《智能信息处理》课程考试

基于新闻领域本体的检索方法

张新健

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

2021 年 12 月 18 日

DOI:10.7544/issn1000-1239.2021.20190785
58(1):1-6,2021

基于新闻领域本体的检索方法

张新健[1] 李冠宇[1,2]

[1]（大连海事大学信息科学技术学院 辽宁省大连市 116026）

[2]（智能信息处理教育部重点实验室（大连海事大学）辽宁省大连市 116026）

（zxj@dlmu.edu.cn）

Retrieval method based on news domain ontology

Zhang xinjian[1] , Li Guanyu[1,2]

[1]（*College of Computer Science and Technology, Dalian Maritime University, Dalian City, Liaoning Province* 116026）

[2]（*Key Laboratory of Intelligent Information Processing Ministry of Education*（*Dalian Maritime University*）, *Liaoning Province* 116026）

Abstract When the semantic web information retrieval system is put into operation, there are some problems in the stand-alone environment, such as limited storage capacity and slow multi-user concurrent query speed. To solve this problem, a distributed Semantic Web retrieval method based on news domain ontology is proposed. Firstly, according to the characteristics of the news domain, referring to the seven step method and skeleton method, the news domain ontology is constructed, and the hybrid semantic similarity algorithm suitable for ontology is studied for semantic expansion. Then they are labeled and the concepts with higher mixed semantic similarity are more similar, that is, they appear at the same time in retrieval. Experimental results show that this method effectively reduces the response time of query keywords and improves the recall and precision of news retrieval.

Key words ontology; News retrieval; Query response

摘要 当语义 Web 信息检索系统投入实际运行时，单机环境存在存储容量有限和多用户并发查询速度慢等问题。针对此问题，提出了基于新闻领域本体的分布式语义 Web 检索方法。首先依据新闻领域的特点，参照七步法和骨架法，构建新闻领域本体，研究适合本体的混合语义相似度算法进行语义扩展。然后将其标注且认为混合语义相似度更高的概念更相似，即在检索时同时出现。实验结果表明，该方法有效地减少了查询关键词的响应时间，提高了新闻检索的查全率和查准率。

关键词 本体；新闻检索；查询响应

中图法分类号 TP391

收稿日期：2019-11-08；**修回日期**：2020-04-16

基金项目：国家自然科学基金项目（60903098）

This work is supported by the National Natural Science Foundation of China (60903098).

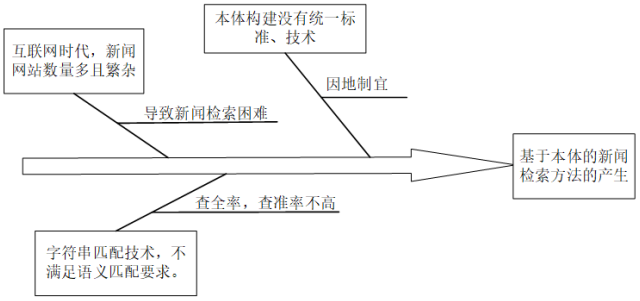
通信作者：李冠宇（liguanyu@dlmu.edu.cn）

由于新闻传播的研究持续深入和对研究成果的普及推广，记载新闻的文字得到了快速网络化，促使传统 Web 向语义 Web 发展，导致语义 Web 的数据量迅速增长。语义 Web 作为当下研究的热点，在西文和中文领域发展迅速，语言作为文化的载体，在文化的传承和研究中起着重要的作用。

随着信息检索技术的不断发展，但是由于新闻网站数量多且繁杂，使得人们在众多网站中查找自己需要的内容很困难。

本文利用构建本体的原则和方法以及本体的相关理论[1][2]构建新闻领域的本体，实现对新闻领域信息检索。

因果链图如下：



1 相关工作

本体[3]的概念最初起源于哲学领域，是共享概念模型的明确形式化规范说明。它的目标是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，并从不同层次的形式化模式上给出这些词汇（术语）和词汇之间相互关系的明确定义。本体的概念有四层含义：概念化(conceptualization)、形式化(formal)、明确(explicit)、共享(share)。它是一种能在语义和知识层次上描述信息系统的概念模型。本体应用的基础是本体的构建。

