

《智能信息处理》课程考试

## 基于本体的信息检索分析

田坤

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

2021 年 12 月 16 日

# 基于本体的信息检索分析

田坤

(大连海事大学 信息科学技术学院, 大连 116026)

**摘要** 随着万维网的进一步普及,人们越来越能方便的获取信息,但是随着网页上信息的爆炸式增长,缺乏语义描述的信息会导致信息检索时结果的遗漏以及大量冗余。本文主要讨论了语义网、本体以及信息检索,并对基于本体的语义检索系统进行了介绍。

**关键词** 本体; 信息检索; 语义网;

## Base on the Ontology of Text Similarity Calculation Analysis

Tian Kun<sup>1)</sup>

<sup>1)</sup>( School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

**Abstract** With the further popularization of the world wide web, people can get more and more convenient information. However, with the explosive growth of information on the web page, the lack of semantic description of information will lead to the omission of information retrieval results and a large number of redundancy. This paper mainly discusses semantic web, ontology and information retrieval, and introduces the semantic retrieval system based on ontology.

**Key words** Ontology; information retrieval; semantic web;

## 1 引言

如今的生活,当浏览网页时,人们通常使用的是万维网。万维网给全球的交流和传播带来了革命性的变化,大大方便了人们进行资源共享与信息交流。但是每一天服务器通过网页发布的信息是海量的,这些信息中,通常有格式的描述,也就是信息资源,但是却没有语义的描述,也就是信息缺乏元数据。以一本书为例,元数据是关键信息点,即书的作者、写书的时间、书的主题等关键信息,信息资源则是书的内容。元数据用来描述信息资源,可以增强各种资源之间的可交换性,提供资源的可访问性,也可以沟通不同的数据格式。万维网的检索方式是关键字检索,在检索时,万维网也只能进行语法匹配,而不能进行语义匹配,导致检索结果存在大量的冗余信息,也会存在很多遗漏的信息。这主要是由于万维网的格式是显式的而语义是隐式的,导致机器无法理解语义,不利于信息处理自动

化,检索效果也比较差。

为了解决以上的问题,语义网应运而生。语义网认为,因特网上的信息应该有助于更有效的搜索、访问和应用。语义网提供计算机对互联网信息的可理解意义,满足 AGENT 的异构分布信息的有效检索和访问,实现网络信息资源在语义层的全面互联,实现互联网信息的实现了更高层次、基于知识的智能应用。语义网络采用语义搜索的方式,其形式是单独的,意义是明确的,机器能够理解语义,信息处理可以自动化。

## 2 语义网的概念及体系结构

### 2.1 语义网的概念

语义网的概念是由万维网联盟的蒂姆·伯纳斯-李在 1998 年提出的。语义网就是能够根据语义进行判断的智能网络,实现人与电脑之间的无障碍沟通。它好比一个巨型的大脑,智能化程度极高,协调能力非常强大。在语义网上连接的每一部电脑不

但能够理解词语和概念，而且还能够理解它们之间的逻辑关系，可以干人所从事的工作。它将使人类从搜索相关网页的繁重劳动中解放出来，把用户变成全能的上帝。语义网中的计算机能利用自己的智能软件，在万维网上的海量资源中找到你所需要的信息。语义网络的核心是通过在多维网络的文档中添加计算机能够理解的意义“元数据”，整个因特网成为共同的信息交换介质。

## 2.2 语义网的体系结构

蒂姆·伯纳斯-李在 2000 年提供出的语义网的体系结构如图 1 所示<sup>[1]</sup>。该体系结构共有七层，从底层到高层依次为 Unicode/URI、XML、RDF、本体、逻辑、证明以及信任。

