《智能信息处理》课程考试

基于本体的教育资源语义检索系统研究

王晓丽

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

2020 年 12 月 6 日

基于本体的教育资源语义检索系统研究

王晓丽

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘　要　随着互联网上教育资源的快速增长，越来越多的用户通过网络进行学习研究。传统的信息检索采用的是基于关键字匹配方式，检索结果准确率较低，因而用户要从海量的资源中检索到需要的资源将变得越来越困难。为提高信息检索的准确率和全面性，在对语义检索技术和领域本体的构建进行充分研究的基础上，本文在网络教育资源的检索工作中应用了本体语义的概念，提出对用户输入的查询条件进行基于本体的查询扩展算法, 建立了基于本体的教育资源语义检索模型。本文模型及方法能有效地提高检索质量，较好地弥补传统关键字检索缺乏语义的不足。

关键词　本体；语义检索；教育资源；语义检索系统；

Research on Educational Resources Semantic Retrieval System Based on Ontology

Wang Xiaoli

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract　With the rapid growth of educational resources on the Internet, more and more users are learning and researching through the Internet. Traditional information retrieval is based on keyword matching, and the retrieval result accuracy is low, so it will become more and more difficult for users to retrieve the required resources from the massive resources. In order to improve the accuracy and overall rate of information retrieval, on the basis of studying the relevant technologies about ontology constructing and semantic retrieval, This paper applies the concept of ontology semantics in the retrieval of network education resources, proposes an ontology-based query extension algorithm for user input query conditions, and establishes an ontology-based semantic retrieval model for education resources. The model and method in this paper can effectively improve the retrieval quality and make up for the lack of semantics in traditional keyword retrieval.

Key words　Ontology; Semantic retrieval; Educational resources; Semantic retrieval system

1 引言

随着我国科学技术以及互联网技术的快速发展，使得更多的用户需要通过网络来获取自己学习所需要的教育资源。网络上的海量信息多种多样，怎样快速并有效的返回给用户其需要的信息成为研究的焦点。目前存在的一些信息检索技术基本上都是基于关键词的，用这种方法检索出的结果中往往含有许多用户不需要的信息，准确率比较低；对于查全率，这种检索方式不对同义词和相关类等进行处理，所以也不能保证较高的查全率。这就要求将检索方式从关键字层面提升到语义层面。将信息和语义结合起来，变成计算机可识别的知识，从而使系统自动识别出用户的需求，并挖掘出与用户查询相关的一些信息或者一些用户可能感兴趣潜在信息，将它们一并返回。

本体是一种可以对概念进行建模的工具，能够在语义层面描述信息，同时也可以很好的描述概念的内涵以及概念与概念之间的关系，具有很好的概念层次，支持逻辑推理，因此在信息检索特别是语义检索领域得到了普遍应用。本文以教育知识本体构建、语义关联度算法研究及语义检索技术应用为基础, 结合教育领域专业特征, 构建了基于本体的教育资源语义检索模型。本文将首先阐述相关的理论及技术， 然后对基于本体的教育资源语义检索系统模型做相关介绍。

2 本体概念及相关理论

2.1 本体的概念

“本体最初起源于哲学领域, 被哲学家用来描述实物的本质. “本体”在中国文化中的基本含义是事物的主体或自身, 事物的来源或根源. 对于本体的定义有多种理解, 在计算机界, 明确本体的定义精力了一个过程. 1993 年， Gruber 给出了本体的一个最为流行的定义, 即“本体是概念模型的明确的规范说明”. 后来, Borst 对此稍作修改, 提出: “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, 其表示的含义独立于具体的环境状态; “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(即能被计算机处理); “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, 它所针对的是团体而不是个体[1]. 该定义具有丰富的概念层次结构及较强的逻辑推理能力, 得到了许多专业领域的认可.

2.2 本体的分类

通常研究领域不同，本体的分类方法也不尽相同，在某特定领域中构建本体的目的能确定本体的类型，进而确定本体具备的功能。依据本体表达的概念模型所在的层次，可对本体做如下分类：知识本体：将知识表示抽象化，将领域概念以及它们之间的关系规范化，重点研究内容是实现语言对知识的表达描述。通用本体：从认识论出发，表达抽象的概念。如初始概念的时间、空间、事物、数量、状态、属性，以及这些初始状态组成的下位概念。这类本体主要用在应用常识知识。领域本体：抽象出特定领域内的知识，对其中的概念和它们之间的关系、在领域内进行的活动、这些活动相关领域的领域知识和原理研究 等进行描述。应用本体：描述与领域内特定任务相关的概念和它们之间的关系。有关任务的描述要依据特定的领域知识本体。图 1 表示各本体间关系。

