《智能信息处理》课程考试

基于领域本体的语义检索研究现状

尹鹏

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

2021 年 12 月 20 日

基于领域本体的语义检索研究现状

尹鹏

[1)]（大连海事大学 信息科学技术学院, 大连 116026）

摘　要　随着互联网的快速发展，网络上的信息资源种类不断增多，信息量也呈几何级数增长。传统的检索系统往往采用基于关键字的字符串匹配方式进行检索，不考察检索词或检索短语的语义，导致返回的检索结果难以让用户满意。在面对一词多义或多词同义等情况时，这种情况表现得更加明显。与此同时，数据量更大、价值密度更低的大数据时代的到来，使人们更加迫切地需要一种有更高查准率和查全率的检索系统，为此基于领域本体的语义检索系统应运而生。

关键词　领域本体 ；　语义检索 ；

Research on Semantic Retrieval Based on Domain Ontology

Yin Peng

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

Abstract　With the rapid development of the Internet, the types of information resources on the Internet continue to increase, and the amount of information also increases geometrically. Traditional retrieval systems often use keyword-based string matching methods for retrieval, and do not examine the semantics of retrieval words or retrieval phrases, resulting in returned retrieval results that are difficult to satisfy users. This situation becomes more obvious when faced with situations where a word has multiple meanings or multiple words have synonymous meanings. At the same time, the advent of the era of big data with larger data volume and lower value density has made people more urgently need a retrieval system with higher precision and recall. For this reason, semantic retrieval based on domain ontology The system came into being.

Key words　Domain ontology; semantic retrieval;

1　本体相关理论

1.1　本体的定义

本体(ontology)最开始是哲学研究中的概念，由 17 世纪郭克兰纽（R.Goclenius)首次提出，在苏格拉底思想中萌芽，亚里士多德进一步推动发展，关心的是客观现实的抽象本质和组成，且能够对客观存在进行解释和说明。总的来说，哲学领域的本体(ontology)研究两方面内容，一是存在的本质，二是研究客体对象的理论定义和基本特征，而且不依赖某种具体的语言。

人工智能(AI，ArtificialIntelligence)领域成功将本体概念引入，这也是研究人员首次将本体这一概念引入到信息工程领域中来，将其作为知识表示和组织的载体。当对本体的研究更加深入时，本体的内涵也随之发生了变化。按照时间先后顺序列出本体概念发展过程中出现的几个代表性人物：

(1)Fikes，Gruber 等人他们将本体(ontology)定义为给出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规则的定义；

(2)随后 Gruber 经过研究后，给出本体(ontology)是概念模型的明确的规范说明这一更为规范的定义；

(3)Borst 在前者基础之上,提出:本体(ontology)是共享概念模型的形式化规范说明；

(4)Studer，Benjamins 等人进行了更为深入的探索,认为：本体(ontology)是共享概念模型的明确

的形式化规范说明。

其中最为著名、影响最大的并且引用最为广泛的是第 4 种定义，对这个定义进 行分析可以发现该定义包含以下的信息：本体实质上是一种说明，该说明是关于共 享概念模型的，而且是明确的、形式化的。具体来说共享概念模型指对现实世界中 的某些现象的概念进行基于大众认可的抽象而得出的模型；所谓明确性指涉及到的 概念和概念的应用规则是清晰确定的，不能模棱两可；而定义中的形式化表示计算机能够识别该说明并进行分析处理。通俗地说，本体就是收集领域内的重要知识和普遍认可的词汇, 并对这些知识和词汇的进行常识性组织, 最终得到多角度多层次的词汇及其相互关系的规范化说明。

1.2 本体的类别

如何对本体（ontology）进行分类，研究人员给出了各种方案，其中得到广泛 认可的是 Guarino 于 1997 年给出的根据本体描述的详细程度或本体的领域依赖度来进行本体分类。前者指本体能够描述目标对象到什么水平，且没有绝对的衡量 标准。详细程度较高的称之为参考本体，反之为共享本体。后者指本体对该领域的 依赖程度，大体上可分为顶级本体、领域本体、任务本体和应用本体四类:

(1)顶级本体指代抽象程度最高的本体，研究对象仅限于最抽象普通的概念及其相互关系，典型示例为空间；

(2)领域本体的研究范围较顶级本体小，只针对具体领域中的概念及相互关系， 如中医药领域、图书领域；

(3)任务本体处理的是关于任务的概念及概念之间的关系；

(4)应用本体的应用目标则进一步细化，将研究范围限定在那些与特定领域和任务密切相关的概念及其相互关系上。

后来，Perez 等人在总结以往研究成果的基础上进行探索，特别是对 Guarino 提 出的分类法进行更加详细的分解，提出将所有本体分为 10 种[19]： 知识表示本体、普通本体、顶级本体、领域本体、元本体、语言本体、任务本体、领域-任务本体、 方法本体和应用本体。缺点是这样划分出来的本体之间有交叉，且层次结构不甚分明。

