《智能信息处理》课程考试

基于 protégé的排序算法本体搭建

李锦峰

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

2020 年 12 月 06 日

基于 protégé的排序算法本体搭建

李锦峰

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要 排序算法作为计算机程序设计中的一种重要操作，广泛的应用于不同的领域。针对不同的应用场景，所使用的排序算法也是不同的，传统的排序算法检索是基于语法匹配的，并不能满足不同应用场景下的排序算法的检索。本体作为一种有效表现概念层次结构和相互关系的模型，能够有效地表示排序算法与各个应用场景之间的联系，使得信息检索能够在语义的层面上进行，有效的解决了传统信息检索中查不准，查不全的问题。本文借助本体构建工具 Protégé进行排序算法本体的搭建，将排序算法与其应用场景联系起来，从而达到在特定应用场景下准确查找排序算法的目的。

关键词 排序算法 本体 Protégé

Construction of sorting algorithm ontology based on protégé

LiJinFeng

(Computer science and technology，Dalian maritime university, LiaoningDalian,116026,China)

Abstract As an important operation in computer programming, sorting algorithm is widely used in different fields. For different application scenarios, the sorting algorithms used are also different. The traditional sorting algorithm retrieval is based on grammatical matching, and cannot satisfy the sorting algorithm retrieval in different application scenarios. As a model that effectively expresses the hierarchical structure and interrelationship of concepts, it can effectively express the relationship between the ranking algorithm and various application scenarios, enabling information retrieval to be carried out at the semantic level, effectively solving the traditional information retrieval Inaccurate, incomplete investigation. This paper uses the ontology construction tool Protégé to build the sorting algorithm ontology, linking the sorting algorithm with its application scenarios, so as to achieve the purpose of accurately finding the sorting algorithm in a specific application scenario.

Key words Sorting Algorithm;Ontology; Data mining; Protégé;

1 引言

随着信息技术的快速发展，互联网每天都在产生大量的信息，这些信息中大部分都是无用冗余的信息，怎么从互联网每天产生的大量信息中检索有用的信息就至关重要。传统的信息检索技术都只采用语法匹配，而没有利用数据之间的语义关系，导致它们的查全率和查准率往往达不到预期的效果[4]。在本体出现之后，传统检索出现的查不准，查不全得到了有效得解决。所谓本体，通俗来讲，是用来描述某个领域甚至更广范围内得概念及概念之间的关系，是概念和概念关系的集合。也正是因为本体中包含了概念之间的关系，所以基于本体的检索，不在是只通过语法匹配，而是在语法匹配的基础上又充分的利用了概念之间的语义关系，大大提高了信息检索的查全率和查准率。

排序是计算机程序设计中的一种重要操作，广泛的运用于不用的领域，如数据调度，信息检索，信息储存等。而针对不同的应用场景，所选择的排序算法是不同的，传统的排序算法检索是基于语法匹配的，并不能满足不同应用场景下的排序算法的

检索。本文采用 protégé5.5 本体编辑工具，构建排序算法本体，解决传统排序算法检索中出现的查不准，查不全问题。

2 本体概念及其构建方法

2.1 本体定义

本体是知识的集合，它可以实现在特定领域中的概念共享、知识重用和语义理解。目前在自然语义处理、人工智能、信息检索、知识表达等相关领域已经有本体理论的成熟应用。目前在世界范围内，由于不同专家学者对本体的理解不同，所以对于不同的定义也有所差异。本文采用了中国学者张晓林教授(2002)关于本体的定义[1]，张晓林教授认为本体是概念集，是特定领域公认的关于该领域的对象及其关系的概念化表述。

概念化、形式化、可共享、明确、描述领域知识是本体的五大特征。其中形式化是指将本体中的概念信息用形式化的语义准确描述出来。概念化是指本体用抽象方法对客观世界进行建模。可共享是指本体中的概念是领域中公共认可的通用概念。明确性是指，本体的概念信息以及其之间的语义关系能够被明确的阐述，让计算机能够精确处理信息。描述领域知识指的是本体是对特定领域的信息表示。

