《智能信息处理》课程考试

基于本体的语义检索技术研究

田鑫

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

2020 年 12 月 7 日

基于本体的语义检索技术研究

田鑫[1)]

[1)]（大连海事大学 信息科学技术学院， 大连 116026）

摘　要　随着互联网上信息数量的激增， 传统的基于关键字的信息检索技术已无法满足用户的需要。本体作为一种能在语义和知识层次上描述信息系统的概念模型建模工具，能够提高检索的准确率，缩小搜索范围和减少冗余的返回结果。文章介绍了本体的概念及其在语义检索领域中的作用，在此基础上提出了一种基于本体的语义检索模型，并对该模型的检索思想、体系结构、检索机制及功能特点进行了阐述。

关键词　本体；语义检索

Research on Semantic Retrieval Technology Based on Ontology

TIAN Xin[1)]

[1)](School of Information Science and Technology， Dalian Maritime University， Dalian 116026)

Abstract　With the rapid increase in the amount of information on the Internet， the traditional keyword-based information retrieval technology has been unable to meet the needs of users. As a conceptual model modeling tool that can describe information systems at the semantic and knowledge levels， ontology can improve the accuracy of retrieval， narrow the search scope and reduce redundant return results. The article introduces the concept of ontology and its role in the field of semantic retrieval. On this basis， a semantic retrieval model based on ontology is proposed， and the retrieval ideas， architecture， retrieval mechanism and functional characteristics of the model are explained.

Key words　Ontology; Semantic retrieval

0　引言

随着全球网络化、信息化的发展，网络上的信息越来越多，对信息检索手段的有效性要求也越来越高。如何从海量的信息中高效、快速、准确地检索到所需信息已经成为计算机领域研究的一个热点问题。目前检索技术主要是基于关键字的全文匹配检索技术或者是按主题进行分类的检索技术。它们的结果往往会返回大量无关信息，这样用户需要花时间排除无关信息，才能找到真正想要的信息；而且传统的信息检索寻找的信息可能仅仅是字面本身的信息，但我们想要的是这个信息的概念及其相关的成分，而不仅仅是字面所表达的信息。本体作为一种能够在语义和知识层次上描述信息系统的概念模型建模工具，具有良好的概念层次结构和对逻辑推理的支持，可以在用户提问检索式构造过程中增加语义制导， 赋予检索式语义表达功能，便于检索结果的共享和重用。从而使信息检索从目前基于关键字的层面提高到基于知识的层面，提高了信息检索的查准率和查全率。本文首先概述了信息检索的分类、本体的概念以及本体在信息检索中的作用，然后对基于本体的检索思想、语义检索系统的体系结构、检索机制以及功能进行了分析。

1　本体概念及相关理论

1.1　本体的概念

本体[1]（Ontology）的概念最初起源于哲学领域，它在哲学中的定义为 “对世界上客观存在物的系统地描述，即存在论”，是客观存在的一个系统的解释或说明，关心的是客观现实的抽象本质。

后来随着计算机科学技术的发展， 人工智能学者把本体这个概念应用到了人工智能领域。在人工智能领域，关于本体的概念最为流行的一种定义是 Studer 等人提出来的：本体是共享概念模型的明确的形式化规范说明。这说明本体的概念包含四层含义:概念模型、明确、形式化和共享。“概念模型”指通过抽象出客观世界中一些现象的相关概念而得到的模型。概念模型所表现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，即本体针对的是团体而非个体的共识。

从知识共享的角度来说，本体是对客观存在的概念和关系的描述。它是通用意义上的概念定义集，是关于概念和关系的词汇表。本体的目标是捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇(术语)和词汇间相互关系的明确定义。所以，本体能够将领域中的各种概念及概念之间的关系显示地、形式化表达出来，从而将术语的语义表达出来，因而在语义查询方面发挥着重要作用。

