《智能信息处理》课程考试

基于本体的 WEB 语义检索

赵茂昌

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

2020 年 12 月 8 日

基于本体的 WEB 语义检索

赵茂昌
（大连海事大学　信息科学与技术学院　大连　116026）

摘　要　随着信息时代的到来，计算机技术取得了突飞猛进的发展，与之密切相关的检索系统也得到了进一步地开发和发展，传统的关键字搜索方法缺乏对搜索内容在语义上的处理，搜索出的结果可能会出现不全面、不准确的问题，从而导致搜索结果无法达到用户的要求。基于本体的 Web 语义检索系统是检索系统的一个十分重要的分支，本文在现有的语义检索的基础上，以本体为依据，对语义检索进行简单分析与研究。

关键词　本体;语义检索;WEB; 检索系统

Web Semantic Retrieval Based on Ontology

Maochang Zhao

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

Abstract　With the advent of the information age, computer technology has made rapid development, and the retrieval system closely related to it has been further developed and developed. The traditional keyword search methods lack of semantic processing of the search content, and the search results may appear incomplete and inaccurate problems, resulting in the search results unable to meet the requirements of users. Ontology based Web Semantic retrieval system is a very important branch of retrieval system. Based on the existing semantic retrieval, this paper analyzes and studies the semantic retrieval based on ontology.

Keywords　ontology；semantic retrieval；web；retrieval system

1　引　言

万维网由于其丰富的信息内容和方便的访问方式，已经成为一个能够自动处理信息和动态显示数据的多功能平台。然而，目前的搜索引擎多采用字符匹配算法。用户在搜索的时候，会有很多无用的信息。传统的文本信息检索一般使用查全率(Recall)与查准率(Precision)来对检索效果进行量化评价，但是在信息海量的互联网上，信息检索用查全率与查准率来衡量检索效果不太合适。针对目前检索的不足，Tim Berners Lee 提出了语义网的概念[1]。它扩展了现有的万维网，将信息嵌入到机器可读、表示某种知识的注释中，使计算机和用户能够理解，使计算机和用户能够协同工作。这一理论为解决目前遇到的问题提供了一条新的途径。

本体不仅能表示层次化的知识结构，还可以表示各种复杂的关系，同时经过推理还可以表示隐含的各种数据之间的关系。这样的表示方法有利于数据的有效整合。对于 Web 搜索来说，通过本体的定义，搜索程序可以进行基于语义的精确搜索而不是模糊的关键词搜索．通过本体把页面上的信息与某些知识结构和规则链接起来，对用户检索进行扩展、推理。搜索引擎就可以进行语义级别的 Web 分析和信息抽取。从而提高系统的查准率与查全率。因此。使用本体技术改进信息检索，是实现知识检索和语义检索的关键[2]。

2 本体概念及相关理论

2.1 本体的概念

本体论最先起源于哲学领域。柏拉图认为本体就是理念,康德认为的本体是“自在之物”。本体可以被看作是一个客观存在的元系统,用于解释或说明。对于本体的定义有多种理解,在计算机界,明确本体的定义经历了一个过程。1993 年，Gruber 给出的本体定义为:Ontology 是概念模型的明确的规范说明[3]。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。Studer 在 Gruber 基础上提出“Ontology 是共享概念模型的明确的形式化规范说明”,其中包括四个层面:概念模型(conceptualization)、明确(explicit)、形式化(formal)和共享(share)[4]。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 本体的组成

Perez 等人根据分类对本体进行组织，总结出本体论的五个基本要素,即本体是由概念、关系、函数、公理和实例组成的，本体更适合于表达整体内容;而语义网络应用范围更宽泛且多用于专家系统知识表示，同时不需要专家参与其建模,语义网更适合于表达具有属性的断言。概念反映出现实对象的基本属性，包括实体本身的功能、事物之间的关联以及时间的行为推理关系等；关系是概念之间相互作用的定义和表达。判断事物之间关系的公理，是判断事物之间关系是否正确的基础；公理是判断事物之间关系的基础。实例表示具体的对象和概念的表示。

2.3 本体在语义检索中的应用

目前，本体已广泛运用于知识管理、知识检索、个性化服务等。尤其是在语义检索中，本体的引入能有效解决传统检索存在的检索效率低下、缺乏语义推理能力和无法实现智能检索等方面的不足，为提升各领域语义检索的智能化、精准度和召回率提供有效的技术方案，也为语义检索在智能化、精准度和召回率等方面带来新变革。基于本体的信息检索的设计思想是：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指定的格式存储在元数据数据库中。然后根据本体将用户提交的信息查询请求转换为指定的数据格式。最后语义推理模块对解析后的检索信息进行推理，检索出满足用户需求和条件的数据，并将结果返回给请求者[5]。

3 基于本体的语义检索基本思想

早在 1994 年，Voorhees 就提出了基于本体的查询扩展。2004 年，Navigli 提出了一种基于本体注释的查询扩展方法，利用 WorldNet 中的概念并对注释进行了扩展。经过不断的开发和探索，研究人员意识到为了更好地利用语义，应该充分利用本体中的属性等关系，最大限度地提取文档和用户查询中包含的有价值信息，提出了一种基于本体的语义检索模型。

3.1 语义网的层次结构

(1)Tim Bemers-Lee 在 2000 年提供出了语义网的层次结构。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。Unicode 和 URI 层.

