《智能信息处理》课程考试

本体论及其在信息检索中的应用

曾 敏

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

2016 年 11 月 10 日

本体论及其在信息检索中的应用

曾 敏

（大连海事大学信息科学与技术学院 辽宁大连 116026）

摘 要 本文介绍了本体论的哲学含义以及它在计算机科学技术领域中的意义。并通过分析本体论在计算领域中某些方面的典型应用，论述了本体论的起源和发展、定义、描述语言和应用。 然后具体分析了本体论在信息检索中的应用，构建基于 Ontology 的 Web 信息检索系统。同时提出了基于本体论的信息检索模型。该模型支持用户查询的导引，并按领域分类有选择地返回查询信息。
关键词 本体论，信息检索系统，全文检索，领域分类，概念化

Ontology and Its Application in Information Retrieval

Zeng Min

（Dalian Maritime University，Computer Science and Technology，Liaoning，Dalian，116026）

Abstract This paper introduces the philosophy meaning of ontology and its significance in the field of computer science and technology, and discuss the origination, development, definition, description language and application of ontology through the analysis of ontology in some typical applications in computing. Then the application of Ontology in the information retrieval are analyzed in detail, and the Web information retrieval system based on Ontology is constructed. At the same time the paper puts forward the model of information retrieval based on ontology which supports to be guided by the user's query, and selectively return query information according to the field of classification.
Key words Ontology, the system of Information Retrieval, full-text retrieval, classification of domain, conceptualization

随着信息技术的发展，特别是 Internet 应用的普及，人们已从信息缺乏的时代过渡到了信息极大丰富的时代。Internet 上信息分布在位于不同位置的站点上，据统计到 1997 年夏季已经有 1.5 亿个 Web 主页分布在 65 万个站点上[1] 。

目前网络上的搜索引擎一般使用 2 种技术来实现信息检索:一是使用网站分类技术，即把网站进行树状的归类，登录的网站属于至少一个类别,对每个站点都有简略的描述。雅虎采用了这种方法。为了分类科学准确，需要有一支由各科人才组成的维护队伍。二是使用全文检索技术。全文检索技术处理的对象是文本，它能够对大量文档(这里是大量网页数据)建立由字(词)到文档的倒排索引，在此基础上，用户使用关键词来对文档(网页)进行查询时，系统将给用户返回含该关键词的网页。

一般来说，由于使用了专家来对网站进行归纳和分类，网站分类技术为网络信息导航带来了极大的方便，受到人们的欢迎。但是它维护成本较高，而且对网站的描述也十分简略，其描述能力不能深入网站的内部细节，因此用户不能查询网站内部的重要信息，造成了信息丢失。

全文检索是一个很成熟的技术，它能够解决对网页细节的检索问题。从理论上说，只要网页上出现了某个关键词，就能够使用全文检索用关键词匹配把该网页查出来，但是这又导致了它的缺陷:返回的信息太多。更严重的是，除了综合性的搜索引擎站点有这个现象之外，现在较大的站点对自身站内信息的检索也会返回大量的网页。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。在一些场合，高的查全率带来的成千上万的命中网页。在网页爆炸性增长的今天，没有一个用户有时间和精力来浏览搜索引擎查到的网页。当前的搜索引擎的缺点是不支持用户的信息导引。本文提出了基于本体论的信息检索，支持领域的分类，并按领域分类有选择地返回网页,提高了检索的效率。

1 本体论

1.1 本体论的起源以及定义

本体来自希腊词汇，onto 表示 being，logos 表示 to reason,最初是哲学上的一个分支，用来表示事物的本质和组织，哲学家用它来回答事务研究的基本问题。在现代信息和计算机领域，本体论主要应用在计算机原理、人工智能和信息技术的发展，这三个领域的发展都需要对特定的领域进行通用的描述，这一点正是本体所可以回答的问题。