1.1 本体构建需求分析与方法

(1) 需求分析

新闻是当下各用户获取信息最直接的载体，而近几年随着政府对新闻推广力度的加大，新闻逐渐增多，新闻网站也层出不穷，从未来发展的角度来看，对这些网络资源进行集成处理将是网络建设者亟需解决的新课题。

传统的基于关键字的信息检索技术会产生“词语孤岛”问题，这种方法往往存在检索出的结果不全、不完整和质量不高的问题，从而无法满足用户的需求，用户要在海量数据中找到自己需要的信息困难较大。而语义 Web 信息检索技术可有效提高获取信息的查准率和查全率，减少信息搜索的时间，是满足用户检索需求的好方法。语义 Web 的技术核心是领域本体，通过对本体网的语义相似性计算，将关键词进行语义扩展之后再进行检索，就可以在一定程度上提升检索的查全率和查准率。

(2) 本体构建方法

目前为止，本体构建仍没有统一的标准，在构建领域本体的过程中，还是需要有领域专家的参与和协同工作，即需要人工构建本体。对于研究人员或者用户来说，构建本体的方法有很多种，没有特定的方法去建立本体，且不同领域的本体语义不相同。对于任意一个学科领域，应针对学科领域的特点和需求来选择某一种适合的方法。本文选择七步法结合骨架法来构建新闻领域本体。