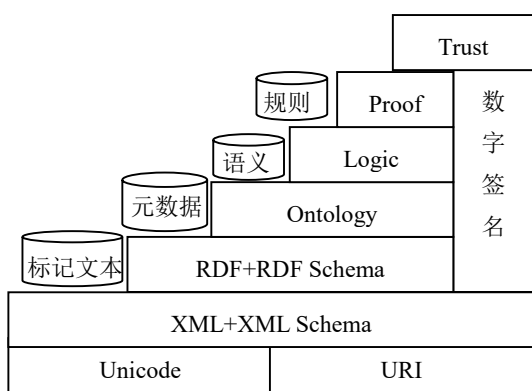


图 1 语义网的体系结构

(1) Unicode/URI 层，也就是“字符集”层。这一层是整个语义网体系结构的基础，其中 Unicode 负责处理资源的编码，URI 负责资源的标识。Unicode 是一个字符集，它基本上包括了世界上所有语言的字符。URI 即统一资源定位符，用于唯一标识网络上的一个概念或资源。

(2) XML 层，也就是根标记语言层。该层是语义网络架构的重要组成部分，该层在语法上表示数据的内容和结构，并且使用标准语言来分离网络信息的表示形式、数据结构和内容。XML (Extensible Markup Language) 是一种简单的、开放的可拓展的标记语言，由文档头、文档体和文档尾构成。XML 支持用户自定义标记，支持异构数据的交换，支持多源数据的集成，支持以统一标准定义字描述数据，并且 XML 的信息内容与信息显示分离。

(3) RDF 层，即“资源描述框架”层。将该信息描述为可用于网络的信息和可用于该资源层的信息。提供了用于实现因特网资源描述的通用框架和数据集成的元数据解决方案。

(4) 本体层。该层是基于 RDF 层定义的概念及其关系的抽象描述，描述应用领域的知识，描述各种资源与资源的关系，实现词汇表的扩展。主体层分离信息的结构和内容，进行信息完全形式化的说明，使因特网信息具有计算机能够理解的意义。在该层中，用户不仅可以定义概念，还可以定义概念之间的丰富关系。

(5) 逻辑层。这一层主要负责提供公理和推理规则，为智能推理提供基础。

(6) 证明层。这一层负责执行逻辑层提供的公理和推理规则，通过逻辑推理对资源、资源之间的关系以及推理结果进行验证，证明其有效性。

(7) 信任层。这一层通过数字签名交换以及数字签名，建立一定的信任关系，从而证明语义网输出的可靠性以及其是否符合用户的要求。

## 3 本体的概念及相关理论

### 3.1 本体的概念

关于本体的定义有许多不同的解释。本体这一概念最早来源于哲学领域。在哲学领域，本体是“存在及其本质和规律”，本体就是底层的、起支撑作用的东西。之后，本体这一概念被引入到计算机领域人工智能方向。Neches 等人首先给出了本体的定义，即“构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则”。之后，越来越多的人开始研究本体，越来越多的人也给出了本体的新的定义。在 1993 年，Gruber 提出本体是“概念模型的明确的规范说明”。这一定义得到了人们的广泛认可<sup>[2]</sup>。后人在这基础上不断对这一定义进行完善。Borst 提出了本体是“共享概念模型的形式化规范说明”。强调了共享概念模型，提出了本体是形式化规范说明。Studer 结合前人的观点，进一步给出本体新定义，即本体是“共享概念模型的明确的形式化规范说明”。

将本体的定义拆解开，可以看到，概念模型是领域的主要概念，概念模型指的是语词及其相互关系；共享指的是领域内共识；明确指的是概念有明确的定义，概念是反映对象的特有属性的思维形式，概念是从对象的属性中抽出特有属性概括而成，概念的语言表达形式是词或词组，概念又内涵和外延；而形式化指的是计算机可读。

本体的目标是获取、描述和表示相关领域的知识，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语词和语词之间相互关系的明确定义。本体的作用是使领域知识显示化、明确化，使领域知识与操作性知识分离，得到对领域知识的一致理解。

本体论是反映客观存在的概念模型，也就是术语和术语之间的关系。概念模型的规范说明为“语词及其相互关系”，它的特点是明确、形式化并且可共享。

### 3.2 本体的层次分类

根据不同的划分标准，可以将本体分为不同的类别。

根据形式化的程度，可以分为以下几类：非形式化、半形式化和形式化。其中，非形式化主要包括自然语言，形式化则是形式语言，而半形式化则包括结构化的自然语言和程度低的形式语言。