本体以一种明确的、形式化的方式表示领域概念及其之间的关系，成为人、机器、应用程序对概念语义达到共同理解的媒介，在应用间实现知识的共享及重用。其获得承认最广泛的定义为 1993 年美国 Sanford 大学的知识系统实验室(KSL, Knowledge systems Laboratory) 的学者 ThomasGruber 给出的定义。表述为:“ 知识的形式化表达的基础是概念化,概念化包括研究领域内的对象、概念和其他实体,以及他们之间的关系;因此，概念化是我们为了某些目的用来表示世界的一种抽象、简化观点，每一个知识库、知识系统、基于知识层次的 Agent 都明确的或是隐含的遵守某个概念化。本体论是对概念化明确的规范说明，在 AI 领域，存在是可以表示。当领域知识以一种明确的形式化进行描述的时候，被表示的对象的集合构成论域，对象的集合及其可描述的关系通过知识表达语言的词汇描述。因此,在 AI 领域，我们可以通过定义只是表示的术语集合来定义程序的本体。在这样的一个本体中，论域中实体的名字通过定义与人可以理解的文本相关联，定义描述了名字包含的意义、公理，公理限制了这些术语的解释及形式化使用。从形式化角度来说，本体是一个逻辑理论的陈述性描述。”

1.2 本体论的描述语言

从本体论的定义可以看出，本体的特征为概念化、定义和描述，因此在本体描述语言是本体论研究中的重要方面。因为本体只有被形式化后，才能被机器理解，作为一种独立于语言的概念化表示方法，可以使用不同程度的形式化方法对其进行描述。并且可以根据形式化程度的不同,选择用自然语言来表示本体，或者使用框架、语义网络或逻辑语言来描述，采用相应的推理机制实现概念隐含语义的明确化。从根本上来说，各种本体表示语言主要侧重于两个方面即 ontology 的共享和表示。它们的语法主要基于 XML 和 LISP，语义基于概念图和框架表示。一个理想的本体交换语言应该具有比较容易学习的语法、语义，及很强的表示能力。由于 XML 语法比 LISP 更容易学习,且 XML 是当前 Web 交换数据的标准格式，可以在计算机之间方便的解析各种类型的数据，框架表示比概念图更容易确定,且 RDF 和 RDF Schema 提供了描述 Web 信息资源的通用框架。许多本体表示语言都采用 XML 语法及 RDF 及 RDFSchema 描述 Web 资源。

1.2.1 XML

XML 是一个严格符合 SGML 格式的、结构化的语言，实现了文档的显示和数据分离。这种结构化的数据易于使用、携带、传递，是目前 Web 数据交换最好的语法格式。XML 提供 DTD、XML Schema 对文档结构的进行有效性验证，通过描述/ 约束文档逻辑结构实现数据的语义。XML 对本体的描述，就是利用 DTD 或 XML Schema 对本体所表达的领域知识进行结构化定义，然后再利用 XML 文档结构与 XML 内容之间的关系对本体知识进行描述，从而提供对数据内容的语义描述。但是由 DTD 自身描述能力有限、没有数据类型的支持、约束定义能力不足，无法对 XML 实例文档做出更细致的语义限制等。因此无法表达概念间的继承关系，并不能完全满足 XML 自动化处理的要求。XMLSchema 虽然解决了 DTD 存在的问题，但是 DTD、XMLSchema 为 XML 文档提供的约束机制时只是用限定 XML 文档所用到的标记和这些标记之间的结构关系，语义仍然是隐含的。因而 XML 所表示的本体是轻量级(Light weight)的本体，只能保证人们使用相同的词汇，是一种较低层次的本体的应用，本体中不包含有用的语义信息。

1.2.2 RDF

RDF 被认为是表示和处理半结构化数据的一种极好选择，它为描述 Web 元数据(metadata) 的语义提供了数据模型，用来描述对象(资源) 以及其之间的关系。由于语义 Web 只是对已知的事实和结论进行表达，因此 RDF 在这个层次上，RDF 是一个完备的、形式化的系统。RDF 对资源的描述基于这样的思想:利用当前现有的 Web 体系结构中的标识符 UIRs 作为标识符系统来标识事物，用简单的属性及属性值来声明资源。但是，RDF 只是提供了一个用于领域无关的机制来描述元数据，描述资源属性及其相关关系,但是并没有提供按照类的机制描述信息资源、声明属性、

