《智能信息处理》课程考试

基于本体的科技文献检索技术研究

王娅菲

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

2021 年 12 月 19 日

基于本体的科技文献检索技术研究

王娅菲

（大连海事大学　信息科学与技术学院　大连　116026）

摘　要　传统的关键字检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。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。语义信息的缺乏将会导致科技文献检索数据的错误描述，不同来源的数据结构会使得数据信息出现差异和重复性。使用基于语义的科技文献检索机制去除传统检索模型的缺点，将检索内容与语义关联算法相结合，引入概念的语义相似性与基于原则的语义推理，对于本体的构建和规则的制定推理出适当的对象的属性，并且构建出检索原型，对现实的研究具有积极的意义。

关键词　本体;信息检索;语义 WEB;查询扩展

Research on Technology Literature Retrieval Based on Ontology

Yafei Wang

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

Abstract　The traditional keyword search method lacks the semantic processing of the search content, and the search results may be incomplete and inaccurate. Based on the existing semantic retrieval, this paper makes a simple analysis and research on semantic retrieval based on ontology. The lack of semantic information will lead to the wrong description of the retrieval data of scientific and technological literature, and the data structure of different sources will cause the difference and repetition of the data information. Use the semantic-based scientific literature retrieval mechanism to remove the shortcomings of the traditional retrieval model, combine the retrieval content with the semantic association algorithm, introduce the semantic similarity of concepts and principle-based semantic reasoning, and make appropriate reasoning for the construction of the ontology and the formulation of rules. The attributes of the object and the construction of the retrieval prototype have positive significance for the study of reality.

Keywords　ontology；information retrieval；semantic web；query expansion

1　引　言

在大数据时代，要从海量的信息数据中检索出符合自己查找的内容，传统的检索技术手段变得越来越低效[1]。传统的信息检索手段一般是基于关键词匹配，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，用查全率与查准率来衡量检索效果不太合适，且利用这种方式获得的结果大多是松散的，不成体系的。针对目前检索方式的不足，Tim Berners Lee 提出了语义网的概念[2]。语义技术的核心是本体技术，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。相较于机械式的检索方式，根据语义的检索，能够合理地对检索结果进行概念扩展，使得检索结果的查全率和查准率获得极大的提高。同时伴随着 Lucene、Nutch 和各种爬虫技术的高速发展，全文检索的搜索引擎也被大量使用在检索过程中，基于本体技术的层次化的语义相似度思路上的研究成果也十分丰硕。本文通过阐述信息语义共享和本体技术的运用，分析和实现语义

检索模型的语义扩展和规范化推理过程，对检索词的量化扩展，可以给用户提供令人满意的信息检索效果。

2 科技文献检索及其语义问题

2.1 语义概述

语义网的概念最早由 Tim Berners Lee 提出[2]。语义 Web 主要是为了说明两个实体间的关系而产生的，主要是用于网页数据。从那时起，语义 Web 的概念就一直在扩展。目前，语义学的重要意义是用包含语义学的链接来描述世界上两个实体之间的关系，形成一个包罗万象、具有推理能力的庞大知识库。语义网扩展了当前互联网的功能，显示出事物都是相互联系的。语义网可以理解为进行人与计算机交互的实体[3]，可以促进人们更好地利用互联网中的数据。

关联概念模型用于知识建模、知识存储、知识共享和推理知识生成新知识。语义Web包括xml、rdf、owl、本体等重要概念。本体是对现实世界的抽象描述，它只包含有价值的数据。语义Web的总体结构及其不同层次的语义表达功能，如图 1 所示。

科技文献数据模型多，语义模糊，数据稀疏，难以建立固定的结构化模型。利用语义本体对科技文献数据进行建模，可以较好地解决这些困难。不同的信息检索模型采用不同的语义本体对科技信息的数据进行描述，语义本体是这种模型的基础，该模型可以统一地管理这些元数据。对语义本体来说，这样的做法更加精确，可以发挥出更好的效果。

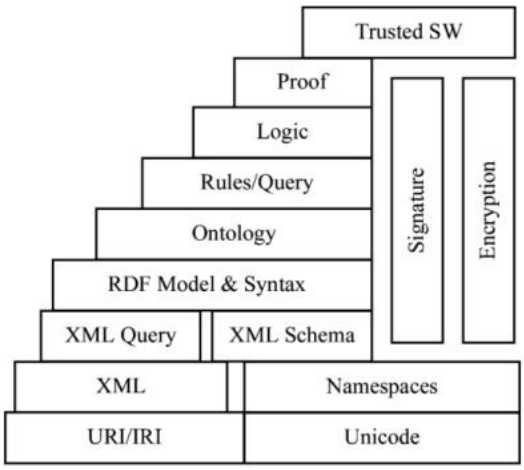

图 1 语义网层次结构

2.2 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。

2.3 语义信息共享

支持语义的信息检索模型主要是为了对数据进行操作，而数据共享主要考虑语义共享和语法共享两个重要的知识点。语义共享注重信息的内容，语法共享考虑对数据访问的问题。在数据领域中，需要保证在数据信息共享的过程中，语义转换的信息不会产生数据丢失，对应于上下文的语义环境对数据进行语义描述。