根据应用领域的不同，主体研究的重点也不同。具有普遍意义的能接触到客观世界常识的本体被称为顶级身材。特定学科领域的主体被称为领域主体。解决某个问题的解决方案主体被称为任务主体。与问题解决相关的主体被称为应用主体。

其中，顶级实体，表现着对客观世界存在的实体的种类进行区分，具有普遍意义的观念。独立于特定的问题和学科领域，是实际世界的常识。领域本体表现了关于特定学科领域的知识。提供了某学科领域概念的词性和概念的关系，以及与该学科领域相关的重要理论。任务主体可以对某个任务共享，解决与领域无关的问题。应用本体包含属性关系。描述了根据特定领域解决问题的方法<sup>[3]</sup>。各本体间关系如图 2 所示。

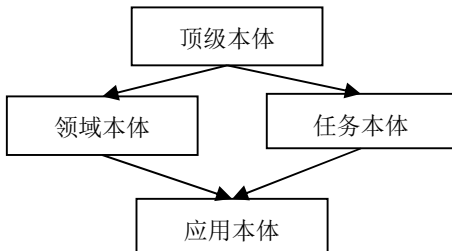


图 2 本体分类及关系图

### 4.2 构建本体的准则

关于构建本体的准则，不同的研究专家给出了不同的解释。Gruber 在 1995 年给出了自己的构建本体的准则，即清晰性、完全性、一致性、可扩展性和最小承诺以及最小编码偏好。其中，最小承诺准则为只定义最必要的术语且只定义约束最弱的关系；而最小编码偏好准则则是不指定术语形式化

用何编码。Arpirez 则提出了另一种构建本体的准则，即概念命名标准化、概念层次多样化和语义距离最小化这三个规则。

## 4 基于本体的语义检索系统

### 4.1 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<sup>[2]</sup>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。

### 4.2 基于本体的语义检索模型

基于本体的语义检索模型<sup>[3]</sup>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。

#### 4.3 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。

#### 4.4 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

的研究还没有形成自己的专业特色和体系。

语义网是互联网的未来,相信随着语义网相关技术的成熟,并推动信息检索的进一步发展。

#### 参 考 文 献

- [1] 杨方颖,蒋正翔,张姗姗.基于本体结构的语义相似度计算[J].计算机技术与发展,2013,23(7): 52-56.
- [2] 张云中.基于形式概念分析的领域本体构建方法研究[D].吉林大学,2009.
- [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] 李桂华,汪学明.基于本体的语义信息检索的研究[J].电脑知识与技术:学术交流,2010.
- [5] 刘树林.基于领域本体信息检索的研究及其实现[D].2009.
- [6] 杨建林.基于本体的文本信息检索研究[J].情报理论与实践,2006,029(005):598-601.
- [7] 刘吉双,陈乙雄.基于本体的文档信息检索模型优化[J].2018.
- [8] 李雪竹,周国祥.基于本体的语义网技术在信息检索中的研究[C]//全国安全关键技术与应用学术会议.2009.
- [9] 王进.基于本体的语义信息检索研究[D].中国科学技术大学,2006.

## 5 总结

信息具有语义描述、信息具有元数据的语义网是提高信息检索效率的有效途径。万维网中的信息因为只具有信息资源,缺乏了对语义的描述,而且它的检索方式也只是关键字检索,所以在信息检索过程中产生了大量冗余以及信息遗漏。同时在万维网中语义是隐式的,机器无法理解语义,这也就不利于信息处理实现自动化。

本文介绍了语义网以及本体的基本概念,系统说明了信息检索中的相关技术以及基于本体的语义检索系统。语义网是近年来出现的一个新概念,越来越多的学者开始研究语义网,语义网作为未来互联网的发展趋势受到了人们的广泛重视。但是,现阶段对语义网的研究还不够深入,例如,关于语义网的理论研究还不够充分,大多是在简要介绍语义网及其相关技术的基础上,对语义网的体系结构、方法和原则、构建策略等方面的探讨较少。语义网