七步法：

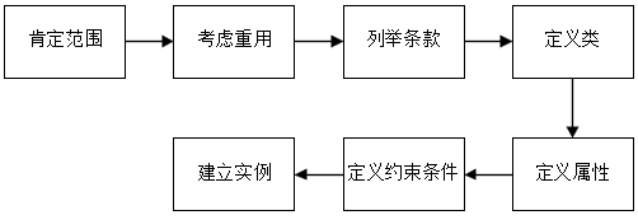


图 1.1 七步法流程

骨架法：

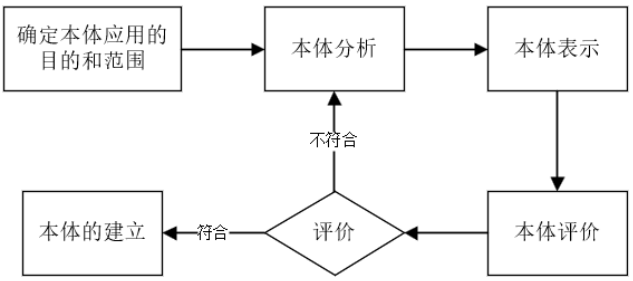


图 1.2 骨架法流程

技术路线如下图所示：

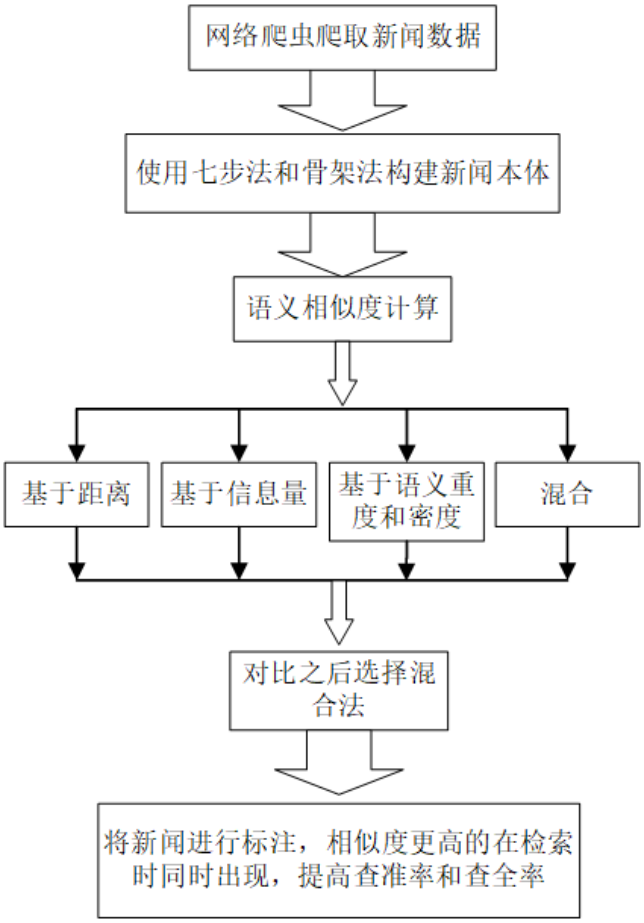


图 1.3 技术流程图

2 确定领域概念及本体层次结构

2.1 领域本体概念之间的关系及层次结构

在确定新闻领域核心概念之前，首先使用网络爬虫收集各新闻网站数据，分析、确定不同概念之间的逻辑关系。在新闻领域，类与类之间不仅存在父类与子类的继承关系、上下位关系和包含关系，还有兄弟类之间的相似关系和平行关系，也有类之间的成员关系和部分整体关系。再结合国际出版电讯委员会设计的三层结构的新闻主题分类模型，通过词频统计找出呈现频率最高的 5 个核心概念为经济、政治、社会、文化和教育，并作为本体的 5 大类 1 级概念，其结构如图 2.1 所示。