描述属性语义及与其它资源之间的关系。也是说 RDF 不能描述领域相关资源的语义关系，如同义词、一词多义等。虽然 RDFS 弥补了这方面的缺陷，RDF 在语义的描述上深度仍然欠缺。

1.2.3 OWL

OWL 语言是一个定义和示例 Web 本体的语言，是 RDF 的扩展，它既是 Web 标识语言，又是本体描述语言，在 Web 上发布和共享本体。和 XML Schema 相比，OWL 语言是知识表示，不是信息表示格式;和 RDFS 相比，OWL 不仅可以用更复杂的方法描述类，如 disjoint，而且扩展了 RDFS 属性，允许表示属性的 transitive，Symmetric 及 functional 性质，表达了更强的概念语义信息，支持描述逻辑推理。

1.3 本体论的应用

Ontology 的功能主要是实现某种程度的知识共享和重用，它能使计算机对信息和对语言的理解上升到语义层次。所以，Ontology 在一些涉及到信息的互操作、知识理解等方面的领域具有很大的应用前景。在信息检索方面，Ontology 使得传统的基于关键字的检索，上升到语义检索的高度。其基本思想是:先建立相关领域的 Ontology，根据 Ontology 对收集的信息进行标注，用户的检索请求按照 Ontology 转换成规定的格式，在 Ontology 的帮助下匹配出符合条件的数据集合返回给用户。在信息集成方面，分布式信息集成面临的主要问题是结构、设施的异构和缺乏统一的语义集，借助 Ontology 可以在一定程度上解决语义异构的问题。采用语义方法进行信息集成的特点是扩展性好、适应动态信息源、支持语义级查询。在机器翻译方面，通过把某种语言中的词汇映射到 Ontology 中的概念，可以支持在源语言分析时进行歧义消解和目标语言生成时的词汇选择，并可以作为源语言和目标语言的中间表示的概念来源。另外，在知识获取、数据挖掘、软件工程等相关方面,本体论(Ontology) 都有广泛的应用，但是由于 Ontology 构件的困难，以及构建技术的不成熟等原因，真正意义上的应用还没有完全展开，大部分项目尚在试验阶段。

1.3.1 在智能信息检索中的应用

对于普通的 WWW 用户,“信息过载” 已经成为一个日益严重的问题。目前广泛使用的信息检索或者是依赖编码过程(即对于给定的内容使用特定的观点或分类方法进行描述)或者是进行全文检索。由于编码的描述只能反映内容的一部分，单个词汇的出现更是难于反映文献的内容，所以以上方法都难于确保检索内容的精确匹配。

在实际应用中，人们逐渐认识到使用语义进行检索是一种解决精确查询的有效途径。但是语义检索依赖显示标注的信息资源，或是完整、正确的自然语言理解系统。Ontology 在智能检索信息系统中提供了形成查询和资源描述所必需的元语，以 Ontology 技术为核心建立领域语义模型,为信息源提供语义标注信息，使系统内的所有 Agent 对领域内的概念、概念之间的联系及领域内的基本公理知识有一个统一的认识，从而能够显著地提高系统的联想能力和精确性，有望快速、高效、精确地检索出用户所需的有价值的信息，同时也提供给系统内的所有用户对领域的一个全面的共同视图[3] 。Ontology 已逐渐成为一种智能信息检索系统的知识表示，是系统集成的核心部件。

1.3.2 在面向对象分析中的应用[4]