2.2 本体的描述语言

本体描述语言实质上是一种可以对本体进行描述的代码，有助于更好的描述领域内的本体。本体描述语义可以划分为两类：基于谓词逻辑的本体描述语义和基于 Web 的本体描述语言。

基于谓词本体描述语言主要包括[1]Ontlongua、OCML、LOOM、Cycl 和 Flogic。其中,Ontolingua、OCM L 和 Flogic 是基于一阶谓词逻辑和框架模型的本体描述语言,LOOM 是基于描述逻辑的,Cycl 是在一阶谓词逻辑基础上进行扩展的二阶逻辑语言。基于 Web 的本体描述语言主要包括[10]，XOL、RDFS、SHOE、OIL、DAML+OIL 和 OWL。

随着计算机科学技术的发展，基于 Web 的本体描述语言逐渐成为主要的本体描述语言。由于 OWL 是 W3C 的推荐标准，符合 PDF/XML 标准语法格式，并且能够与多种本体描述语言进行兼容和交互，所以应用范围很广，深受用户的喜爱。

2.3 本体构建方法

本体建设的方法学还没有成熟的理论做指导，而且目前的本体构建方法都是针对具体的项目提出的，这就导致各种本体构建方法的出现[10]。目前最具有代表性的本体构建方法有[1]，骨架法、IDEF5 法、七步法、五步循环法、METH-ONTOLOGY 法、KACTUS 法、SENSUS 法和循环获取法。七步法是基于本体构建工具 Protégé 的本体构建方法，也是本文构建本体的方法。

2.4 本体构建工具

为了解决本体构建复杂和工作量大的问题，研究者们开发出许多本体构建工具，目前市面上比较流行的本体构建工具主要有[8]Web Onto、Ontolingua、Onto Sarus、Onto Edit、和 Protégé。由于 Ontolongua，Web Onto，Onto sarus，OntoEdit 源码都是不开放的，所以本文选择 Protégé 作为本体的构建工具。

Protégé 软件是斯坦福大学医学院生物信息研究中心基于 Java 语义开发的本体编辑和获取软件。Protégé 提供了本体概念类，关系，属性和实例的构建，并且屏蔽了具体的本体描述语言，用户只需要在概念层次上进行领域本体模型的构建。Protégé 工具通过 Class（类），Object Properties（对象属性），Data properties（数据属性），Annotation Properites（注释属性），Datatypes(数据类型)，Individuals（实例）6 个选项来描述本体。Class 选项用来定义本体的类及其子类。Object Properties 选项用于设定类之间的关系，可以设定作用的定义域，值域以及属性特性，其中属性特性包括传递性，函数，逆函数和对称性。Data properties 选项用来设定类的某一个方面的特性，比如在算法类中，算法的时间复杂度是算法类的一个特性，其定义域是不同的算法，而值域则是它的数据类型。Individuals 选项中，用来对前面建立的类创建实例，在创建实例时，需要对前面建立的相应类的属性进行值的设定[2]。

3 排序算法本体知识相关概念及联系

3.1 排序算法定义

所谓排序，就是使一串记录，按照其中的某个或某些关键字的大小，递减或递增的排列起来的操作。排序算法，就是如何使得记录按照要求排列的方法。排序算法大体可以分为内部排序和外部排序

