《智能信息处理》课程作业

基于领域本体的信息检索研究

刘锦

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

2020 年 12 月 3 日

基于领域本体的信息检索研究

刘锦

(大连海事大学 信息科学技术学院, 大连市 中国 116033)

摘 要 随着网络信息的急剧增加，人们对于信息检索的要求越来越高。人们希望可以借助信息检索工具，快速、准确的找到所需要的信息。但是，当前基于关键词匹配的信息检索方法在查全率和查准率方面具有很大的局限性，具体表现为：一方面，检索到的信息过于繁杂，且大多数是无用信息；另一方面，相当多的有效信息在检索过程中被遗漏。为了改进基于关键词的信息检索方法，本文提出了依据领域本体的信息检索思想，把信息检索从基于关键词的层次提高到概念层次，有效提高了信息检索的查全率和查准率。

关键词 信息检索；领域本体

中图法分类号 TP312

Research on Information Retrieval Based on Domain Ontology

Liu Jin

（Dalian Maritime University Department of Information and Science , Dalian China 116033）

Abstract With the rapid increase of network information, people have higher and higher requirements for information retrieval. People hope to use information retrieval tools to quickly and accurately find the information they need. However, the current information retrieval methods based on keyword matching have great limitations in recall and precision. The specific manifestations are: on the one hand, the retrieved information is too complicated and most of it is useless information; on the other hand, On the one hand, quite a lot of valid information was missed in the retrieval process. In order to improve the method of information retrieval based on keywords, this paper puts forward the idea of information retrieval based on domain ontology, which improves information retrieval from keyword-based level to conceptual level, which effectively improves the recall and accuracy of information retrieval.

Key words Information retrieval; Domain ontology

1 引言

随着 Internet 的迅速发展与推广，网络上的信息已变得非常庞大，特别是大数据时代的到来，进一步提高了对信息检索的要求。如何进行高效、准确得信息检索已经成了人们最关心的话题之一 。常听到人抱怨，利用现有检索工具来查询某一信息，得到的却是一堆垃圾信息，很少有他们想要的东西。原因就在于目前主流的信息检索技术主要是传统的基于关键词匹配的检索技术，在检索过程中这些关键词字符只能从字面上来理解其含义，而词汇的内在概念无法表示出来，所以在信息检索过程中时常会出现检索结果不全、检索结果还会出现一些用户不需要的信息;同时检索结果也很难检索到关键词背后潜藏的信息[1]。

2 信息检索

2.1 信息检索概述

信息检索(Information Retrieval)是信息资源与信息需求的匹配过程[2]，是通过一定的算法寻找信息资源与信息需求的交集的过程*(*如图 *1* 所示*)*。由于信息资源空间和信息需求空间的不确定性，信息检索是信息需求向信息资源不断靠近的过程，是一个摸索的过程，是一个逐步求精的过程。

作者简介：刘锦，男，1996 年生，硕士研究生在读，研究方向为图像处理. Email:liujin@dlmu.edu.cn

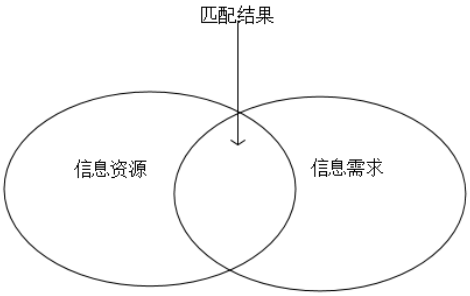

图 1 信息搜索逻辑模型

根据不同的标准，信息检索有不同的分类。按照检索技术可以将信息检索分为三类：全文检索（Text Retrieval）、数据检索（Data Retrieval）和知识检索（Knowledge Retrieval）。

全文检索是指计算机索引程序通过扫描文章中的每一个词，来对其建立索引，指明该词在文章中出现的次数和位置;当用户查询时，检索程序就根据事先建立的索引进行查找，并将查找的结果反馈给用户的检索方式[3]。数据检索的特点是查询要求和信息系统中的数据都遵循一定的格式，具有一定的结构，允许对特定字段检索。数据检索需要有标识字段的方法，其性能取决于所使用的标识字段的方法和用户对这种方法的理解，因此具有很大的局限性。数据检索支持语义匹配的能力也较差。知识检索强调的是基于知识的、语义上的匹配，因此在查准率和查全率上有更好的保证。目前知识检索是信息检索研究的重点，特别是面向 Web 信息的知识检索。

