信息检索的一个发展方向。随着互联网的飞速发展，每个人的对信息的需求将不再满足于标准化、单一化的大众需求。不同的人需要不同的服务，如残疾人士对网络信息检索的要求就要区别于常人，要是信息检索能很好的识别语音检索就能很有效的满足他们的信息需求。如何使用户更方便、快捷地使用各种检索工具，满足用户各种检索要求，个性化服务也会成为网络信息检索重要的发展方向。

4 结束语

本实验中使用的本体属于计算机领域，特别侧重于熟悉的计算机概念和相关教育资源搜索的扩展。 因此，查询扩展结果是理想的。 如果扩展到其他领域和专业领域，扩展规则需要改进，例如对应于专业缩写 PC 的“个人计算机”，一旦扩展到其它领域，其含义就会改变，例如在军事领域PCAnd 在网络游戏中代表“玩家角色”; 那么我们需要结合定义级别和用户资源描述文件进行个性化过滤，以确保搜索结果的正确性，这是下一步将开展的研究工作。

参考文献

[1]张榕宁.国外网络信息检索研究现状[J]. 图书馆论坛,2006,(8):188—190.

[2]杨青,王瑞菊.浅析网络住处检索中的问题与对策[J].图书馆学研究,2004,(6):82—83.

[3]李爱红.网络搜索引擎的比较研究[J].中国信息导报.1999(1):25 一 26.

[4]徐亨南.人工智能与智能信息检索.信息检索(江西图书馆学刊),2005,35(1):53 一 54.

[5]金海,袁平鹏.语义网数据管理技术及应用[M].北京：科学出版社,2010.

[6]黄果,周竹荣,周亭.基于语义网的信息检索研究.西南大学学报(自然科学版),2007,29(1):77 一 80.