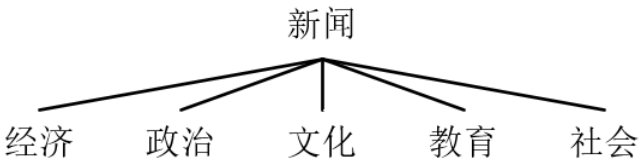


图 2.1 新闻本体 1 级概念结构图

依据类之间的上述关系，确定经济概念下的 2 级至 4 级概念本体结构如图 2.2 所示。

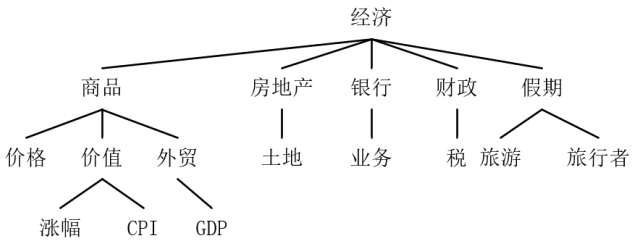


图 2.2 经济概念下的本体概念结构图

确定政治概念下的 2 级至 4 级概念 本体结构如图 2.3 所示。

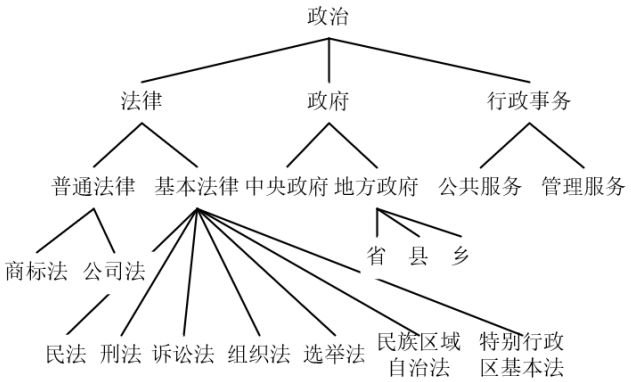


图 2.3 政治概念下的本体概念结构图

确定社会概念下的 2 级至 3 级概念本体结构如图 2.4 所示。

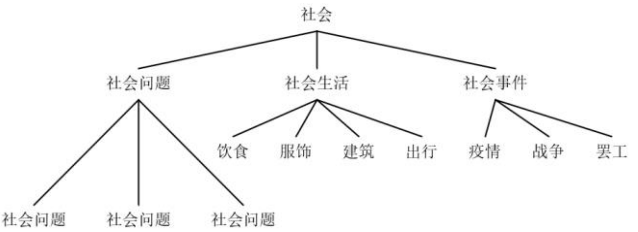


图 2.4 社会概念下的本体概念结构图

确定文化概念下的 2 级至 5 级概念本体结构如图 2.5 所示。

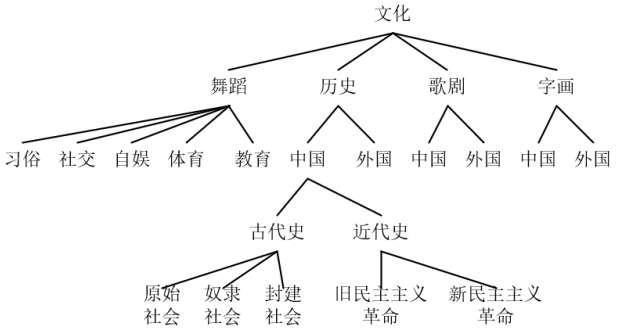


图 2.5 文化本体 1 级概念结构图

确定教育概念下的 2 级至 3 级概念本体结构如图 2.6 所示。

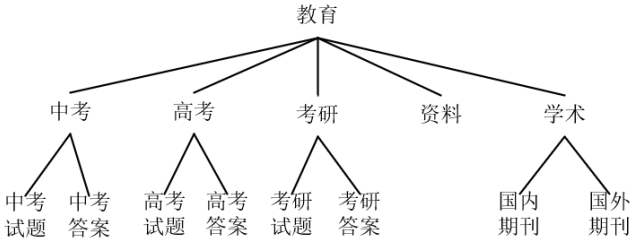


图 2.6 教育本体 1 级概念结构图

2.2 定义类的属性

本体概念也可以看作是一个类。本体概念建立之后，就要确定类属性。属性既可以表示类和类之间的差异，也可以体现类之间的关系。确定类的属性可以减少类的重复定义，减少数据冗余。定义了属性之后，可以对不同概念类进行描述和归纳，子类要继承父类的所有属性，还具有自己的一些新特性。通过分析，蒙古语新闻领域类和类之间大多表现为子类和父类、兄弟类的关系，少量表现为同义类、等价类的关系，类和类之间的关系让领域本体概念之间具有了关联性。新闻领域共有的数据属性有：标题、正文、时间、作者、来源和链接等，数据属性用来描述类所具有的数据。本体类的主要对象属性有：父节点、子节点、部分、兄弟节点、相似概念和相同概念，对象属性可明确不同类之间的关系。

3 语义相似度计算

语义相似度是衡量词语之间关系的一个重要指标，也是语义 Web 中最重要的一部分。语义相似度是指几个词语或词语之间可能替代或相互交换的程度。通过语义相似度的计算结果实现隐藏概念的扩充和语义的扩展，利于检索关键词的语义解析和检索范围的扩大，从而提高检索的查全率和查准率[4-6]。

3.1 基于语义距离的语义相似度计算

首先依据本体的层次结构，从顶级概念开始，依自上而下、从左到右的次序赋予每个概念节点的 id 序号为 $1,2,3,\ldots$ 。在本体层次结构中用路径长度和概念之间的深度差来表示概念之间的语义距离。概念之间的语义距离和语义相似度成反比，当 2 个概念之间有较大的语义距离时，它们的相似度较低；相反，则它们的相似度较高。语义距离与路径长度和深度有关，用 $Distance(a,b)$ 表示在本体中概念 a 到达概念 b 中所需要的最少的边数。$Depth(a,b)$ 表示概念 a 与概念 b 的深度差，计算深度值又与自身深度和共有祖先深度有关，2 个概念深度差越大，说明 2 个概念距离越远，语义相似度越低；反之，语义相似度较高。用式(1)表示为：

$$Depth(a,b)=\frac{d\big(NCA(a,b)\big)}{d(a)+d(b)-d\big(NCA(a,b)\big)} \tag{1}$$

其中，$d(NCA(a,b))$ 代表共同祖先深度，$d(a)$ 代表概念 a 的深度，$d(b)$ 代表概念 b 的深度。令根节点新闻的深度为 1，则 4 层本体结构的概念深度最大为 5。任意 2 个概念 a 和 b 之间的语义距离 $Distance(a,b)$ 应介于 $0\sim 2*Depth(T)$，$Depth(T)$ 是树的最大深度。再考虑到距离和深度的关系应是同增同减，那么基于距离深度的概念 语义相似度计算如式(2)所示为：

$$Sim_{distance(a,b)}=\frac{2*Depth(a,b)}{2*Depth(a,b)+Distance(a,b)} \tag{2}$$

其中，$Sim_{distance(a,b)}\in[0,1]$ 。

3.2 基于信息量的语义相似度计算

在本体层次结构中，对于父节点来说，它的每一个概念子节点都是对其概念的细化和具体化，所以可以通过比较概念之间所包含的信息量和公共祖先概念节点的信息量来衡量概念之间的相似度。信息量的计算方式如式(3)所示：

$$IC_{(c)}=-1bP \tag{3}$$

其中 P 代表概念节点 c 出现的概率，用概率来表示概念的信息量。概率 P 的计算如式(4)所示：

$$P=\begin{cases}\dfrac{1}{countleaves}, \text{当}p\text{为叶子节点}\\ \displaystyle\sum_{i=1}^{C(p)}P(c_i), \text{当}p\text{为非叶子节点}\end{cases} \tag{4}$$

