《智能信息处理》课程考试

基于本体的信息检索学习

学　　院： 信息科学技术学院

专　　业： 软件工程

姓　　名： 姜俐伶

学　　号： 1120200323

授 课 老 师： 李冠宇

考核	课程成绩
得分	

大 连 海 事 大 学

2020 年 11 月 22 日

基于本体的信息检索学习

姜俐伶
（大连海事大学 信息科学技术学院 大连 116026）

摘 要：当今的在线教育平台中有大量的学习资源，但也存在搜索学习资源的困难，搜索结果的冗余以及缺少对关键字搜索的语义支持的问题。 经验表明，语义方法可以为我们所要探究的领域提供更好的技术。 通过将语义网的核心技术本体引入学习资源的检索中，通过本体分析，本体推理和本体中的概念相似度计算，设计了一种以学科知识点本体为知识描述的学习资源语义检索系统。 语义检索学习资源，以减少检索结果的信息冗余并实现高度相关资源的显示。

关键字：本体;语义网;信息检索

Research on Text Information Retrieval Based on Ontology

Jiang Liling
(Department of Information Science Technology, Dalian Maritime University, Dalian 116026)

Abstract: There are a large number of learning resources in today's online education platform, but there are also difficulties in searching for learning resources, redundancy in search results, and lack of semantic support for keyword search. Semantic methods can provide better technology for the areas we are exploring. By introducing the core technology ontology of Semantic Web into the retrieval of learning resources, through ontology analysis, ontology reasoning and conceptual similarity calculation in ontology, a learning resource semantic retrieval system based on subject knowledge point ontology is designed. Semantic retrieval of learning resources to reduce information redundancy of search results and enable the display of highly relevant resources.

Keywords: Ontology; Semantic net; Information retrieval

1 本体概念及相关理论

1.1 本体的概念

本体这个词源于哲学领域，且一直以来存在着许多不同的用法。在计算机界，明确本体的定义经历了一个过程。1993 年，Gruber 提出“本体是概念模型的明确的规范说明”。随后，Borst 提出：“本体是共享概念模型的形式化规范说明”。之后 Studer 等认为本体是共享概念模型的明确的形式化规范说明。这个定义包含 4 层含义：概念化、明确、形式化和共享。“概念”指通过抽象出客观世界中一些现象的相关概念而得到的概念模型；“明确”指所使用的概念及使用这些概念的约束都有明确的定义；“形式化”指本体是计算机可读的；“共享”指本体中体现的是共同认可的知识，反映的是相关领域中公认的概念集[2]。本体形式化后是概念、属性、关系的一组定义，是提供表示领域知识的一个符号集合，表示 了一个特定领域中各知识库间保持不变的领域知识。

1.2 本体的分类

本体基本上可以分为以下几类。可以根据要求选择不同的本体：通用本体，描述最通用的概念，例如空间，时间，事件，动作等，与特定的问题和领域无关，因为公共通信工具可以说是常识。目前，信息检索技术分为三类：全文检索，数据检索和知识检索。本体的基本特征是清晰的概念层次和强大的逻辑推理。在信息检索应用方面，基于本体的信息检索的设计思想如下：首先，在领域专家的帮助下建立相关领域的本体；其次，收集信息资源中的数据，并将收集到的数据以规定的格式存储。在配置数据库中；然后用户提交的信息查询请求，并根据本体将请求解析为规定的数据格式；最后通过语义推理模块对解析后的检索信息进行推理，检索出符合用户需求并满足条件的数据并将结果反馈送给请求者[4]。

现实世界；领域或任务本体，定义或描述特定领域中的相关知识，领域本体就像专家的专业知识一样，每个专业知识记录领域中的事物；应用程序本体，使用属性和关系定义和描述现实世界中依赖于特定领域和主题的知识。这类本体与解决问题的方法相关联[5]。

2 语义网

2.1 语义网的概念

语义网是由万维网联盟的蒂姆·伯纳斯-李在1998年提出的一个概念，它的核心是：通过给万维网上的文档添加能够被计算机所理解的语义，从而使整个互联网成为一个通用的信息交换媒介。语义万维网通过使用标准、置标语言和相关的处理工具来扩展万维网的能力。语义网是由比现今成熟的网际搜索工具更加行之有效的并且自动聚集和搜集信息的文档组成的。其最基本的元素就是语义链接。