1.3 本体的描述

本体的描述可以从建模原语和描述语言两个方面来说明。

(1)建模原语（Modeling Primitive）方面。为更严谨地组织和描述本体，本体研究人员普遍会引入由 Perez， Benjamins 等人归纳总结出的 5 个基本的建模原语， 这几个要素有各自的功能，相互结合可以更好地对本体进行描述：类(classes)或概 念(concepts)，关系(relations)，函数(functions)，公理(axioms)和实例(instances) 。 下面进行详述。

①建模原语中的概念比较抽象，没有任何限制，可以对各种现象进行描述。

②建模原语中的关系表示概念和概念之间是如何发生关联的。该关系可以用 n 维笛卡儿积子集的方式表示出来即

$$R: C_1 \times C_2 \times \cdots \times C_{\mathrm{n}} \quad (2,1)$$

③这里的函数实质上也是一种关系，在这种关系中某元素前的所有元素能够决 定且唯一确定该元素，形式上表示为:

$$F: C_1 \times C_2 \times \cdots \times C_{n-1} \rightarrow C_n \quad (2.2)$$

④公理实质上是永远为真的断言。

⑤实例则代表元素。

从语义上来分析建模原语（Modeling Primitive），实例代指对象；概念可以说是 实例的集合，该部分通常使用层次框架结构来表述，框架包括概念自身的属性描述和其他概念的关系；关系相当于对象组的集合，最基本最常用的关系有部分与整体 关系，继承关系，实例关系，属性关系。在具体的应用中, 研究人员可以根据需要灵活选择合适关系，并且可以使用上述基本关系之外的其他关系甚至自定义关系。

(2) 描述语言方面

当前主流的本体（Ontology）描述语言按形式化程度分为，非形式化语言、非 形式化语言、半非形式化语言、形式化语言[21]。而且自然语言、逻辑语言、框架、 语义 Web 等等多种方式都可以用来描述本体的各方面信息，其中比较专业且适用性 好的是两类逻辑语言。

第一类是以逻辑分析为基础，主要有以下 3 个：Ontolingua 是一种基于 KIF(Knowledge Interchange Format 以一阶逻辑为基础的形式语言)的统一规范的本 体构建语言，它在本体设计、构

造等阶段的均实现了计算机可处理；Cycl 是 Cyc 系统的一种基于一阶谓词演算的知识描述语言，结构体系庞大但不失灵活性，功能 强大，能够满足各种推理方面的需求如等价推理、缺省推理，此外还能拥有二阶谓词演算等其他功能；Loom 是一种基于一阶谓词逻辑的高级编程语言和描述语言，表达能力强、演绎推理能力强大、编程风格多样。

第二类主要描述概念和概念之间的关系，该类语言常用于处理 Web 信息，主要有以下几个：RDF/RDFS，RDF（Resource Description Framework）是一种基本数据 模型，采用“资源-属性-属性值”这种主谓宾三元组结构来表示网络上的资源，但 只能提供和领域无关的机制来描述元数据、资源属性及相关关系，描述领域相关的关系乏力 ，核心类为 Resource/Class/Property、核心特性 Type/subPropertyOf/subClassOf，核心约束 domain/range/ConstraintResource 等，RDFS 在 RDF 基础之上进行了扩展，能够提供更复杂的语义约束，讨论类和属性的关系、 子类和子属性的包含关系等；OWL，这也是目前 W3C 机构推荐的本体描述语言标准，其在表示信息之间的语义关系方面能力较其他描述语言更加强大，而且 OWL 语言还可以结合 RDF Schema 实现对 RDF 模型进行描述，从而实现同时对信息的层 次结构和属性特征进行刻画。不仅如此 OWL 还提供三种子语言 OWL Lite、OWL DL 和 OWL Full，开发人员可以根据需要选择：

下面将按照表达能力的由弱到强的顺序介绍这三种子语言：OWL Lite 提供简单 的分类层次结构和约束，OWL DL 具备强大的表达能力且能保证计算的完备性和可 决定性，OWL Full 在语法方面最为灵活，对 RDF 语法无任何限制，缺点是可计算 性方面稍差。总体来看，OWL DL 语言综合性能（在描述和推理能力两方面）表现较好，故在进行本体描述时一般都会选择 OWL DL 语言。

2 语义检索相关理论

2.1 语义检索和信息检索

信息检索的概念在穆尔斯于 1948 年提出来之后，含义不断得到扩展，在起初定义为信息检索是一种“延时性通讯形式”，之后维克利提倡信息检索是用各种方式检索到含有对检索者有价值信息的文献的过程，再到信息检索是利用用户的查询请求检索到相应信息资源的过程，即网络上的信息资源和用户的查询需求之间，在预先设定的算法处理之下形成匹配的过程。