其中，$countleaves$ 是叶子节点的个数。当 p 为叶子节点时，p 出现的概率就是总叶子节点数的倒数；当 p 为非叶子节点时，p 出现的概率就是 p 的子节点出现的概率之和。其中 $C(p)$ 是 p 节点的子节点个数，c_i 是 p 节点的第 i 个子节点，而 $P(c_i)$ 是 c_i 出现的概率。根据分析可知，概念 a 和 b 信

息量的相似度也取决于最近共同祖先节点的信息量，所以本文基于内容的语义相似度计算方式如式（5）所示：

$$Sim_{ic(a,b)} = \frac{2 * IC\big(NCA(a,b)\big) + \sigma}{IC_{(a)} + IC_{(b)} + \sigma} \tag{5}$$

其中，σ 是平衡因子，以确保分子不为 0。经过实验得知，平衡因子取 0.5 时，语义相似度计算的结果与主观判断的结果相符度高。

3.3 基于语义重合和语义密度的语义相似度计算

语义重合度是指概念拥有相同的祖先节点的个数，相同祖先节点越少，表明它们不在一个分支，则相似度越低；反之，语义相似度越高。而语义密度是指概念所拥有的子节点的个数。在本体层次中不同分支节点拥有的子节点的数量不同。如果在本体中，某一概念的节点密度越大，说明对该节点概念的具体化、细化程度越高，语义相似度越高。基于语义重合和语义密度的语义相似度计算如式（6）所示：

$$\begin{aligned} Sim_{pro(a,b)} = {} & Count(a \cap b) + Dendity(a \cap b) \\ & + Count(T) + Density_{max}(T)/2 \\ & * \big(Count(T) - Count(a \cap b) \\ & + Density_{max}(T) \\ & - Dendity(a \cap b)\big) \end{aligned} \tag{6}$$

其中，$Count(a \cap b)$ 代表概念 a 与 b 的共同父节点数；$Density(a \cap b)$ 代表概念 a 与 b 的共同子节点数；$Count(T)$ 则是整棵树的父节点数；$Density_{max}(T)$ 则是整棵树密度最大的节点数，也就是拥有子节点数最大的那个概念的子节点数。

3.4 混合式语义相似度计算

本文采用混合式语义相似度计算方式，根据实际情况进行混合，同时考虑了概念词的位置信息、概念词的信息量和概念词的密度等，如式（7）所示：

$$\begin{aligned} Sim_{(a,b)} = {} & \alpha * Sim_{distance(a,b)} + \\ & \beta * Sim_{ic(a,b)} + \gamma * Sim_{pro(a,b)} \end{aligned} \tag{7}$$

因为每一个因素对语义相似度的影响不同，所以所占的比例也不同，故要对 α、β、γ 这 3 个参数的值进行实验确定。已知 $\alpha + \beta + \gamma = 1$，用函数生成随机的 60 组满足条件的 α、β、γ。通过实验得到，当 $\alpha = 0.35, \beta = 0.2, \gamma = 0.45$ 时，既可以表达出父子节点的关系，也可以看出距离、信息量和语义密度以及重合度对语义相似度的影响。

表 2 中给出了 4 种语义相似度计算方法的实验结果。

本体概念a	本体概念 b	相似度			
		基于距离	基于信息量	基于语义重度和密度	混合
经济	政治	0.25	0.344	0.533	0.379
经济	社会	0.25	0.285	0.533	0.370
经济	文化	0.25	0.326	0.533	0.376
经济	教育	0.25	0.326	0.533	0.376
经济	房地产	0.57	0.611	0.605	0.564
经济	银行	0.57	0.589	0.725	0.615
经济	财政	0.57	0.468	0.533	0.542
经济	法律	0.14	0.259	0.533	0.329
经济	政府	0.14	0.259	0.533	0.329
经济	行政事务	0.14	0.258	0.533	0.324
经济	社会问题	0.14	0.224	0.533	0.324
经济	社会生活	0.14	0.224	0.533	0.324
经济	社会事件	0.14	0.224	0.533	0.324
经济	舞蹈	0.14	0.224	0.533	0.324
经济	历史	0.14	0.224	0.533	0.324
经济	歌剧	0.14	0.224	0.533	0.324
经济	旅游	0.33	0.468	0.533	0.443