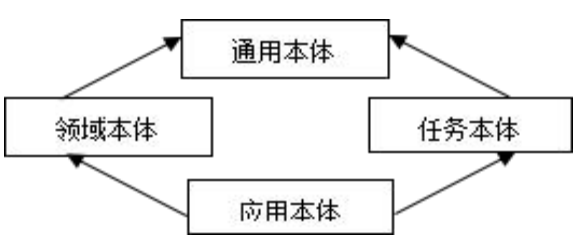

图 1　本体分类及关系图

2.3 本体的分类

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.本体的基本特征是具有清晰的概念层次结构和功能强大的逻辑推理能力.在信息检索应用方面,基于本体的信息检索的设计思想为: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,把收集来的数据按照规定的格式存储在元数据库中;然后用户提交的信息查询请求,并按照本体将该请求转化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].

3 语义网的概念及层次结构

3.1 语义网概念

“语义网”(Semantic Web)这一概念出现在人们的生活中，是受 Web 的创始人 Tim Berners-Lee 的影响，他曾于 1998 年首次提出语义网的基本思想，并对其做了如下定义：语义网首先是一个网，在这个网中有许多描述事物之间关系文档，同时也包含了相关语义信息，使机器可以自动处理。所谓“语义”就是文本的含义，主要用来对网络中文本的内容与结构进行说明，与文本的格式无关。有了语义，网络就可以进行“判断”，因此我们说语义网是一种具有智能的网络，它能理解自然语言并且可以实现人与计算机在思想上的交互。语义网在当前万维网的基础上进行扩展，为网络中的信息加入对应的语义，使计算机可以理解并自动处理。它是如何让计算机理解这些信息的呢？这就是语义网研究的关键，也就是说，我们需要向网络中添加一些知识并使机器理解，这些知识就是我们所说的概念。

3.2 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[2]。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。

（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。

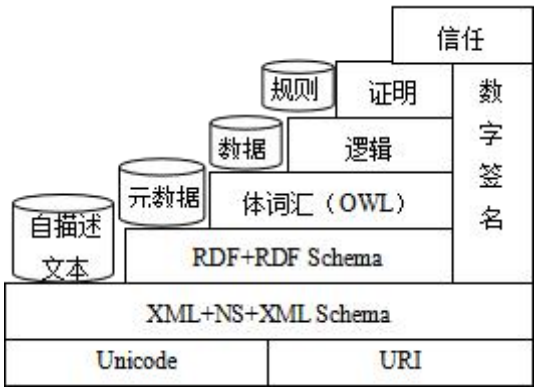

图 2　语义网的层次模型

（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，为网上资源描述提供了一种通用框架和实现数据集成的元数据解决方案。

（4）本体层。该层用于描述各种资源与资源之间的联系，本体揭示了资源本身以及资源之间更为复杂和丰富的语义信息， 从而将信息的结构和内容分离，对信息作完全形式化的描述，使网上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。证明层执行逻辑层产生的规则，并结合信任层的应用机制来评判是否能够信赖给定的证明。

（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。

4 基于本体的教育资源语义检索系统研究

4.1 基于本体的语义检索模型构建

4.1.1 本体建立

本体库的建立是语义检索系统设计的基础，同时也是语义检索系统中最重要的组成单元查询扩展算法设计的基本依据。由于自动生成本体方法有许多不足，因此笔者以面向对象程序设计课程的主要知识点为数据源，用 Protégé 与 OWL (Web Ontology Language) 相结合的方式创建本体[3] 并将领域本体通过Protégé进行编码，作为测试数据蓝本。

4.1.2 教育资源获取

基于本体的语义检索能准确高效实施的首要前提是要有足够可供检索的信息保存于信息库，因此教育资源信息库的建立是整个检索系统设计与实现的基础。在笔者设计的模型中，主要利用网络爬虫从网络上获取各种诸如名词术语、课件、案例、试题、参考资源、网址资源以及其他基础资源的相关教育资源，并最终通过 Lucene[4]对预处理后的网络资源进行索引库的建设和存储。