面向对象分析在当前需求分析中最具代表性，面向对象分析是把图和语义网络模型与面向对象程序设计语言中的概念结合在一起而形成的分析方法。这个方法采用了实体、关系、属性等信息造型中的概念，同时采用了密封、对象、类的结构和继承性等面向对象程序设计语言的概念。面向对象分析本质上是自底而上的过程，通过对具体事物的认定和抽象，归纳概括出共性，区分出个性，用类和类层次结构加于表示。由此可见，面向对象分析方法是以对象和对象类为中心进行的,对象和对象类组成了一定的层次关系，这种垂直的组织方式表示了元素之间具有的父子关系，而其它方面的内容，如对象间的关系，对象间的消息传递，则相对处于次要地位。但在现实世界中，项目和软件可能极其庞大和复杂，要确切掌握不同对象和对象类之间的各种关系比理解单个对象模块更为困难，软件工程师往往需要付出更多的精力和时间来分析对象类之间的关系，而不仅仅只分析对象本身。软件开发者的注意力不应被个体对象的具体内容所吸引，而应集中考虑如何利用所获得的大量可重用软件构件去构造新的软件。因此软件构件之间的关系描述应受到足够重视。

本体是领域概念的显示表示。根据本体论的思想，某个领域的本体就是关于该领域的一个公认的概念集，其中的概念含有公认的语义，这些语义通过概念之间

的各种关联来体现。本体通过它的概念集及其所处的上下文来刻画概念的内涵。由此可见，本体强调相关领域的本质概念，同时也强调这些本质概念之间的关联。在面向对象分析中，使用本体的思想和本体描述现实世界的方法,可以将对象之间的各种关系用形式语言充分刻画出来。

1.3.3 在软件构件重用中的应用

基于构件的软件开发技术，旨在通过重用技术，提高软件的开发效率，避免一些不必要的重复劳动。可重用的软件构件和相关信息通常被存储在各种各样的数据库中，由此可见，存储构件的数据库是分布式的，并且是异构的。如果软件开发者对领域中的重用构件一无所知，自然检索空间则是整个 Internet 网络。为了快速有效地找到所需构件，必须设计基于 Internet 的搜索引擎，该搜索引擎能够根据用户输入的关键字，快速准确地返回相关的信息资源表。由于可重用的构件库是分布式和异构的，所以，为了便于检索，必须在用户和软件构件库之间建立中介层。在中介层，人们将领域本体的概念用在软件构件的组织与管理中，通过本体集成统一的构件数据库[5]，从而提高了软件的开发效率。

1.3.4 在知识工程中的应用

一般在开发基于知识的系统(Knowledge -based System ， KBS)时，知识工程师很难定义系统在应用领域中具体、完整的工作方式，因此一般的 KBS 系统都采用进化的原型系统方法进行开发。知识工程师将本体论概念引入知识工程，详细说明模型中涵盖的概念、实例、关系和公理等实体，并以此建立领域本体。通过使用元属性对属性进行分析，并对属性提出了一种针对本体建模概念化分析的形式化方法，解决了知识共享中的一些问题，有效地促进了来自不同领域的研究人员和组织间的交流[6] 。

由此可见，将本体论应用于计算机科学技术领域，将问题领域中的术语、术语间的联系及领域中的公理组织起来，建立本体，并提供形式化方法和工具，则能使所建立的本体被方便地共享和重用，从而解决了计算机领域中的许多难题。

2 信息检索中本体论的应用

2.1 传统信息检索中的瓶颈

理想的信息检索要求快速、准确、全面。人们为了实现这些要求，开发了多种技术，传统信息检索技术大致可分为 3 类:全文检索、数据检索和主题检索。全文检索把用户的查询请求以关键词的形式与全文中的每一个词进行比较，检索方式主要基于词频分析技术，比较有代表性的是 Google 和百度。这种方式检出信息量大、自动程度较高,缺点是返回信息过多，检准率很低。数据检索主要针对结构化信息系统，查询要求和数据都遵循一定的格式，具有一定的结构，允许对特定的字段检索。各种商业数据库多采用此种检索方式。数据检索依赖于编码的质量，检索花费大，检出的信息相对准确，但检全率很低。主题检索以人工方式或半自动方式收集信息，对访问文档进行描述,将之添加到合适的事先确定好的主题类别中。用户从基本大类，一级一级地向下访问，得到检索结果。代表为 Yahoo！这种检索方式的优点就是便于用户自己掌握检索进度，但是费时较长，且更新不及时。上述信息检索方式所面临的困难实质在于传统的信息检索技术缺乏知识处理能力和理解能力。人们认识到必须将信息检索从目前基于关键词层面提高到基于知识(或概念) 层面，从而提出了知识检索的概念。知识检索强调的是基于知识的、语义上的匹配。目前知识检索，特别是面向 Web 信息的知识检索是信息检索研究的重点。

