基于本体的语义检索研究

任梦圆

（ 大连海事大学 信息科学技术学院, 大连 116000）

摘要：

传统信息检索主要是基于关键词，将用户的检索请求和全文中的每一个词进行比较，由于字义本身与其概念的延伸不在同一级上，这使得查询结果仅仅是与字面意义或某层意义相匹配，不能准确理解用户查询意图，导致查全率和查准率不高。本体具有良好的概念层次结构和对逻辑推理的支持，能够通过概念之间的关系表达概念的语义信息。因此，基于本体的语义检索能够理解用户的查询意图，并挖掘出语义信息，从而检索出与此概念相关的、用户需求的信息，在语义层次上实现检索，克服了传统信息检索技术的局限性。

关键词:本体;语义检索;语义信息;语义扩展

中图法分类号：TP301

Research on Ontology Based Semantic Retrieval

RenMengyuan

Abstract Traditional information retrieval is mainly based on keywords, they compares the user's retrieval request with each word in the context. Because the original meaning and the meaning extended of the concept is not in the same level, which it makes the search engine misunderstand the intention of the user accurately. As a result the solution only matches the literal meaning of the concept, what's more, the recall and precision ratios of the retrieval are not very quite high. With a good concept hierarchical structure and logical reasoning function support, the ontology can express semantic information based on relations between concepts. Therefore, ontology-based semantic retrieval can understand the user's query intention accurately and exploit semantic information which not only is relevant to the concept but also meet the user's need. The ontology can achieve retrieval in semantic level, which overcomes the limitations of traditional information retrieval technology.

Keywords: Ontology; Semantic Retrieval; Semantic Information; Semantic extension

1 背景

1.1 选题背景

现有的信息检索基本上是基于关键词的，在检索时会查找出一堆用户并不需要的信息，准确率很低;在查全率方面，传统方式缺乏词与词之间同义、相关之类的关系，从而查全率也不能令人满意，出现“忠实表达”、“表达差异”和“词汇孤岛”等问题。具体困难表现为:搜索结果往往是大量的没有价值的信息，很多检索结果和用户查询毫无关系，信息过量，返回太多的无关内容;即使给出我们相关页面，仍需要自己去浏览相关网页，从大量的信息内容中获取需要的相关信息，查询结果不是信息查询，而只是位置查询;基于关键词的检索方法不能理解用户的需要，目前的检索技术仅对关键词进行字面上的匹配，不能根据用户查询目的进行查询的内容扩展，缺乏语义理解和关联。因此，面对数字图书馆的海量信息，如何提高用户在信息检索时的查全查准率，一直是信息检索领域内的一个热点问题。为解决上述问题，应把信息检索从传统的基于关键词层面提高到基于语义的层面。

作者简介：任梦圆（1994-）女，辽宁人，硕士研究生，研究方向：人工智能

1.2 本体在语义检索中的地位

如何让信息表示成机器可识别的知识成为语义检索中的核心问题。本体本身是一个哲学概念，它作为一种能在语义层次上描述信息的概念模型建模工具，能够很好地描述概念的内涵以及概念与概念之间的关系，具有良好的概念层次结构和对逻辑推理的支持，因而在语义检索中得到了广泛的应用。它使信息检索从基于关键词的层面提高到基于语义(或概念)层面上成为了可能。将本体融合到传统信息检索技术中，不仅可以从语义层次上对文档中的信息进行处理，还可以结合用户的检索条件进行语义推理，进而得到较为准确的结果。语义检索利用本体，通过对信息进行语义层次上的分析和理解，力图真正理解检索者的信息请求。

2 语义检索

2.1 信息检索

莫尔斯在 1950 年发表了《把信息检索看作是时间性的通讯》一文，不仅首次提出了信息检索这个概念，并认为“信息检索是一种时间性的通讯形式”。这种认识强调了用户需求的重要性，至今依然对信息检索服务具有很强的理论指导和实践意义。从信息处理的角度来看，如何处理信息和信息的结构是信息检索的基本问题，这种认识把信息检索视为计算机科学与技术的一个分支，认为不仅仅是文档和文献，声音、图像、数据等也都能反映信息。把信息检索看作是一种信息处理的认识，强调了如何构造以及利用什么形式来构造信息结构的问题。在当今因特网迅速发展，网络信息量庞大、参与者众多的情况下，这种认识对于信息检索工具的设计和组建具有指导意义。在信息检索领域，有一种传统的支持者众多的主流观点，那就是从文献查找角度来看的信息检索，这种观点认为查找出含有用户所需信息的文献的过程就是信息检索的基本过程。