首先看基于距离的语义相似度，经济与政治、文化、教育、社会是兄弟关系，根据距离与深度计算的结果都是 0.25，因为社会含有的子节点比较多，所以说经济和社会的相似度应该比其他兄弟本体概念的语义相似度高；再看基于信息量的计算结果，当 2 个本体概念所包含的子节点个数相同但处于不同的分支时，比如经济和财政，经济和旅游，它们基于信息量的语义相似度一样，所以该方法没有办法考虑到位置信息；房地产和银行均是经济的子节点，法律、政府和行政事务是经济兄弟节点的子节点，而根据语义重合度的计算结果，无法判断哪个本体概念是兄弟节点，哪个本体概念是兄弟节点的子节点，所以也不适用于本文的本体。经过对表 2 的分析，把数据结果大于 0.5 的

本体概念取出，认为本体概念 a 与 b 相似。由实验结果得出，基于语义距离深度的语义相似度计算方法简单，易于实施，但是不能很好地体现节点密度和节点信息量之间的关系。基于信息量的语义相似度计算相对比较客观，能综合反映本体概念之间的相似性和差异，但是又不能完全分辨 2 个本体概念的位置信息。基于语义重合和语义密度的语义相似度计算必须依靠具有完备概念的概念集，本体概念语义重合和语义密度越大语义相似度越高，但不能反映概念节点距离和信息量之间的差异。混合式的计算方法是将前 3 种计算方法混合起来，并且把每一个因素按照不同占比相加，最后所得出的语义相似度结果比单一因素的语义相似度结果更加准确。所以，本文采用混合式的语义相似度计算方法来对查询关键词进行语义扩展，然后在已经标记索引的本体实例库中进行搜索，从而得到所需的查询结果，最终充分提高检索的查全率和查准率。

结束语

面对日益增长的新闻网站，如果能快速、准确、全面地获取到需要的新闻网络信息资源，就能很好地满足人民在新闻领域中对信息检索的大量需求。本文在建立新闻领域本体的基础上，研究合适的混合式语义相似度计算方法进行语义扩展，有效提高了新闻的检索速度，同时也提高了新闻关键词查询的查全率和查准率。

参 考 文 献

[1] 李景，孟连生．构建知识本体方法体系的比较研究[J]．现 代 图书情报技术，2004，9(7)：17-22

[2] 李善平，尹奇韡，胡玉杰，等．本体论研究综述[J]．计 算 机 研 究与发展，2004，41(7)：1041-1052

[3] Formica A .Ontology-based concept similarity in formalconcept analysis , Information Sciences , 176(2006) , 2624~2641

[4] 朱征宇，孙俊华．改进的基于《知网》的词汇语义相似度计算[J]．计算机应用，2013，33(8)：2276-2279．

[5] 张思琪，邢薇薇，蔡圆媛．一种基于 WordNet 的混合式语义相似度计算方法[J]．计算机工程与科学，2017，39(5)：971-977．

[6] 李峰，李芳．中文词语语义相似度计算—基于《知 网》2000[J]．中文信息学报，2007，21(3)：99-105

[7] 赵俊生,王鑫宇,尹玉洁,张林.基于蒙古语新闻领域本体的分布式检索方法[J].计算机工程与科学,2021,43(03):560-570.

Zhang Xinjian, born in 1998. M.S. His main research interests include Web mining, information retrieval, machine learning.

张新健，1998 年生，硕士。主要研究方向为 Web 挖掘、信息检索和机器学习。

Guanyu Li He is currently a professor in Dalian Maritime University, China. His primary research interests are in Semantic Web, Ontology Engineering, Internet of Things, Knowledge Graph, etc. He has published more than 60 papers in refereed journals and conferences

李冠宇 现任大连海事大学教授。 主要研究方向为语义网、本体工程、物联网、知识图谱等，在权威期刊和会议上发表论文 60 余篇。