(1) Unicode 和 URI 层是整个语义网的基础,Unicode 处理资源的编码,保证国际通用字符集的使用，实现互联网上信息的统一编码。

(2)XML+Name Space+XML Schema.

XML 层具有名称空间和 XML 模式定义，通过 XML 标记语言（XML）将联机资源信息的结构、内容和表示分离开来，确保语义 Web 的定义，并支持与其他基于 XML 的标准的无缝集成。

(3)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。

(4)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。

(5)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。

(6)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。

本体从语义网的分析结构出发，揭示了资源本身和资源之间的语义信息，从而将信息的结构和内容分离开来，使信息得到充分的形式化描述，使联机信息具有计算机可理解的语义。然而，传统信息检索技术的缺点在于语义理解有限，查全率和查准率较低。因此，有必要引入本体作为构建语义层的核心组件。本体的语义层主要是对资源和用户需求的语义信息的形式化描述，以及不同层次之间的语义聚合。这样，用户就可以利用鼓动来扩展和扩展关键字搜索，实现搜索的自动化和智能化。

3.2 信息检索的定义及基本原理

信息检索是指以一定的方式组织和存储信息，并根据用户的需要查找相关信息的过程。它包括两个部分：（1）存储：将大量分散无序的信息集中在一起，经过处理和排序，使其有序化、系统化，成为一个可以进行查询的信息集；（2）检索：通过查询语言，从集中搜索出所需的信息。这是广义的信息检索，狭义的信息检索仅指第（2）部分，即从信息集中找到所需信息的过程。

从本质上讲，信息检索是信息集与需求集的匹配与选择。为了实现匹配和选择，首先要对信息集进行特征化，即对信息集进行人工或计算机处理，并将原始的隐含特征和不可识别特征明确化。这种处理称为内容分析和索引，其中用于表示文档特征的术语称为索引词。另一方面，在检索过程中，也要分析用户的信息需求，并提取出概念或属性，并使用与索引过程系统相同的标识系统（检索语言）来表达包含在概念和属性中的需求，然后通过匹配和选择机制，进行相似度比较这套要求和信息收集，最终按照一定的标准符合信息的需要。

3.3 基于本体的语义检索流程概述

与传统的关键字检索相比，基于本体的语义检索具有反映语义信息、准确表达用户查询意图的优点。为了更好的体现语义检索过程中的语义。查询扩展是基础，因为语义检索的前提是充分理解用户提出的查询请求，即语义处理用户检索请求。查询扩展是用户查询处理中的一个重要步骤，主要是指如何为用户输入的检索请求分配语义。利用本体论的知识、推理机制和简单的自然语言处理技术来处理用户提出的问题，分析检索输入的类型和目标，利用本体的语义关系（如义、上下关系等）和推理机制进行查询语义扩展，从而更准确地了解用户的查询需求，提高查询效率。

3.4 语义扩展概述

分词后，用户输入的查询最终转换为包含单个关键字和多个关键字，其中多个关键字占主导地位。它们用于描述用户的查询意图，通常包含检索到的对象的关键字和关键属性。对于单关键字查询，就是选择合适的领域本体对查询词进行处理。它是将转换后的查询词与本体库中的概念、属性、关系和实例进行匹配，通过扩展形成新的查询词。有两种情况需要处理：一种是查询词是本体中的一个概念；另一种是查询词在本体中不存在，只是一般的词。对于更多的关键字组合查询，它是通过本体库对用户的查询进行规范化，规范化了词的概念，同时，根据用户输入的概念、属性，或者作为实例，利用本体的语义关系，推理出相关的语义信息，并基于信息内容检索返回相关知识。

4 总 结

数据作为信息资源不断积累，使得互联网已经成为一个杂乱无章的信息仓库，这给用户检索信息带来了不便，用户使用的方法也有很大的局限性，目前基于本体的信息检索技术不仅局限于关键字匹配，还将能够实现处理与语义相关的查询，为提高检索效率和深度提供了一种新的方法，为构建基于多本体语义的检索系统奠定了基础。如何有效地构建一个能够快速嵌入到特定领域的通用本体，如何利用本体快速检索出具有真实价值的可重用和再创造的隐藏知识，而如何构建一个满足用户需求的系统将成为未来该领域实际应用的主要方向。

参 考 文 献

[1] Berners-Lee T, Hendler J. The Semantic Web. Scientific American, 2001,258(5):34-37.

[2]. 陈振标，基于本体的语义检索技术研究 福州大学，情报探索 2011

[3] Thomas R,Gruber.A Translation Approach to Portable Ontology Specifications[J].Knowledge Acquisition,1993(6):199-220.

[4]裴培,丁雪晶.基于本体的语义相似度计算综述[J].合肥学院学报(综合版),2020,37(05):68-74.

[5]张婷,段跃兴,张月琴.基于旅游领域本体的语义检索模型[J] 太原理工大学学报,2020,51(02):220-225.