2.2 本体论在信息检索中的作用原理

人们对知识(或概念) 检索进行了积极的探索，提出了各种各样的思想，例如概念聚类、空间向量等等,不一而足。而本体论因为所具有的良好的概念层次结构和对逻辑推理的支持，因而在信息检索领域得到了广泛地应用,成为研究热点。

从本体论的概念上进行理解，本体就是把现实世界中的某个领域抽象成一组概念以及概念之间的关系,进而构造出这个领域的主体，在信息检索领域中，本体在某种意义上就是领域知识模型。

基于 Ontology 的知识检索的工作原理可以阐述为:在领域专家的帮助下，建立基于领域概念知识的领域 Ontology。收集信息源中的数据，并参照已建立的领域 Ontology ,把收集来的数据按规定的格式存储在元数据库(关系数据库、知识库等) 中。对用户检

索界面获取的查询请求，查询转换器按照 Ontology 把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。检索的结果经过定制处理后，返回给用户。

2.3 本体论在信息检索中的具体应用

一般说来，基于 Ontology 的信息检索分为三个步骤:用户提出语言申请→基于请求的信息匹配→匹配信息的输出，本文从这三个步骤分别说明 Ontology 技术在信息检索中的应用。

2.3.1 基于Ontology 的检索请求处理

传统的信息检索工具提供给用户的主要是基于关键字的检索接口，但是在很多情况下用户真正的检索意图很难用几个关键字表达清楚，这也是导致现有检索系统的精度不高的原因之一。因此为了能够更好的让用户表达出他的检索意图，我们提供给用户的检索接口是自然语言的表达方式。用户可以以自然语言的方式向系统提出问题，然后我们利用 Ontology 领域中的知识和一些简单的自然语言理解的技术对用户的问题进行分析，提取主题词，得到用户真正的检索意图，然后将检索请求提交给系统的检索部分。在进行处理的过程中，首要的问题就是建立基于 Ontology 的分词数据库，然后对用户的问题进行概念类型识别和问题类型识别。概念类型识别的作用是根据句法分析的结果和领域 Ontology 中的概念类型模板，识别出该问题所描述的概念的类型。通过概念类型识别之后我们可以知道该问题所关心的是人物这个概念中的某个属性，有的放矢,从而减少信息检索的处理时间。问题类型的识别是指将用户的问题根据问题类型库划分到一个指定的类型中。在用户提交问题之后，系统就需要结合领域 Ontology 中所表述的词汇的语义知识，分析判断问题的类型。得到问题的概念类别和类型之后，系统就可以根据主题词库从用户问题中提取出检索关键词并将它们提交给系统的检索部分。

2.3.2 基于Ontology 的文本预处理

信息检索的目的，就是根据用户的检索要求，从大量的信息中找到满足用户要求的信息，并对检索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。在基于 Ontology 信息检索系统中，文本预处理的目的是从非结构化的文本信息中提取出文本中的有用信息并根据领域 Ontology 的概念类型模板形成信息实体，从而将这些非结构化的文本信息转化成具有一定结构的信息实体。在对文本内容进行分析处理之前，我们事先将整篇文本划分若干小段文本,然后再进行分词与词性标注的处理，并且在分词过程中进行概念和逻辑的本体描述，为后一步的文本分析打下基础。文本分析包括实体特征选取和信息实体构建两个方面，Ontology 技术都参与其中，并且根据领域 Ontology 中的领域抽象概念类的描述进行特征和实体的区别，从而在繁杂的信息中准确地检索出匹配信息返回给用户。