2.2 传统信息检索技术的特点

传统信息检索技术主要是关键词匹配的语法进行匹配，关键词搜索引擎先将信息资源进行分析，抽取关键词，并建立索引数据库。其优点是查询速度快，且容易实现。但同时也有一定的局限性，即当用户在输入框输入想要查询的关键词后，搜索引擎会把数据库中包含该关键词的所有信息作为匹配结果一股脑地全部返回给用户。首先，这就会导致原本对用户没有太大作用的信息也显示了出来，用户再自己去筛选，既浪费了时间也占用了过多的网络资源，效率低下；其次，网络信息发布的随意性较大，很多信息更是良莠不齐，无疑增加了信息的不确定性和用户的不安全感，使信息质量和精度降低，其可靠性、权威性和利用价值受到质疑，令用户无所适从；再次，导致了“词汇孤岛"问题。在人的大脑中，概念并不是孤立存在的，它总是与其他概念之间存在各种各样的联系，正是这种联系造就了五彩缤纷的现实世界，而在传统信息检索中，这种概念之间的语义联系是很难描述的。

3 本体及领域本体概述

3.1 本体概念及表示

本体(Ontology)，最初是哲学领域概念，当时的人们把本体当作从柏拉图到黑格尔的西方传统哲学的主干，这就意味着，本体其实是形而上学的哲学分支，它对客观世界的事物进行分解，发现其基本的组成部分，进而研究客观事物的抽象本质。自从 20 世纪 90 年代，本体被入人工智能、知识工程等领域以来，关于本体的定义，学术界一直没有统一的定论，但存在基本的共识，即本体包括:概念化、可明确、形式化、可共享这四大特征。“概念化”是从实际现象中抽象出相关概念，进而得到其相应的模型;“可明确”是指抽象出的概念以及对概念的使用是有明确的规定;“形式化”是指本体可以利用计算机进行相应的处理;“可共享”体现对知识领域的共同认可程度，同时也反映了大家共同认可的概念集[4]。

本体可以通过一个五元组（C，R，A，P，I）来表示。其中：C（Class）是本体中的类集；R（Relationship）是类的层次关系集合；A（Axiom）为公理；P(Properties)为属性；I（Instances）是实例[5]。

（1）类或概念：领域内的概念，可以是实际存在的真实事物，也可以是抽象的概念。

（2）关系：用于描述类与类之间的关系。

（3）函数：是一种特殊的关系，该关系中 前 n-1 个元素可以唯一决定第 n 个元素。

（4）公理：表示永远成立的事实。

（5）实例：某个类中真实存在的个例。

根据以上的表示方法，用简单的有向图表示本体。如图 2。

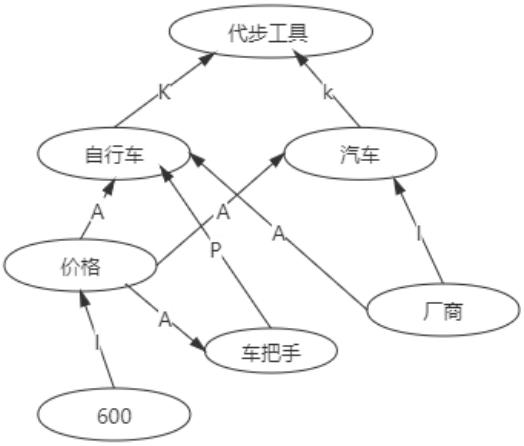

图 2 本体的例子

3.2 领域本体的概念及形式化描述

领域本体是本体的一个研究层次，是对领域共享概念的明确的、形式化的和规范化的说明[6]，它的目标在于捕获相关领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并且按照层次关系给出这些词汇和词汇之间相互关系的明确定义，以实现软件系统对这些概念的共享和重用。