两大类。图 3.1 是常见的排序算法的划分图。

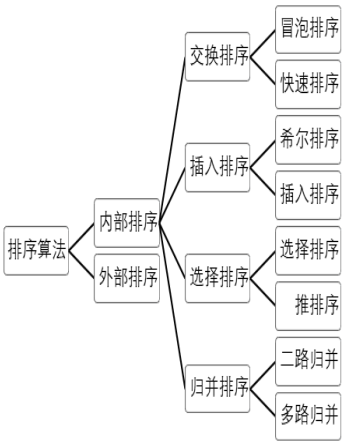

图 3.1 排序算法划分图

3.2 排序算法的时间复杂度和空间复杂度

在不同应用场景中如何选择合适的排序算法，有两个指标至关重要，这两个分别是算法的时间复杂度和空间复杂度。算法的时间复杂度和空间复杂度是用来描述算法效率的两个指标。同一场景下使用不同的算法，尽管最终得到的结果都是一样的，但在过程中消耗的资源和时间却会有很大的区别。图 3.2 展示了各个排序算法的空间复杂度以及时间复杂度。

排序方法	时间复杂度（平均）	时间复杂度（最坏）	时间复杂度（最好）	空间复杂度	稳定性
插入排序	$O(n^2)$	$O(n^2)$	$O(n)$	$O(1)$	稳定
希尔排序	$O(n^{1.3})$	$O(n^2)$	$O(n)$	$O(1)$	不稳定
选择排序	$O(n^2)$	$O(n^2)$	$O(n^2)$	$O(1)$	不稳定
堆排序	$O(nlog_2n)$	$O(nlog_2n)$	$O(nlog_2n)$	$O(1)$	不稳定
冒泡排序	$O(n^2)$	$O(n^2)$	$O(n)$	$O(1)$	稳定
快速排序	$O(nlog_2n)$	$O(n^2)$	$O(nlog_2n)$	$O(nlog_2n)$	不稳定
归并排序	$O(nlog_2n)$	$O(nlog_2n)$	$O(nlog_2n)$	$O(n)$	稳定
计数排序	$O(n+k)$	$O(n+k)$	$O(n+k)$	$O(n+k)$	稳定
桶排序	$O(n+k)$	$O(n^2)$	$O(n)$	$O(n+k)$	稳定
基数排序	$O(n*k)$	$O(n*k)$	$O(n*k)$	$O(n+k)$	稳定

图 3.2 排序算法的空间复杂度以及时间复杂度

3.3 排序算法的应用

根据上述排序算法的空间复杂度和时间复杂度可以归纳出排序算法的使用场景。

（1）数据的规模较小的时候，可以采用直接插入排序或直接选择排序；

（2）若文件初始状态基本有序，则应选用直接插入或冒泡排序；

（3）当数据规模较大时，应用速度最快的排序算法，可以考虑使用快速排序。当记录随机分布的时候，快速排序平均时间最短，但是会出现最坏的情况，这个时候的时间复杂度是 O(n^2)，且递归深度为 n,所需的占空间为 O(n)；

（4）堆排序不会出现快排那样最坏情况，且堆排序所需的辅助空间比快排要少，但是这两种算法都不是稳定的，要求排序时是稳定的，可以考虑用归并排序；

（5）归并排序可以用于内部排序，也可以使用于外部排序。在外部排序时，通常采用多路归并，并且通过解决长顺串的合并，加上长的初始串，提高主机与外设并行能力等，以减少访问外存额外次数，提高外排的效率；

（6）特殊的桶排序、基数排序都是稳定且高效的排序算法，但有一定的局限性：

根据上述总结的排序算法使用场景，建立排序算法和各个应用场景的联系，该联系是构建排序算法本体的关键。

4 protégé 中排序算法本体的搭建

本文借助本体构建工具 Protégé 对排序算法进行本体构建，选用的版本为 protégé5.5 版本。构建过程主要分为类的构建，对象属性定义以及数据类型的构建，通过本体的构建展示排序算法知识及其相关关系[7]。图 4.1 是本体构建工具 Protégé 的主页面图。