2.1 语义网的层次结构

蒂姆·伯纳斯-李在2000年提供出的语义网的层次结构如图1所示[6]。该结构从底层到高层依次为Unicode和URI、XML、RDF和RDF Schema、本体、逻辑、证明和信任。

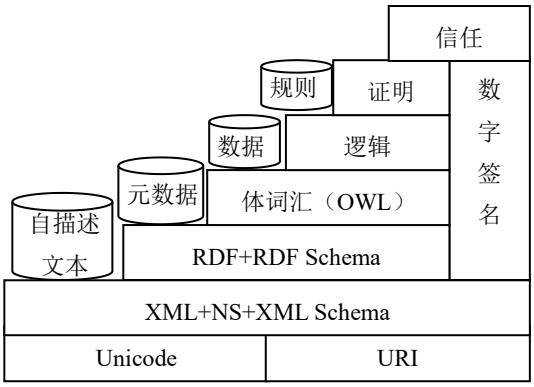

图 1　语义模型

（1）Unicode 和 URI 层。Unicode 和 URI层是整个语义Web的基础,其中Unicode处理资源的编码，保证使用的国际通用的字符集，实现网上信息的统一编码。URL 是 URI 的超集，URI 支持语义网的对象和资源的精细标识，从而使精确信息检索成为可能。

（2）XML+Name Space+XML Schema。XML层具有命名空间和XML Schema定义，通过 XML 标记语言将网上资源信息的结构、内容与数据的表现形式进行分离，确保语义网的定义，并支持与其他基于 XML 的标准进行无缝集成。

（3）RDF+RDF Schema 层。该层用于描述万维网上的资源及其类型，为在线资源描述和用于数据集成的元数据解决方案提供了通用框架。

（4）本体层。该层用于描述各种资源与资源之间的关系。本体揭示了资源本身与资源之间更复杂，更丰富的语义信息，从而分离了信息的结构和内容，并完整地描述了信息。使在线信息具有计算机可理解的语义。

（5）逻辑层。逻辑主要提供公理和推理规则，为智能推理提供基础。

（6）证明层。证明层执行逻辑层生成的规则，并结合信任层的应用机制来判断它是否可以信任给定的证书。

（7）信任层。通过数字签名、证书、基于Agent社区成员间相互推荐等机制和方法来实现 Web 环境中的信任管理。Web 是否能够发挥出最大潜在功能取决于用户是否能够信任 Web 提供的服务和信息。

3 基于本体的学习资源语义检索系统研究

3.1 系统架构

针对当前学习资源检索系统缺乏语义支持和冗余信息的不足，本文设计了基于资源的基于本体的学习资源检索系统。它以主题知识点本体为基础，对系统进行基础知识描述，并在此基础上实现本体知识点本体的推理，分析和相似度计算，为学习资源的语义检索提供语义支持，并通过相似性进行获取和过滤。提供语义检索的学习资源列表。

本体是概念上的规范。对于电子商务安全本体，它包括各种术语，关系和术语的语义。语义处理模块主要进行语义处理并为其他模块提供语义支持，主要包括本体推理、本体解析和相似度计算三个子模块。本体推理是利用学科知识点本体中声明的概念和公理系统以及自创的规则系统来推导出蕴含的关系的过程，是知识完整性的需要。在推导出的本体模型的基础上，就可以进行本体的解析和相似度计算，本体的解析是进行本体中概念的查询、关系的查询和其他一些更为复杂的查询工作，它是本体与其他部件的连接纽带。概念相似度计算模块是领域本体中概念的相似度计算。访问本体可以从访问用户的小对等视图或侧面访问用户描述，例如访问用户类型，行为，状态等。演示不同属性和访问用户属性之间的关系。使用访问用户本体作为访问用户知识的显示模型可以改善业务系统和访问用户之间基于语义的协作，从而实现对用户信息访问的高度共享和重用。

3.2 学习资源语义检索

学习资源语义检索的作用是将学习资源存入学习资源库，提取用户提交表单中有关学习资源描述的信息来形成学习资源的元数据。为了提高查询的准确率，我们将采用基于关键字和语义的综合查询方法。所以首先对资源的描述信息进行分词和词频统计，选取满足条件的词汇作为查询关键字加入到元数据中，对于学习资源的语义标注，其过程如下：