2.3.3 信息实体的索引与检索

信息检索的最终目的是让用户能够快速而准确的得到其所需的信息，而系统的信息索引模型和检索机制的好坏则会直接影响整个系统的性能，因此对于信息检索系统来说，一个好的信息索引模型和检索机制是必不可少的。对信息实体进索引的首要工作就是要进行信息实体特征项的选取。实体特征项可以是文本中的各种语言单位,对于中文来说可以是字、词、短语，甚至是句子或者句群等更高层次的单位。因此，特征项的选择只能由处理速度、精度、存储空间等方面的具体要求来决定。在基于 Ontology 技术检索过程中，为了能够通过特征项快速查找到信息实体，文本信息中所有的特征项用一个链表连接起来，每一个特征项节点指向一个信息实体的链表，实体链表中的每个节点记录实体库中的实体号。通过实体号系统就可以很方便的从实体库中查找到该实体的信息。

在文本预处理阶段，我们将自然语言文本根据领域 Ontology 中的知识转化成大量的信息实体;在问题处理阶段，我们将用户的问题转换成对某个信息实体的属性的查询，通过这两部分的处理之后，我们就将自然语言检索的问题转换成了对信息实体检索的问题。信息实体的检索分为三个阶段。首先，检索部分得到从问题处理部分提交过来的检索关键词，把关键词都作为信息实体的特征项，寻找它在特征项链表中的位置;然后，我们去掉重复实体号之后，到信息实体库中就可以查找到实体的所有信息;最后，把剩下的信息实体按照它们所包含的检索关键词的数量进行排序，根据在问题分析阶段中分析出来的结果属性，提取出实体中对应属性的值作为检索结果返回给用户。

2.4 基于Ontology 的Web 信息检索系统

根据上文的 Ontology 技术应用的具体分析，我们可以总结出基于 Ontology 的 Web 信息检索系统包

含 5 个主要处理模:Ontology 管理模块、问题处理模块、文本预处理模块、信息检索模块、库文件管理模块。5 个模块相互协作,共同完成用户问题的回答任务。对于用户来讲，他直接把问题提交给问题处理模块，然后等待处理的结果。系统在接受了用户的问题后，首先进行问题分析和处理，得到检索关键词和问题类型。检索关键词将提交给检索部分去获取信息实体，问题类型将被转换为结果属性提交给检索模块，用于从信息实体中提取结果属性值。检索模块提取出实体的结果属性值以一定的形式返回给用户。

3 基于本体论的信息检索模型

我们提出了基于本体论的信息检索模型，如图 1 所示.

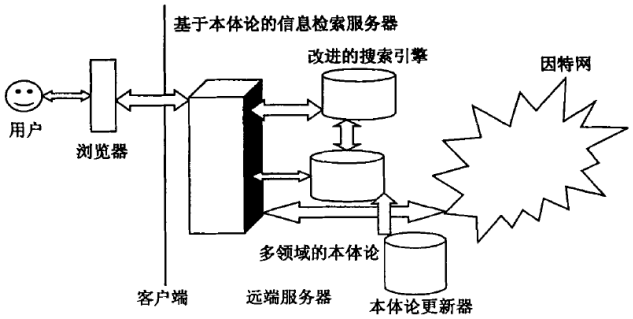

图 1　基于本体论的信息检索模型

Fig.1 Information retrieval model based on ontologies

3.1 多领域的本体论

领域分类 $D = \{d_1, d_2, \cdots, d_n\}$，本体论 $O = \{O_1, O_2, \cdots, O_n\}$,本体论 O_i 对应于领域 d_i，由此我们建立了本体论到领域的对应关系。多领域的本体论是本体论的集合。