领域本体结构 O 可以表示为一个三元组 O = {C，R，A°}。其中，C 表示概念的集合;R 表示关系的集合；A°表示本体公理的集合。按照领域本体结构 O 来表达某一具体领域本体的词汇空间 L 可以表示为一个四元组:L={ L^c， L^R， F， G }[7]。其中，L^c 是用来表达领域本体概念的词汇的集合; L^R 是用来表达领域本体关系的词汇的集合; F 是建立 L^c 与 C 中实体之间对应关系的映射的集合; G 是建立 L*与 R 中实体之间对应关系的映射的集合。需要说明一点，L^c 与 C 以及 L^R 与 R 之间的映射并不要求严格一一对应。即可以存在若干词汇表达某一概念或关系，也可以存在若干概念或关系都可以以某一词汇表达。

3.3 领域本体在信息检索中的优势

首先，领域本体的结构呈树状结构显示，有效支持了概念之间的逻辑推理。其次，领域概念被予以形式化和规范化地表示，实现了软件系统之间、人与软件系统之间对这些概念的共同理解，同时也实现了机器自动求解的推理过程。

3.4 本体构造准则

在构建本体时，特别是领域本体，构造过程繁杂，因此必须要具体问题具体分析，同时严格遵循构造准则，其中最有影响的是 Gruber 于 1995 年提出的 5 条准则[8]。即：

（1）清晰性（Clarity）、明确性和客观性，即本体应该用自然语言对所定义的术语给出明确的、客观的语义定义，即必须有效地说明所定义术语的意思。而且，当定义可以用逻辑公理表达时，它应该是形式化的。

（2）完全性(Completeness)，即所给出的定义是完整的，完全能表达所描述术语的含义。

（3）一致性 (Coherence)，即由术语得出的推论与术语本身的含义是相容的，即支持与其定义相一致的推理，不会产生矛盾: 所定义的公理以及用自然语言进行说明的文档也应该是一致的。

（4）可扩展性(Extendibility)，即向本体中添加通用或专用的术语时，不需要修改其已有的概念定义和内容，即支持在已有的概念基础上定义新的概念。

（5）最小承诺(Minimal ontological comitment) 和最小编码偏好(Minimalencoding bias)，所谓最小承诺，即本体约定应该最小，因此只定义最必要的术语和约束最弱的关系。所谓最小编码偏好，即不需要指定术语形式化用何编码。

3.5 领域本体构建方法和步骤

本体是描述概念和概念间关系的模型。因此，在领域本体模型构建时，首先要考虑到本体的组成部分。通常情况下, 一个领域本体由以下几个方面组成, 即该领域本体的层次体系、对应属性及属性的取值范围、本体层次间的语义关系、层次之间的推理规则。当前比较前沿和有效的构建领域本体方法是丁晟春、李岳盟[9]等提出的基于顶层本体构建领域本体的构建方法。该方法从本体工程方法论的成熟度和领域本体构建的特点出发, 借鉴了 Mike UschoId & King 的“骨架”法和斯坦福大学的“七步”法, 并融合了叙词表和顶层本体资源, 对概念体系的规范化校

验和本体的标准化处理提出了具体的方法和步骤。

该研究方法的核心思想是, 从本体工程的基本思想出发, 借助词表法对选词进行规范化处理, 并选择合适的顶层本体, 对领域本体构建进行标准化处理, 最后将领域本体嫁接入顶层本体中。基于顶层本体的领域本体构建框架如图 3 所示。