1.2 本体的分类

本体基本可分为以下几类，可以根据需求选用不同的知识本体：通用本体，描述最一般化的概念，例如空间、时间、事件、行动等，独立于特定的问题与领域，作为大众沟通的工具，可以说是真实世界中的常识；领域或任务本体，定义或描述特殊领域的相关知识，领域本体如同专家的专门知识，每一份专业知识都记载该领域中的事物；应用本体，使用属性、关系进行定义与描述真实世界中既依赖于某个特定领域又依赖于某项课题的知识。这类本体与解决问题的方法相关联。图 1 表示各本体间关系。

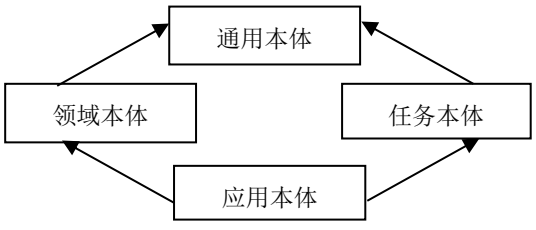

图 1 本体分类及关系图

1.3 语义检索

当前信息检索技术主要有全文检索、数据检索、知识检索三大类，其描述列举如表 1 所示：

表 1

分类	特点	缺点
全文检索（Text retrieval）	把用户的查询请求和全文中的每一个词进行比较，不考虑查询请求和文件语义上的匹配。	虽然可以保证查全率，但是差准率大大降低。
数据检索（Data retrieval）	查询要求和信息系统中的数据都遵循一定的格式，具有一定的结构，允许对特定字段检索（例如：作者=“张三”）。需要有标示字段的方法。	性能取决于所使用的字段标示方法和用户对方法的理解，具有很大的局限性，支持语义匹配的能力较差。
知识检索（Knowledge retrieval）	基于知识的、语义上的匹配，在查准率和查全率上有更好的保证。是信息检索的重点，特别是面向 Web 信息的知识检索的重点。	

基于知识的检索即语义检索，是把信息检索与人工智能技术、自然语言技术相结合的检索，它从语义理解的角度分析信息对象与检索者的检索请求，是一种基于概念及其相关关系的检索匹配机制。语义检索之所以能够提供比全文检索更为智能化、知识化的服务，其根本在于拥有比全文检索更为丰富的知识描述空间，即概念空间。到目前为止，关于概念空间并没有明确的定义，但综观各类资料可以将概念空间理解为一种集语义关系于一体的、用于计算机识别操作的概念集合，它是语义检索研究的关键。

1.4 本体在语义检索中的作用

本体作为一种能在语义和知识层次上描述信息系统的概念模型建模工具，它在语义检索中的作用可概括为以下几点：

(1) 本体为语义标注和扩展提供了标准的词汇库；

(2) 检索中所进行的推理工作必须在本体中进行；

(3) 本体可以明确领域假设，使领域公理得到明确描述而达成共识。

2 语义网的概念及层次结构

2.1 语义网的概念

语义网[1]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。[4]

2.2 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[4]。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。

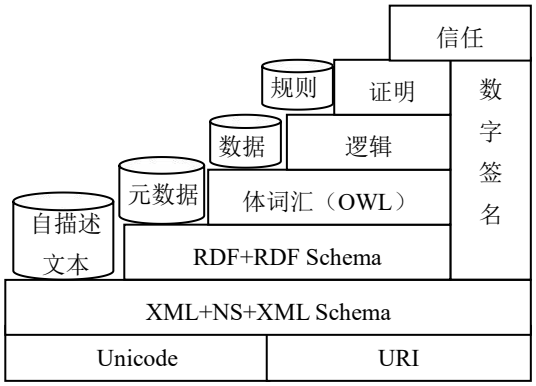

图 2 语义网的层次模型

（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。

3 基于本体的语义检索系统

3.1 基于本体的语义检测的基本思想

学者们一般将基于本体的语义相似度计算方法划分为4类：

由于 Ontology 具有良好的概念层次结构和对逻辑推理的支持， 因此在信息检索， 特别是在基于知识的检索中得到了广泛的应用。

基于 Ontology 的信息检索的基本思想[2]可归纳如下：

（1）在领域专家的帮助下，建立相关领域的Ontology 。

（2）利用本体中的概念来标引相关的信息资源并以特定的格式存储，标引的过程与传统的方法类似，可以用手工标引，也可以采取自动和半自动的方式，而结果通常是以 RDF 文档的方式以特定的格式存储。