3.2 改进的搜索引擎

改进的搜索引擎采用全文检索技术。全文检索技术处理的对象是文本，它能够对大量文档(这里是大量网页数据)建立由字(词)到文档的倒排索引。改进的搜索引擎加上了由关键字到领域的领域索引表。比如:例　有一篇文章，如果它出现的词是比如地球、月球、太阳的词等等，那么这篇文章出现的词:“火星”，极大可能的意思应是:九大行星之一，一般不会是火中的火星的意思。在改进的搜索引擎中，建立了从关键字到领域的索引，支持领域分类，如图 2 所示。

表 1　关键字的索引表

WebLink1	HTML 文件	出现该字(词语)的片段
WebLink2	HTML 文件	出现该字(词语)的片段
…	…	…
WebLinkn	HTML 文件 n	出现该字(词语)的片段 n

表 2　WebLink1 对应的领域索引表

领域 1	领域 1 领域相关度
领域 2	领域 2 领域相关度
…	…
领域 n	领域 n 领域相关度

图 2　从关键字到领域的索引表

Fig.2 Index table from key word to domain

3.3 本体论更新器

Nicala Guarino 认为应该按照层次关系，建立不同的本体论。在建立了顶层本体论之后，就可以着手建立领域本体论了[5] 。Nicala Guarino 对建立本体论的方法学进行了讨论[3] 。

本体论是世界的反映。因此它必然随着现实的发展而变化。更新本体论的方法有 2 种方式:人工方式和系统在已有的知识上对因特网上信息学习自动更新。搜索过程:当用户提交一个查询后，比如输入了“火星”，由浏览器交给了远端的基于本体论的信息检索服务器。远端的基于本体论的信息检索服务器通过查询本体论，得到这个关键字的信息，比如:这个关键是否是一个术语，以及这个术语在不同领域的含义等等。如果不是一个术语或者说不是一个概念，这只好交给改进的搜索引擎检索，按传统的搜索引擎方法对它进行检索。如果是一个术语或者说是一个概念,则在本体论中(可能是很多领域的本体论集)有它的入口。在本体论中得到这“术语”的信息，如:属于某个领域集合及该领域集的定义、用法示例、相关的主题、同义词，如果本体论支持多语言，还有其它语言的同义词等等。把这些信息返回给用户，用户可以根据它关心的领域对查询结果进行过滤，这就缩小了查询的范围。也可以选择关键字的概念，由系统作概念到领域的映射。把选择的结果交给远端的基于本体论的信息检索服务器，基于本体的信息检索服务器对结果进行处理后，交给改进的搜索引擎。最后，搜索引擎把查询结果返回给用户。

3.4 关键字到本体论的映射

基于本体的信息检索模型中，当用户提交了一个关键字的查询后，基于本体的信息检索服务器将在本体论集中得到该关键字的信息，如:属于那个领域、同义词、定义、还有示例、语义关系等。

一个本体论可以表示为一个有向图 G=(V,E)，其中 V 是结点，E 是有向边，其类型有多种，比如:属于那个领域、定义、同义词、和其它概念的联系、是什么概念的子概念等。

把关键字映射到本体论集，如果本体论中出现了这个词汇，则取出领域、同义词、定义、示例等信息。

3.5 网页通过本体论映射到领域集合

关于这个问题，我们提出了两种方法，第一种方法是针对网页有关键字的情况;第二种方法针对网页没有关键字的情况。

方法 1 如果一篇文章有关键字，可以采用以下的方法[6]:

函数 Terminology(Oi)从领域 D_i 对应的本体论中求出该领域的术语集(包括同义词);函数 Definition(O_i ,Keyword)从本体论 O_i 中求出关键字 Keyword 的定义;函数 Relation(O_i)从本体论 O_i 中求出由概念关系构成的语义网络集。

设 O_1 , O_2 , … , O_v 分别是领域 D_1 ,D_2 , … ,D_v 的本体论，术语集 T_i = Terminology(O_i)，其中(0 ≤ i ≤ n)，Ks ={Key_1 ,Key_2 ,… ,Key_n }为被检索文档 Doc 中所给出的关键字。

任一文档中所给出的关键字应体现该文档最核心的内容，这些最核心的内容若不出现在该领域的本体论中，则说明该文档与这一领域无关，即 $K_s \cap T_i = \psi \Rightarrow Doc \notin D_i$ 这里 $1 \leqslant i \leqslant n$。