4.1.3 查询扩展

查询扩展是语义检索区别于关键字检索的重要指标。所谓查询扩展是指以本体库中的本体为依据，将与用户检索条件相关的词条添加于该检索条件，扩展为新的检索关键字，进而以新生成的检索关键字进行检索的方式。这种检索方式可从一定程度上弥补用户提供的查询条件不准确的缺陷，改善待检索信息的检索准确率和全面率。如何利用查询扩展词，用合适的条件去扩展查询，是实现查询扩展的核心问题。基于本体的教育资源语义检索模型是对查询条件的语义扩展。在扩展查询之前分为以下几个步骤:第一，本体库的建立是查询扩展环节进行的基础;第二，将本体库中的信息资源中概念间的关联程度进行量化；第三，根据概念量化的结果来确定查询的扩展范围。在扩展查询阶段，通过推理逻辑关系，将建立好的领域本体库进行语义推理，以对查询条件进行语义扩展。

4.1.4 检索引擎与结果排序

在语义检索的系统构建中，检索引擎在其中发挥着非常关键的作用。Lucene 检索引擎作为源代码完全开放的全文检索引擎工具包，又是软件基金会的子项目，不管是从哪种角度出发该检索引擎都是最佳选择。Lucene 检索引擎应用了 Java 语言已实现的开源软件项目，具有很好的可操作性、可拓展性以及实用性。该检索引擎通过屏蔽检索环节中复杂的检索过程，使得在检索系统构建中比较倾向于建立资源的业务领域。与此同时，Lucene 提供的是一些高效的功能，例如，将检索条件按照语义相关度进行排序、生成文档引擎库等多项功能。通过以上四方面的论述可以得出基于本体的语义检索模

型结构, 如图 3 所示。

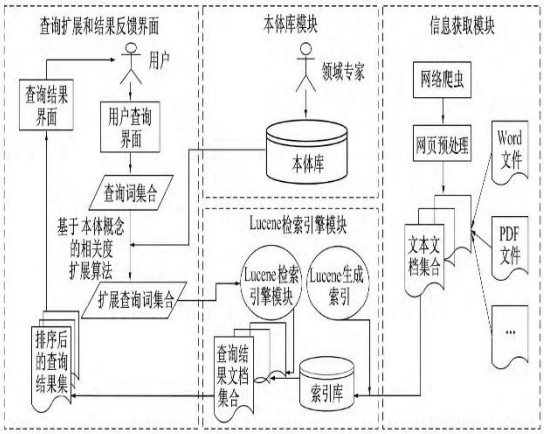

图 3 基于本体的语义检索模型结构

4.2 基于本体的教育资源语义检索系统的设计

网络爬虫构件、语义推理引擎构件和检索引擎构件是组成语义检索系统的 3 个核心构件。

4.2.1 网络爬虫构件

服务器端和客户端共同组成了该系统的网络爬虫构件。服务器端保存和管理爬虫的配置文件, 通过对配置文件进行解析调度客户端。客户端主要通过响应接口完成对网络资源的抓取工作。

4.2.2 语义推理引擎

语义检索系统中, 用户输入的查询条件会交由语义推理引擎, 语义推理引擎负责将这些查询条件与本体库中的领域本体进行概念匹配, 进而将匹配到的概念与原始查询条件结合并进行查询扩展, 最终, 将扩展后得到的新查询条件交给检索引擎进行更准确的查询检索工作。对用户输入的检索条件进行基于本体的、高效准确的语义查询扩展的重要前提和核心保障是设计科学合理的语义关联度计算方法, 下面论述该模型所采用的语义关联度计算方法。

（1） 语义相似度算法。语义相似度是指两概念在语义上的一致程度, 目前, 研究人员主要采用综合相似度的计算方法[5]基于领域本体概念结构图的本体概念语义相似度计算主要受以下因素影响: 语义距离、两个概念的最近共同父节点深度和语义重合度等[6]。其中, 基于语义距离的计算方法如下：

$$S_{Simdege}(C_x,C_y)=\frac{2L_{max}-L(C_x,C_y)}{L_{max}} \quad (1)$$

其中L_{max}为本体结构图中从头结点到其他所有节点的路径最大值, $L(C_x,C_y)$为两个本体概念间的最小距离。

由于本体概念结构图对基于距离的本体语义相似度计算结果有很大影响, 因此, 为弥补基于距离相似度计算方法的局限性, 在计算相似度时, 结合了另一种影响语义相似度的因素—两个概念的最近共同父节点深度计算语义相似度。计算过程如下。

$$S_{Simlin}(C_x,C_y)=\frac{2log\,P_{MRCA}(C_x,C_y)}{\log P_{C_x}+\log P_{C_y}} \quad (2)$$