语义映射还可以应用于更广泛的领域，提供了该领域概念的统一描述。本文提出的语义配置和映射是相似的，但又有所不同。语义映射是一种解决由不同系统表达的相同概念的方言问题。其次，语义网通过融合语义的扩展定义来扩大当前的 Web，尤其针对语义 Web。进行语义的科技文献检索中，十分需要向体系结构以及语义实现技术的结合。

3 本体技术及语义检索模型

在本体技术中，重点考察的是客观事物的抽象本质，本体是相关研究领域的词汇关系与术语的综合，是共享概念模型的形式化描述。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]。概念模型的表现含义独立于语义的环境状态。明确的意义是指概念上的定义的约束形式。数据共

享保证相关领域的概念集合，针对的是某种概念的总体集合而非独立的个体信息。针对本体概念的形式化描述，具有不同的构建方式，本体具有的几种特征要素分别是其声明、公理、概念、属性以及关系[5]。

本体的描述语言 OWL 是在 DAML 描述语言的基础上发展而来的。OWL Lite 保证用户的简单约束，表现一个分明的层次分类方法，其转换速度更为迅速。而 OWL DL 则是支持推理功能的系统，利用推理方法增加计算的完全性与可靠性，提供良好的逻辑处理方法与可推理性的计算性质。OWL FULL 提供丰富的表达能力，在 OWL FULL 中，自身既可以作为语义个体存在，也可以作为多个个体的集合，还可以在本体的基础上，支持预定义推理成分。OWL Lite、OWL FULL 都可以作为 RDF 的约束化扩展。而 RDF 则可以作为 OWL FULL 的文档。

相比于传统的科技文献检索，基于语义检索的模型具有更高效的检索过程、更加准确的检索结果的手段。传统的科技文献检索包含截词检索、全文检索、布尔逻辑检索以及字段限制检索[6]。而本体在科技文献检索方面则具有明显优势。本体可以用来表示丰富多彩的相关领域现象的知识的逻辑抽象，本体对知识的获得和积累是等级结构严密的、知识描述全面和概念规范化的机器推理和自动化处理方式[7]，并且保证知识的不断的动态更新。本体对相关领域的知识具有删除、修正和改变的可操作性。

4 基于本体的科技文献检索实现

4.1 系统分析和整体架构

需要在相关领域的专家指导下，帮助建立起相关领域的本体，将数据源根据严谨的数据结构方式补充到数据源中的文献当中。对于用户界面的响应将查询个体转换成规范化格式，匹配出相关知识领域的集合，经过定制化处理后，将获得的检索结果呈献给用户，完整地实现了由字面匹配向语义概念匹配的提高。结构的构建图如图 2 所示。

该系统功能的架构流程包括系统的本体构建，建立本体与关系数据库的映射关系，建立区语义索引，实现检索引擎的功能，加入本体索引的文件，利用推理机对用户输入的关键词进行合理化的概念推理和扩展，使用 Jena 工具对本体施行快速化的查询操作。

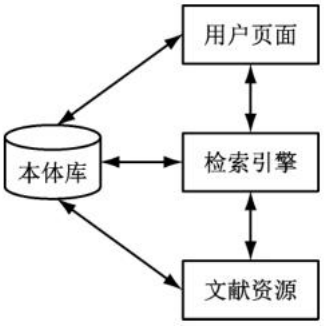

图 2 方案设计结构图

系统的开发语言采用 Java 语言，该语言可移植性强，是面向对象的、分布式的编程语言，开发环境 IDE 采用 eclipse，该开发环境集成了许多语言开发包。本文所使用的全文索引应用包 Lucene 是定义了索引文件格式、基于 Java 语言的语义开发工具，可以兼容不同的文本格式，具有强大的查询引擎，降低了学习扩展的索引能力，默认包含了模糊查询方式、文本布尔操作方式以及分组查询操作。

系统的总体构架包含用户页面模块、本体构建模块、文献映射模块以及检索处理模块，各个模块之间保持着协作和联系的关系状态，共同组成一个完成的总体，实现一个完整的查询功能，其系统体系结构如图 3 所示。