（3）对 RDF 、RDFS 、OWL 等相关文件的解析和推理。其目的是为了将以一般文件存储的本体和信息资源信息从文件中读取出来存储在特定的模型中以便于程序处理，并可以根据一定的推理规则基于本体进行语义推理。

（4）对用户检索界面获取的查询请求，查询

转换器按照 Ontology 将查询请求转换成规定的格式，在 Ontology 的帮助下从元数据库中匹配出符合条件的数据集合。

3.2 基于本体的语义检索模型

基于本体的语义检索模型[3]结构包括人机交互层、信息处理层、信息收集和存储层 3 部分组成。人机交互层包括用户、 查询界面、查询请求是用户和系统通讯的接口，用户的查询请求从这里发送给系统，系统的检索结果也从这里传递给用户。信息处理层主要包括查询处理、语义分析、语义查询、本体推理引擎等模块。信息处理层则是根据领域本体库，把用户的查询请求转换成查询本体，语义分析、查询模块完成概念相关度分析，本体推理引擎模块完成对标注信息的语义闭包求解。信息收集和存储层包括领域本体库、元数据库、元数据基 本信息库、资源库（关系数据库、文档文件、XML 文件、其他格式文件等）。信息收集和储存层则是构建领域本体库，同时利用领域本体库的构建信息资源(包括各种结构化、半结构化、非结构化信息资源等)的本体描述以及元数据库，实现对信息的形式化表示和存储。

3.3 模型的检索机制

（1）在领域专家的帮助下，建立基于领域概念的领域本体库，同时收集数据资源对象，用本体描述语言来描述，并参照已建立的领域本体，把收集来的数据资源对象按规定的格式存储在元数据库中。机器可以理解的带有语义信息的元数据是借助语义标引工具，按照领域本体的概念及关联，对资源对象进行概念分析、分类、标引、描述和处理而形成的。

（2）用户进行查询时，需要通过查询界面，在本体引导下最大限度地表达查询需求，构造查询本体。

（3）根据已经构造的领域本体及查询本体，通过语义逻辑推理模块进行相似度的语义推理，实现检索，获得元数据库中匹配的本体列表，最后再将检索的结果返回给用户。

3.4 模型的功能特点

（1）实现了基于领域本体的检索，改善原有单一关键词检索的语义缺乏问题。

（2）本体库中蕴含了大量的知识和推理规则，能够实现语义推理。

4 总结

本文针对传统的基于关键字的信息检索中缺乏知识表示和语义处理能力的缺陷，提出了一个基于本体的语义检索模型。相对于传统的检索而言，语义检索将用户的输入转换为系统所能认知的知识，因而使得查准率和查全率也大大提高。作为新一代检索技术，语义检索是信息检索目前发展的趋势，而且具有较大的应用价值，如何自动建立本体，处理复杂查询将是下一步研究的重点。

参 考 文 献

[1]李文靖，胡书山，余日季. 基于语义网的数字化家具模型本体设计与检索[J]. 软件导刊，2019，18(08):136-139+143.

[2]王继东，张瑜，李娜.基于本体的语义检索技术研究与实现[J].计算机技术与发展，2017，19(10):134-137.

[3]张继芳. 基于本体的语义检索技术研究[J].科技信息，2011，(10):90-91.

[4]黄 敏， 赖茂生.语义检索研究综述[J] .图书情报工作， 2008，52(6) :63-66.