其中C_x是本体 C 中概念 x 的实例数, 概念 x 和概念 y 的最近公共祖节点用$P_{MRCA}(C_x,C_y)$ 表示, P_{C_x}和P_{C_y}为概念 x 和概念 y 的祖节点。

笔者提出领域本体中任意两个概念间的相似度计算公式为

$$S_{Sim}(C_x,C_y)=\alpha S_{imedge}(C_x,C_y)+\beta S_{Simlin}(C_x,C_y) \quad (3)$$

其中 α 和 β 为调节参数, 且 $\alpha+\beta$ =1; 相似度 $S_{Sim}(C_x,C_y)\in[0,1]$。

（2） 语义相关度算法。语义相关度是指两个概念在语义上的相关联程度, 因此, 计算本体概念的语义关联程度除受语义相似度的影响外, 在一定程度上还受两个概念语义相关度的影响。本文提出的语义相关度计算方法为:在本体推理的基础上, 获取与参照概念相关的概念集合, 然后基于此集合进行语义相关度计算。同时应注意, 有些概念之间的相关度会随传递减弱。综合上述因素, 本系统所采用的语义相关度

$$R_{el}(C_x,C_y)=\prod_{j=1}^{n}R_{el}(e_j) \quad （4）$$

其中$R_{el}(e_j)$表示两个概念相连关系边 e 所对应的语义相关度权值。该权值由专家根据概念的关系类型分别给出。

（3）语义关联度算法。对于不同的两个本体概念之间的概念关系以及所蕴涵的关系, 主要由两个概念之间的语义相关度以及语义相似度所决定。因此, 在语义检索系统中, 可以将语义相关度与语义相似度结合起来, 通过加权将语义关联度表示的不同概念之间的关联程度计算出来, 其计算公式可以表示为:

$$S_{re}(C_x,C_y)=\mu S_{im}(C_x,C_y)+\lambda R_{el}(C_x,C_y) \quad （5）$$

在这个计算公式中, 其中的$S_{re}(C_x,C_y)$ 表示的是两个不同概念之间的语义关联度; $R_{el}(C_x,C_y)$ 表示的是两个不同概念之间的语义相关度; $S_{im}(C_x,C_y)$ 表示的是不同概念之间的语义相似度, 其中的μ和λ分别表示的是调节参数且两者相加的和等于 1, 调节

参数的设定一般是由领域专家直接给出的。同时需要注意的是语义相似度的范围应该在 0~1 之间。

4.2.3 检索引擎

检索引擎中的解析器组件首先负责获取由网络爬虫从网络上收集到的各种教育资源，由检索引擎的索引构建器对解析器获得的资源进行解析处理并建立查询索引库，最后由检索引擎中的查询器结合语义推理引擎提供的扩展查询集合完成最终的资源检索以及检索结果的客户端回显工作。

结束语

语义检索是解决当今网络教学资源检索中存在的各种问题的一个有效的途径，总而言之，检索者在输入查询条件之后，系统会根据查询的条件进行扩展查询计算，并将算法结合检索信息的语义应用到基于本体的教育资源语义检索系统中，正是这种有效的检索方式实现了将 Lucene 对扩展后查询条件进行检索模型的设计。从专家做出的实验中可以得出，基于本体的教育资源语义检索模型的构建，在综合查准率以及综合查全率中分别可以达到 83%和 81%，在这两项指标中相较于传统的关键字查询方式具有很大的提升。语义网是互联网的未来，同时它的相关研究也正蓬勃发展，相信随着语义网相关技术的成熟，各种基于语义网技术的网络教育应用平台也将会陆续出现，并推动网络教育的进一步发展。

参 考 文 献

[1]黄以宝.基于本体的教育资源语义检索系统的实现探讨[J].信息与电脑(理论版),2018(18):182-184.

[2]许鑫,杨佳颖.国外语义网研究现状与动向——基于 2002—2018 年 ISWC会议[J].情报学报,2020,39(07):761-776.

[3]吴国祥,谢大同.国内基于本体的网络教育资源建设研究综述[J].情报探索,2018(10):130-134.

[4]于超,王璐,程道文.基于本体的教育资源语义检索系统研究[J].吉林大学学报(信息科学版),2018,36(02):207-212.

[5] Anna Formica, Elaheh Pourabbas, Francesco Taglino. Semantic Search Enhanced with Rating Scores. 2020, 12(4).

[6]李京杰.语义网在教育资源领域的应用研究[J].中国教育技术装备,2019(04):5-8.