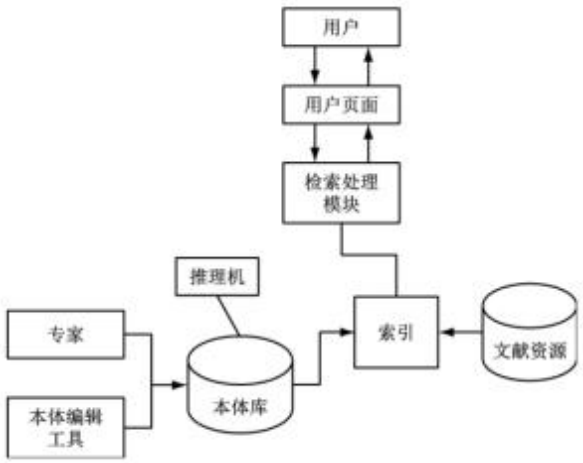

图 3 系统体系结构

用户登录进入用户页面，发送用户请求，系统产生响应模式，用户可以操纵检索处理模块。通过相关领域专家和本体编辑工具构建出本体，本体资源主要由推理机和本体库组成。推理机接收到检索索引提示，通过本体库发出

索引，文献资源单位依据索引，返回出合适的科技文献数据。

4.2 模块设计与算法实现

考虑到本体与科技文献数据库的映射关系，将本体与文献数据相结合，构建起对应的关系。数据库的信息由海量的文献数据信息组成，当本体库与文献数据库联系在一起时，检索该领域知识的某些个体知识就会生成合理的索引构建。每个文献实例的产生会与它检索到的最为靠近的数据概念组合在一起放入到索引文件中去，形成一个完整的总体，其流程如图 4 所示。

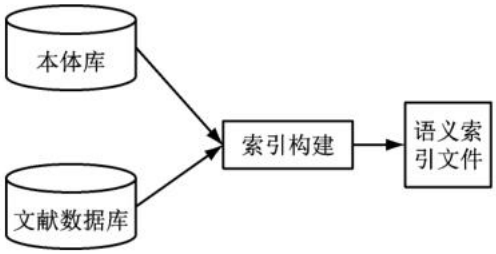

图 4 本体/文献映射模块结构图

在基于 Lucene 的检索引擎设计的基础上进行二次开发，建立起一个面向对象的高效检索引擎，通过语义检索建立的索引文件，采用推理机制实现对原始搜索信息的推理优化，具体如图 5 所示。根据 Jena API 对本体建模语言的数据结构存储分发到数据库后台，通过输入的查询语句，放入到 Lucene 的推理引擎，对查询语句优化后，实现加载推理范式。

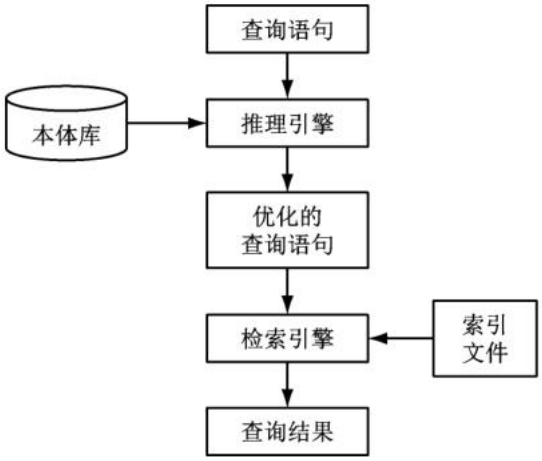

图 5 推理引擎与检索引擎关系

5 总 结

本文从实际出发，与传统的科技文献检索作比较，研究了文献检索与语义 Web 的联系。通过使用语义模型的本体技术对检索方式进行改善，针对科技文献检索的模型，提高检索的准确率，分析了概念之间、实例之间的语义关系。最后通过将全文检索方式、本体检索方式和基于语义推理检索的方式，采用优秀的面向对象的编程方法，以本体技术为核心支撑点，采用逻辑推理方式进行语义检索，得到用户需要的检索结果。在未来的研究中，提高文献数据的清理算法效率，减少重复检索记录和检索时间，需要深入考虑各种语义关系的排序方式，以满足不同用户的检索需求。

参 考 文 献

[1]张孝飞,孔繁秀.基于语义概念分析的科技文献检索研究[J]. 情报理论与实践, 2016 ,39(8):115-118.

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

[3]胡昌平,林鑫.科技文献检索中基于主题词表分面化改造的分面构建[J]. 情报学报, 2015, 34(8):875-884.

[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]谭德坤.模糊粗糙集在科技文档检索中的应用研究[J]. 计算机仿真, 2011, 28(10):168-172.

[6]王莉,梁冰,白海燕. 基于本体的科技文献检索框架与技术实现[J]. 数字图书馆论坛, 2012(7):37-44.

[7]葛慧丽, 叶志飞. 一种基于迭代运算引文排序的科技文献检索系统[J]. 计算机时代, 2011(9):15-18.