2.2 信息检索的定义

信息检索（Information Retrieval）是指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。狭义的信息检索就是信息检索过程的后半部分，即从信息集合中找出所需要的信息的过程，也就是我们常说的信息查寻（Information Search 或 Information Seek)。信息检索（Information Retrieval）是指从信息资源的集合中查找所需文献或查找所需文献中包含的信息内容的过程。

2.3 语义

语义是指数据的含义。语义可以简单地看作是数据(实体、符号)所对应的现实世界
中的事物对象所代表的概念的含义，以及这些含义之间的关系。语义是客观事物对象在人脑中的反映，即人们对客观事物对象的认识，这种认识用数据表示出来就是语义。“数据”由“未被理解的部分—信息”和“已被理解的部分—知识”构成。数据(目标域)=信息(未知域)+知识(已知域)，即 Data = Information + Knowledge。其中未知与已知两部分的划分仅仅是相对于主体(如:人)或载体而言的。

2.4 语义检索

对于语义检索目前没有一个确切的定义，本文认为语义检索就是在语义索引的基础上，利用本体来理解用户查询的语义实现对用户查询的语义扩展和语义匹配，即是对检索过程赋予一定的语义成分。语义检索的目的，就是通过语义理解用户的查询意图，从大量的信息源中找出满足用户请求的信息，并将结果按照与用户请求的相关性大小进行排序后返回给用户。

3 本体

3.1 本体的定义

本体原指哲学中关于客观存在的概念,解释和说明客观存在的系统性,“关心的是客观现实的抽象本质”。后来知识工程学者使用了这个概念,在开发知识系统时获取领域知识.

在思想上指导了人工智能等领域本体的有关研究和应用。

因为人工智能领域的知识建模是在知识库和两个子系统之间建立联系:智能主体行为和环境.然而,有些学者长期以来的做法是根据指定的任务来表达领域知识,这样做虽然只需要考虑相关的领域知识,但是大规模的模型共享、系统集成、知识获取和重用却依赖于领域的知识结构分析.因此,与任务独立的知识库,即本体被提出。

研究人工智能的 Neches 等人将本体定义为“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的规则的定义[7]”。这个定义指出了本体的构建方法,即:先确定领域的术语和术语间关系和规则,再给出术语和关系的定义.按照这种定义,本体中不仅包含显式定义的术语,也包含运用规则推导出来的术语。

3.2 本体的构成

本体的经典定义从一定程度上反映了概念和概念间的关系(即规则)是构成本体的重要元素,但仅仅包含这些还是不够的.Perez 等人认为本体可以用分类方法来组织,并归纳出本体的五个基本组成元素,即认为本体是由概念或类、关系、函数、公理和实例组成的.因此,本体可以用公式的形式描述为:本体=概念+关系+函数+公理+实例,这里的“+”绝非简单的组合,也应该包括这些元素间的内部联系.这五个部分构成了本体的建模元语。

3.3 本体的分类

3.3.1 按照本体研究的主题进行分类

主要分为五类:

(1) 知识表示本体。研究重点是语言对知识的表达能力.具有代表性的是斯坦福大学知识系统实验室提出的一种称为知识交换格式 KIF(Knowledge Interchange Format)的知识描述语言.目前 KIF 被普遍用在专家系统、数据库和智能代理等领域,可以用来实现人与机交互,将不同的执行结果以自身适当的方式展现给用户,如框架、图表和自然语言等。

(2) 通用或常识本体。关注于常识知识的使用.通用知识本体论的研究包括著名的 CYC 工程,其他工作还包括 J.Sowa 的通用本体研究等。

(3) 领域本体.提供特定领域概念和概念间的关系,以及领域中发生的活动和主要理论、基本原理等.在一个特定领域是可重用的。对特定领域本体的研究和应用已经涉及到医学领域、企业本体等。

(4) 语言学本体.是关于语言、词汇等的本体.典型的实例有 GUM (Generalizedupper Model) 和普林斯顿大学研制的 Wordnet。

(5) 任务本体.与以上不同,任务本体研究可共享问题的求解方法,推理方法与具体的领域无关.主要涉及动态的知识,是本体研究的另一个分支。

3.3.2 根据本体研究层次分类

本体的研究和开发工作是在不同层次上进行的.根据本体的研究层次, 可分为:

(1) 顶层本体,研究通用的概念,与具体的问题和领域无关,因此可以在一个很大的范围内共享。

(2) 领域本体,研究特定领域的相关术语或词汇,如医学或企业模型等。

（3）任务本体,研究通用任务或推理活动,如医学诊断.可以通过使用顶.层本体中定义的词汇来描述自己的词汇.与领域本体处于同一研究层次。

⑷ 应用本体,描述某特定应用的,既可以应用特定领域中的概念也可以应用任务本体中的概念。

4 基于本体的语义检索过程

4.1 基于本体的语义检索流程概述

相对于传统的关键字检索，基于本体的语义检索的优势在于体现语义信息，准确表达用户的查询意图。为了更好地体现语义，基于本体的语义检索过程包括查询式扩展和语义检索。

查询式扩展是基础，这是因为实现语义检索的前提是充分理解用户所提出的查询请求，即对用户检索请求进行语义化处理。查询式扩展是用户查询处理的重要步骤，主要是指怎样把用户输入的检索请求赋予语义。利用本体的知识、推理机制以及简单的自然语言处理技术对用户提问进行处理，分析出检索输入的类型和目标，利用本体中的语义关系(如同义、上下位关系等)和推理机制对查询式进行语义扩展，从而更准确的理解用户查询需求，提高查询效率。

具体步骤是:

1.处理用户查询

将用户输入的检索请求进行预处理，得到有意义的关键词组。

2.查询式扩展

就是将转化后的查询词与本体库中的概念相映射，并对其扩展。

语义检索是指使查询结果排序返回，是根据用户查询信息进行排序返回。我们从语义信息的体现着手，通过计算概念间的语义相似度来使查询结果有序输出。

4.2 语义扩展概述

经过分词后,用户输入的查询最终转化为包括单个关键词和多个关键词，其中以多个关键词为主。它们用以描述用户的查询意图，通常包含被检索对象的关键字、关键属性。

1.单关键词查询

对于单关键词查询来说，就是选择合适的领域本体对查询词进行处理，即将转化后的查询词与本体库中的概念、属性、关系和实例等进行匹配，通过扩展形成新的查询词。这时要处理两种情形:一种是查询词是本体中的概念;另一种是查询词不存在与本体库中，仅是一般词语。

2.多关键词组合查询

对于多关键词组合查询来说，就是将用户的查询词通过本体库进行规范化，得到规范化的概念词，同时，根据用户输入的概念、属性或实例，利用本体丰富的语义关系，推理出相关的语义信息，并基于该信息进行检索返回相关的知识内容。例如用户输入查询“计数控制循环语句”的有关信息，在“C++程序设计”双语课程本体库中可以发现“计数控制循环”为一个循环结构名称概念,“语句”为 C 一程序设计的一个概念。利用本体的丰富语义关系，可以推理到属于“计数控制循环”的语句。因为“计数控制循环”是

一个循环结构名称概念，进而可以从本体中找出它的同义词和子概念，如“count-controlled loops ″, ″while 循环”和“for 循环”等;同理对“语句”也可以扩展，方法和上面相同。此外，本文中考虑了概念的属性关系，对于“语句”概念，从本体中可以发现它的属性信息，如“一般形式”、“流程图”、“执行过程”等。进而可以把查询扩展为“循环结构名称概念+语句概念名+一般形式/流程图/执行过程”，充分反应了查询中的语义信息，也明确了用户查询方向。

5 总结与展望

本文分析了研究基于本体的语义检索的必要性，并对此领域的国内外研究现状进行了简要的阐述。通过发现基于本体的语义检索中的不足，来展开本文的内容。由于时间有限，本文只对语义检索过程进行了部分研究。后续的主要工作是在语义检索过程中更深层次地体现语义信息。

6 参考文献

[1]王进，陈恩红，施德明，张振亚一种基于语义相似度的信息检索方法.模式识别与人工智能，2006, 19(6):696-701.

[2]陈泳，林世平.基于本体的语义检索技术 计算机工程与应用，2006：78-80.

[3]王家琴，李仁发，李仲生，唐剑波一种基于本体的概念语义相似度方法的研究 计算机工程，2007, 33(11):201-203.

[4] Borst W. Construction of Engineering Ontologies[D]. PhD thesis, University of Twente, Enschede, 1997.

[5] Studer R. Knowledge engineering: principles and methods [J]. Data and knowledge engineering