现阶段的信息检索按照检索过程复杂性分为两种，即广义上的信息检索和狭义的信息检索。广义的信息检索不仅仅单纯指目标的信息集合中筛选出部分信息的过程，还涉及到信息存储等信息处理的其他方面，以及作为检索首要工作的信息存储和对于音视频等特殊文件的信息转换。狭义的信息检索即信息查找，即不关心信息是如何存储的，仅从使用者的角度出发，通过某种匹配算法查找目标信息。

语义检索索本质上作为一种特殊的信息检索方式，要求我们需要对信息检索的模型以及信息检索的方法进行研究。

2.2 信息检索模型及方法

信息检索的过程是对信息源和用户查询之间进行配对筛选的过程。建立信息检索的数学模型，可以帮助研究人员更缜密和准确地描述这一过程，将信息检索中涉及到的概念知识和信息的处理过程抽象化符号化公式化，这样以来便可以更好地进行演绎、推理等工作，并用实践检验。

传统的信息检索模型的大都包含这样一个过程即对文档库中的每篇文档进行某种算法处理，总结出几个单词或称为索引词来代表该文档。这些索引词通常是文档中的和全文大意关系密切的名词，检索系统以后通过索引词与文档库中的文档产生交互。大多数时候会选择名词，这是考虑到名词的语义识别度较高，更能通过几个单词来表现文章的主旨，而形容词、副词常做补语起到补充说明的作用。但这些索引词在描述文档方面的重要程度也不尽相同，故通常根据需要量化索引词的对文档重要程度并过滤掉部分索引词，使用剩余索引词来摘要文档。量化索引词的重要程度可以采用通过为所有索引词分配数组权重的方式来进行。传统的信息检索模型有以下三种：

(1) 布尔检索模型

布尔检索模型是最早的简单检索模型，基于布尔代树和集合论，以二元逻辑为基础，或者说使用一组能够描述文档特征的二元变量来代表文档。故在布尔检索模型中，给定的查询词，在某篇文档中出现则查询词权值为 1，否则查询词权值

为 0。举例来说所以对于文档集合 C 中的任一篇文档p_j，索引词集合 M 就可以表示为($w_{1j}\ w_{2j}\ \cdots\ w_{nj}$)，其中 w_{ij}(i=1，2…n)的值,计算公式为公式 2.3：

$$w_{ij}=\frac{1(\text{文档}p_j\text{中包含}M_i)}{0(\text{文档}p_j\text{中不包含}M_j)} \quad (2.3)$$

用户输入的句子中的索引词可以用三种布尔逻辑运算符 AND、OR、NOT 连接， 例如可以将一个用户需求表示成运算式 2.4：

$$Q=m1OR(m2AND(NOTm3)) \quad (2.4)$$

在用户提交上述查询之后，包含有检索词 m1 和 m2 且不包含索引词 m3 的文档将被全部检索出来。布尔检索模型结构简单，设计思路清晰，是研究人员可以轻松上手的信息检索系统框架，在信息检索领域中得到了广泛应用。但由于采取检索策略过于僵硬，将很多类似的文档排除在外，检索结果并不能令人满意。

(2) 向量空间模型

向量空间模型是现有应用检索系统中的数学模型中，最能表现文档库中文档相互关系的模型。向量空间模型在表示查询中的检索词和文档中的索引词的权重时使用连续数值，这样就实现了二者之间的部分匹配，克服了布尔检索模型中由于采用二元逻辑所带来的缺陷。向量空间模型在表示文档和用户查询时采用矢量形式，矢量由索引词及索引词权重构成，向量之间的距离代表文档和查询之间的相似程度。一般通过两个矢量的余弦夹角来表示文档 mj 和查询 Q 之间的相似度，计算公式如公式 2.5：

$$Sim(m_j,Q)=\frac{\vec{m_j}\cdot\vec{Q}}{|\vec{m_j}|\times|\vec{Q}|} \quad (2.5)$$

其中$\vec{Q}$表示查询矢量，$\vec{m_j}$表示文档 d_j的矢量，Sim(m_j,Q)的取值区间为[0,1]。基于向量空间模型的检索系统会对检索结果文档进行排序，然后根据用户的需求，优先选取前 N 篇相似度值较大者或者按照预先设定一个阈值筛选文档后返回给用户。向量空间模型在设计上没有考虑到用户所输入的检索词和文档之间的语义关联，故计算出检索词与文档之间的相似度，存在不合理成分。