经过这一步，我们可以滤掉不相关的领域，得到所有可能与该文档相关的领域，其 Ds_1 ,Ds_2 , … ,Ds_k , 其中 $Ks \cap Ts_j \neq \emptyset$, $S_1 \leqslant S_j \leqslant S_k$ 。

接下来进行近似语义网络匹配。首先求出与关键字的定义相关的术语集合。D_S ={dk |(dk ∈Ts_j)∧(dk 出现于 Key i 的定义 Definition(Osj ,Key j)中 Key i ∈ KS , $1 \leqslant i \leqslant m$, $S_1 \leqslant S_j \leqslant S_k$)},然后求与关键字集直接相关的术语对象集合直接相关的术语对象集合 RO ={obj | 存在 x (x ∈ Ks ∩(obj , x)Relation(Osj))), $S_1 \leqslant S_j \leqslant S_k$}。检索整个文档，统计被检索文档里出现在集合 DS ∪ Ks 中元素的频度 freqi ,freqi 体现了该文档中的术语与 Os_j 中的语义网络的近似匹配程度。我们定义 Degree(Os_j)=$freqs_j$, 因此可以再根据 Ds_1 , Ds_2 ,…,Ds_k , 与被检索文档的相关程度的大小 Degree(Os_j)对它们进行排序。通过上述过程，可以依据本体论对文档进行按领域的分类。

方法 2 首先，对网页进行取词，得到了一个词汇集。在本体论的协助下，取出的词或概念都是具有意义的。然后，直接统计这些词汇在领域本体论出现的次数。我们定义:词汇出现的次数和这个网页的词数比称为该词的领域相关度。对领域相关度确定阈值，当领域相关度大于阈值就认为该网页与这个领域相关。于是，可得一个领域集和领域集的领域相关度。

结束语 Ontology 作为一种新的知识组织方式，力图去解决知识的共享和重利用问题，在知识越来越丰富的今天，受到了越来越多的关注，在许多方面有着广泛的应用前景。但是基于 Ontology 的信息检索研究目前仅仅停留在理论研究方面，在具体的实施和系统构建上还是空白，其中最大的原因就是 Ontology 理论和实践的不成熟，而且运行耗费比较大，时间周期长，这些问题还有待于进一步的解决。

参 考 文 献

[1] BRAKE D. Lost in cyberspace[J] .New Scientist, 1997 ,154(2 088):12—13 .

[2] GUARINO N. Formal ontology and information systems[A] .Formal Ontology in Information System[C] . Tren-to :IOS Press, 1998.6-8 .

[3] GUARINO N ,WELTY C.A formal ontology of operties[A] .Proceedings of the ECAI-OO workshop on applications of ontologies and problem soliving methods [C] , Berlin, 2000 .121— 128 .

[4] USCHOLD M , GRUNINGER M .Ontologies: principles methods and application[J] .Know ledge Engeering Review, 1998, 11(2):93 — 155 .

[5] GUARINO N, WELTY C. Onto logical analysis of taxonomic relatiorr ships. Proceeding s of ER-2000:the conference on conceptual modeling [EB/OL] .http ://www .ladsed.pd.cnrit/ infor/ ontology/ papers/ ontology papers.html 2003,04,11.

[6] Regina M .M .Braga , Marta Mattoso and Claudia M .L Werner .The Use of Mediation and Ontology Technologies for Software Component Information Retrieval [Z] .ACM Software Engineering Notes. Vo l 26, No .3, 19 -28 .

[7] Guarino N .Formal Ontology and Information System [M] .Trento :IOS Press1998 .6-8.

[8] Wu Cheng-Gang, An information retrieval server based on ontology and multi-agent, Journal of Computer Research & Development, 2001 , 38(6):641— 647((in Chinese)
(武成岗.基于本体论和多主体的信息检索服务器[J] .计算机研究与发展, 2001 , 38(6):641— 647).