1. 将分词操作后的词汇列表形成词汇向量

将资源的描述信息进行分词和词频统计处理，选取满足词频统计阈值要求的词汇并将其作为查询关关键字加入到学习资源的元数据中，并将其加入词汇向量以进行下一步操作。

2. 学习资源的语义标注

对于词汇中的向量依次确定对象。需要清晰的定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结构是不可预测的，但要探索的问题应 该是有预见的，为了数据挖掘而数据挖掘，是不可能成功的。

3.3 学习资源查询处理模块

学习资源查询处理模块是系统中的主要事务处理模块，搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适合于数据 挖掘应用的数据。研究数据的质量，为进一步的分析做好准备，并确定将要进行的挖掘操作的类型。其具体处理过程如

图 4 所示。

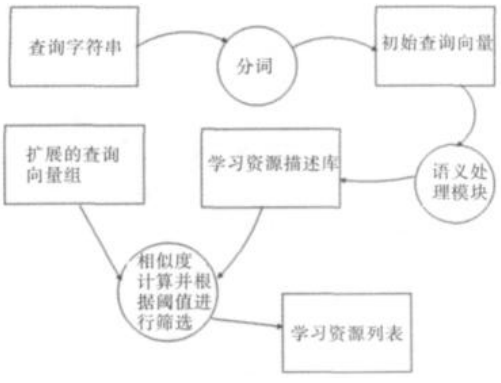

图 4 查询处理模块处理过程图

1. 构造初始查询向量

首先获取用户的查询语句， 并将其进行分词处理，形成初始查询向量。

2. 生成扩展的语义查询向量组

语义处理模块用于执行查询向量中词

汇本体中概念的相似度匹配。 提取满足一定阈值要求的本体中的相关概念，并在原始查询向量中构成查询向量的语义关键单元。词汇表形成新的查询向量，从而形成扩展的语义查询向量组。

3. 查询向量与资源向量之间的相似度计算

将查询向量组和由学习资源元数据中的语义标注部分和查询关键字部分形成的学习资源向量进行相似度运算。

4. 呈现结果列表

最后，将满足阈值要求的学习资源按计算结果由大到小的顺序呈现给用户。

4 实验

以初中物理为例，首先构建初中物理课程知识本体，其步骤为：参考初中物理课本构造概念集合；然后根据概念之间的层次关系构造概念集的层次划分，顶层划分为：知识点、实验、单位、公式等；最后构造概念的数据属性和关系属性，数据属性主要是概念的关键字、重要程度等，关系属性主要包括知识点之间的关系，如光的直线传播是光的折射的先导知识点。构建的本体结构如图 5 所示。

图 5 课程知识本体构建示例图

下面是学习资源语义检索过程的举例：输入查询词，这里为“初中物理中关于光的知识”；经过分词可以得到初始查询向量；初始查询向量经过语义查询扩展处理得到扩展的语义查询向量；最后经过查询向量和资源向量之间的相似度计算得到可用的学习资源列表。处理过程如图 6 所示。

图 6 学习资源语义检索过程示例图

5 结论

语义检索是解决在线教学中学习资源检索中各种问题的有效途径。本文构建了一种基于本体学习资源的语义检索系统，并解释了在网络教学平台上构建语义检索系统的相关技术。除了特定的应用程序外，对于语义查询处理等重要模块，
通过对学习资源注册等方面的详细分析，系统的模块化划分，可以很好地集成到各种网络教学应用平台中进行语义的检索和扩展，并可以利用其本体层进行学习资源的集成和扩展。语义网是 Internet 的未来，其相关研究正在蓬勃发展。
随着网络相关技术的日趋成熟，基于语义网技术的各种在线教育应用平台也将陆续出现，并促进在线教育的进一步发展。

参考文献：

[1] 冯志勇，李文杰，李晓红．本体论工程及其应用[M]．北京：清华大学出版社，2007.5

[2] 宋炜, 张铭. 语义网简明教程[M]. 北京: 高等教育出版社, 2004.

[3] 王洪伟,吴家春,蒋馥.基于本体模型的信息检索机制研究[J].情报学报, 2004,23(1):3-9.

[4]T．R.Gruber．Towards Principles for the Design of Ontologies Used for Knowledge Sharing[J]．International Journal of Human Computer Studies，1995

[5] 张志平, 杨建伟. 语义网及其应用研究综述[J]. 情报学报, 2008, (10) : 721~726.

[6] C.D. Chambers.Enhancement of Visual Selection During Transient Disruption of Parietal Cortex [J].Brain Research, 2006, 1097: 149~155, 2006