(3) 概率模型

概率模型是一种基于概率排序原理的检索模型，该模型假设用户提供的检索词和被检索的文档之间并不是毫无关联而是存在一种相关概率。该模型克服了前一模型中未意识到关键词和文档之间内在关系的问题。在此检索模型中，会将文档库根 据是否相关分成两个部分，然后根据每一个给定检索词在两部分之间的分布情况来计算检索词和文档之间的相关概率。该模型以严格的数学理论作为依据，并且以相关概率大小对检索结果排序。概率模型可以说是一种基于贝叶斯决策理论的模型。

经典的信息检索方法主要分为以下 3 类：

(1) 数据检索

该种方式主要用于通常结构化信息系统，因为该方式对查询和数据都有结构、格式甚至顺序上的严格要求。但结构化数据的格式统一也带来了好处——支持对特定的字段进行检索。数据检索虽然具有较高的查准率，但查全率得不到保证，容易出现漏检文档，并且用户的理解等其他方面也会影响到数据检索的效率。除此之外，单纯的数据检索方式在语义匹配方面数据检索的表现也不尽如人意。

(2) 全文检索

目前应用比较广泛的一种检索方式就是全文检索，该种检索把用户提供的关键字与文档中的所有潜在词进行字符匹配。显然这种检索方式不具有语义推理能力，检索词与文档之间语义上相关与否不会影响检索结果。全文检索优点显而易见，基于词频、检出信息量大、无需人工干预，日常接触较多的 Google 和百度均属于检索方式。缺点是，只能检索文本类资源且不考虑语义匹配，导致会返回大量不相关信息，用户须消耗大量的时间和精力来对检索结果进行手动筛选，大大降低了检索效率。

(3) 知识检索

鉴于全文检索的不足，知识检索这种全新的信息检索方式应运而生。该信息检索方式基于语义匹配，从文本内容的内在含义出发，实现概念级智能化检索。依托于完整的基于语义的知识结构体系，在对文档库中的文档进行一系列逻辑匹配操作之后检出目标文档。知识结构体系利用元数据对网络上信息进行标注。由此可以看出，知识结构体系是其中的重要一环，是基于知识检索的检索系统的必要前提和构建基础。采用知识检索的检索系统在查全率和查准率方面都有较好的

表现。

2.3 检索效果的评价

检索效果评价是指采用某种方式对检索系统的检索成效性或用户的满意程度进行评价。评价指标因人而异，部分用户期待检索结果更加全面即高查全率，部分用户期待查询结果更加准确即高查准率，此外还有系统的响应时间、数据更新频率、 检索系统的文档收录范围等方面的指标。对检索效果进行评价，可以对所提出理论进行验证，找出可能存在的不足之处以进行修改和完善。本文所采用的评价指标结合主流的查全率和查准率两方面。

查全率或召回率是检出结果的全面性评价指标，是对检索系统检出有关文档效果的考量，可以使用已检出的有关文档的数量和文档库中有关文档的总数的比值来表示，如公式 2.7 所示，其中 R 为查全率，r 为已检出的有关文档的数量，N 为文档库中有关文档的总数：

$$R = \frac{\mathrm{r}}{N} \tag{2.7}$$

查准率是对检索系统检索结果准确程度考量，可用已检出的有关文档的数量和已检出的文档总数之间的比值来表示，如公式 2.8 所示，其中，P 为查准率，p 为已检出的有关文档的数量，M 为已检出的文档总数：

$$p = \frac{P}{M} \tag{2.8}$$

4 总结

本文介绍了如下内容：首先是本体，介绍了本体的起源以及含义的演变，并对本体的功能进行了简单描述，进而介绍本体的分类和描述方式，描述方面包括严谨缜密的建模原语和专业且适用性好的逻辑语言等形式化描述方式。本章第二小节重点介绍了信息检索，描述了信息检索的发展，三种经典的信息检索模型，三种常见的信息检索方法，最后给出对检索效果进行评价的思路。任何一个本体语义相似度算法都不可能解决所有问题,因此,要加强相似度融合技术研究,如:如何根据具体任务选择调用相关算法和确定相关参数。

此外,基于本体的语义相似度研究决不是某个领域技术或专家能够解决的问题,因此要加强领域之间的合作。

参 考 文 献

[1] 邓志鸿．Ontology 研究综述[J]．北京大学学报（自然科学版），2002，38(5)：730-738．

[2] 杨月华，杜军平，平源．基于本体的智能信息检索系统[J].软件学报．2015，7（26）：1676-1677．

[3] Borst W N．Construction of Engineering Ontologies for Knowledge Sharing and Reuse. PhD thesis, University of Twente, Enschede, 1997

[4] 刘琳娜，薛建武，汪小梅．领域本体构建方法的构建方法研究[J]．情报杂志，2007（4）：14-16．

[5] 杨利．基于领域本体的语义检索研究[D]．安徽：安徽大学，2012．