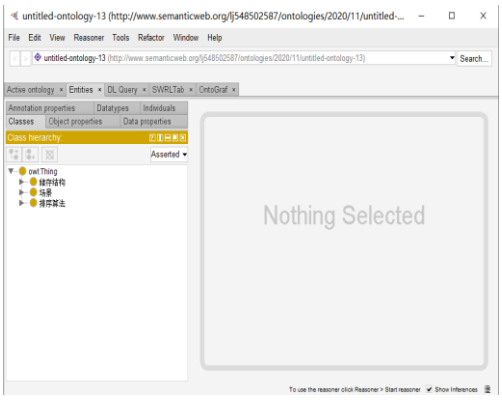

图 4.1 Protégé 主页面图

4.1 排序算法本体类的构建

选择标签 class，创建基本类结构，设置三个大类，包括：排序算法，应用场景，储存结构， 分别定义出各类。其中排序算法又分为内部排序，外部排序。应用场景分为数据储存，数据调度以及数

据检索。储存结构分为顺序存储和链式存储。

4.2 排序算法本体对象的属性定义

选择标签 Object properities 分别对于构建好的排序算法，应用场景以及数据结构进行对象属性的定义，主要定义了以下几个类于类之间的关系：1.使用：表示排序算法使用了什么储存结构。2.应用于：表示排序算法应用于什么现实场景。

4.3 排序算法本体数据类型定义

我们在对排序算法进行描述时，需要通过数据类型的定义进行完善。在本体的构建中，类相当于数据库中的表的形式，而数据类则相当于数据库中的列。这个需要对每一个类进行定义。比如，排序算法的时间复杂度和空间复杂度[7]。

4.4 排序算法本体结构图

我们通过以上对本体类，对象属性及数据属性的定义，可以看到排序算法本体类的本体结构图，本体结构图是对整个排序算法知识结构的描述。其中箭头表示不同的关系，实线箭头表示其父类与子类的关系，虚线箭头则时针对其对象属性的定义，不同颜色的虚线箭头表示不同对象属性。排序算法本体结构图如下图所示：

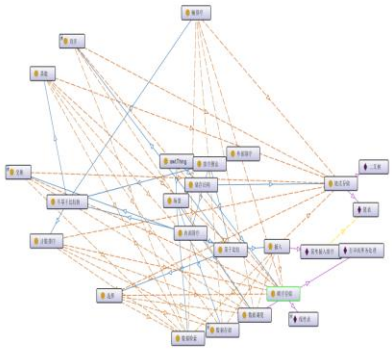

图 4.2 排序算法本体结构图

5 结束语

本文通过对目前排序算法现状及具体知识情况的掌握，对排序算法进行综合的知识汇总，借助工具 Protégé 构建排序算法本体，将排序算法和排序算法应用场景进行关联，最后进行可视化的图像展示，完成了排序算法知识结构构建。对于 protégé 工具在排序算法知识构建方面，深入研究可以实现知识索引，建立更加全面，完整的排序算法知识库。

参 考 文 献

[1]王向前,张宝隆,李慧宗.本体研究综述[J].情报杂志,2016,35(06):163-170.
[2]蔡群英, 黄镇建. 基于protege的课程内容本体的构建[J]. 计算机系统应用, 2012.
[3]范轶,牟冬梅.本体构建工具 protégé 与 KAON 的比较研究[J].现代图书情报技术,2007(08):18-21.
[4]王珊,张俊,彭朝晖,战疆,杜小勇.基于本体的关系数据库语义检索[J].计算机科学与探索,2007(01)
[5]龚资. 基于 OWL 描述的本体推理研究[D].吉林大学,2007.
[6]于娟,马金平,李永.基于 Web 本体语言 OWL 的知识表示[J].计算机工程与设计,2006(22).
[7]张莉,王玉廷.基于 Protege 的糖尿病本体构建[J].科学咨询(教育科研),2020(03):9-11.
[8]赵国梁. 基于石油领域本体的语义关联检索[D].中国石油大学(华东),2018.
[9]李善平,尹奇韡,胡玉杰,郭鸣,付相君.本体论研究综述[J].计算机研究与发展,2004(07):1041-1052.
[10]韩婕,向阳.本体构建研究综述[J].计算机应用与软件,2007(09):21-23.