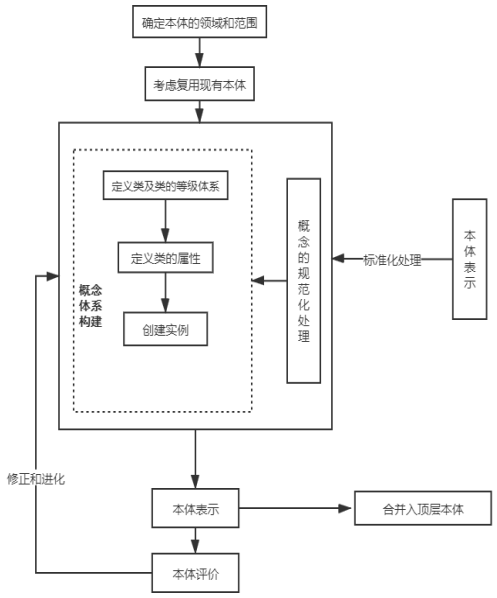

图 3 基于顶层本体的构建方法框架

根据以上构建框架，基于顶层本体构建领域本体的步骤为：

（1）确定本体的领域和范围。

（2）考察复用现有本体。

（3）本体概念体系的构建。

（4）概念的规范化处理。

（5）借鉴顶层本体进行概念体系的标准化处理。

（6）本体表示。

（7）基于顶层本体构建领域本体。

（8）本体评价。

4 基于领域本体信息检索的编辑距离方法

编辑距离是指将一个字符串转换成另一个字符串所需要进行插入、删除、替换等相关编辑操作的次数。两个字符串之间的相似性是通过操作次数来体现的。设两个字符串间的编辑距离为 d,m 为相比较的两个字符串的字符串长度最大值，则两字符串的字形相似度为 $\mathrm{Sim}=\frac{d}{m}\times 100\%$。为了更好地找出与目标字符串字形相似的字符串，可设定一个相似度阈值 r_0，取 $\mathrm{Sim}\le r_0$ 的字符串为最佳匹配。其中，编辑距离 d 可采用动态线性规划方法来计算。设 S= $s_1 s_2 \cdots s_n$ 为待匹配的字符串,而 W= $\omega_1\omega_2\cdots\omega_m$ 为目标字符串,其中 n,m 分别为 S 和 W 的字符串长度。动态计算 S 和 W 的编辑距离的过程可以通过一个 n x m 的矩阵来描述。这个矩阵最右下角的值即为所求的 S 和 W 之间的编辑距离 d。该矩阵元素的计算方法为：

设 $D(s_i,w_j)$ 为 S 中前 i 个字符串和 W 中前 j 个字符串间的编辑距离，是矩阵中的一个值。置初值 $D(s_0,w_0)$ 为 0，则：

$$D(s_i,w_j)=\min\begin{cases} D(s_{i-1},w_{j-1}), & \text{当 } s_i \text{ 与 } w_j \text{ 相同时} \\ D(s_{i-1},w_{j-1})+1, & \text{当在 } S \text{ 中用 } w_j \text{ 替换 } s_i \text{ 时} \\ D(s_i,w_{j-1})+1, & \text{当在 } S \text{ 中插入 } w_j \text{ 时} \\ D(s_{i-1},w_j)+1, & \text{当在 } S \text{ 中删除 } w_j \text{ 时} \end{cases}$$

5 结语

近年来，随着互联网的迅速发展，网络上充斥着各种各样的信息，资源类型数不胜数，当我们想去查某个事物的信息时，搜索引擎返回给我们的信息中，绝大部分是与我们的诉求不相关的。随着本体论的发展，基于本体技术的信息检索模型成为信息检索领域的研究热点，同时 在本体中加入相应的语境，形成领域本体模型，有效改进了传统的基于关键词匹配检索技术在查全率和查准率方面的不足。除此之外，许多方面还有待进行更深入的研究，但由于本人水平有限，还需要进行更加深入地学习。

参考文献

[1]郭维威,褚洪波,李晓艳,刘锋,田铁刚,尹衍林.领域本体模型构建与信息检索方法研究[J].时代农机,2016,43(01):93-94.

[2]道吉草. 基于本体的IFC构件信息检索方法研究[D].上海交通大学,2019.

[3]李彦.一种基于元数据本体计算的网络信息检索方法[J].新技术新工艺,2015(03):41-

43.
[4]刘锋,郭维威.一种优化的基于领域本体语义距离的概念相似度计算模型研究[J].曲阜师范大学学报(自然科学版),2015,41(04):55-59.
[5]孙雪,黄志球,沈国华,王金永,徐恒.基于本体和 BN 的无人车行为决策方法[J/OL].系统工程与电子技术:1-17[2020-12-04].
[6]于碧辉,孙思,李岳.面向电网安全监测的领域本体自动构建[J].计算机系统应用,2020,29(11):243-249.
[7] Nadia Lachtar. 2018. Using Domain Ontology to Classify a Song. In Proceedings of the 7th International Conference on Software Engineering and New Technologies (ICSENT 2018). Association for Computing Machinery, New York, NY, USA, Article 26, 1–5.
[8] T. R Gruber. Towards Principles for the De sign of Ontologies Used for Knowledge Shari ngInternational Joumal of Human Computer Studies, 1995, 43:907-928.
[9]丁晟春,李岳盟,甘利人.基于顶层本体的领域本体综合构建方法研究[J].情报理论与实践,2007(02):236-240.