第 32 卷第 5 期　　河 北 科 技 大 学 学 报　　Vol. 32, No. 5
2011 年 10 月　　Journal of Hebei University of Science and Technology　　Oct. 2011

文章编号:1008-1542(2011)05-0471-06

基于本体的多专业领域知识语义检索系统研究

刘紫玉[1],杨国霞[2],李学会[2]

(1. 河北科技大学经济管理学院,河北石家庄　050018;2. 河北科技大学信息科学与工程学院,河北石家庄　050018)

摘　要:多专业领域是指由多个专业构成的学科领域,如高速铁路领域由工务工程、牵引供电、动车组、运营管理等专业领域构成。对于多专业领域本体的构建,可以先构建各个专业领域本体,然后根据专业领域之间的关系进行本体集成。面向多专业领域,在本体模型的基础上提出了多专业领域本体模型,并给出了多专业领域本体概念语义相似度计算模型,此模型可作为语义扩展的基础。在此基础上,设计了一个基于本体的多专业领域知识语义检索系统,并以高速铁路领域为背景对提出的语义检索系统进行实验验证,从测试结果来看,开发的语义检索系统可以获取较高的准确率和召回率。

关键词:本体;多专业领域;语义检索;高速铁路领域

中图分类号:TP311　　**文献标志码**:A

Research in semantic retrieval system for knowledge of multiple majors domain based on ontology

LIU Zi-yu[1], YANG Guo-xia[2], LI Xue-hui[2]

(1. College of Economics and Management, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

Abstract: In this paper a domain including different major fields is called multiple majors domain. For example, the high-speed railway domain consists of maintenance engineering, traction power supply, EMU and operation management, etc. Ontologies of the major fields are built based on thesaurus and thematic words, and these ontologies are integreted into a unified ontology for the multiple majors domain. Oriented to the domain that consisting of several major fields, this paper gives the ontology model for multiple majors domain, and also builds the model to compute semantic similarity between concepts for multiple major domain. Then, this paper puts out a semantic retrieval system for the knowledge of multiple major domain based on ontology and verifies it. The experimental results show that the developed semantic retrieval system can reach satisfying recall and precision.

Key words: ontology; multiple majors domain; semantic retrieval; high-speed railway domain

本体在知识管理、人工智能、信息检索、Web 服务发现等领域中扮演着越来越重要的角色。根据领域依赖程度,可以将本体分为通用本体、领域本体、任务本体和应用本体[1]。领域本体可以有效地组织领域中的

收稿日期:2011-03-05;修回日期:2011-09-12;责任编辑:李　穆

基金项目:河北省科技支撑计划项目(11213504D);河北省教育厅科学技术研究项目(Z2011275);河北科技大学博士科研基金资助项目(QD201017)

作者简介:刘紫玉(1975-),女,河北赵县人,讲师,博士,主要从事知识管理、本体等方面的研究。

知识，使知识更好地共享、重用。

对于多专业领域本体的构建，可以先构建各个专业领域本体，然后根据专业领域之间的关系进行本体集成。在本文中作以下定义：由多个专业领域构成的领域定义为“多专业领域”，如高速铁路领域；基于每个专业领域构建的本体称为“专业领域本体”，如动车组专业领域本体；由多个专业领域本体集成的统一本体称为“多专业领域本体”，如将高速铁路的动车组专业领域本体、牵引供电专业领域本体等集成后的本体称为“高速铁路多专业领域本体”。多专业领域本体构建方法本文不进行详细说明，可以参见相关作者的文章[2]。

笔者面向多专业领域，在本体模型的基础上提出了多专业领域本体模型，并给出了多专业领域本体概念语义相似度计算模型，此模型可作为语义扩展的基础。在此基础上，设计了一个基于本体的多专业领域知识语义检索系统。最后以高速铁路领域文献资料知识作为实验对象，对本文提出的语义检索系统进行实验分析，从测试结果来看，开发的语义检索原型系统可以获取较高的准确率和召回率。

1 概念语义相似度计算模型

1.1 本体模型

关于本体(Ontology)的定义有许多，目前获得较多认同的是 STUDER 等的解释[3]：“Ontology 是对概念体系的明确的、形式化的、可共享的规范说明”。

定义 1 一个完整的本体应由概念、关系、函数、公理和实例等 5 类基本元素构成。本体可以表示为

$$O=\{C,R,F,A,I\}。$$

其中：C 为概念，概念是指客观世界中任何事物的抽象描述，在本体中通常按照一定的关系形成一个层次结构；

$R\subseteq 2^{C\times C}$，概念之间的关系，如“subclass-of”关系、“part-of”关系等；

$F\subseteq R^n$ 是一种特殊的关系，其中第 n 个元素 c_n 相对于前面 $n-1$ 个元素是唯一确定的，函数 F 可以表示为 $c_1\times c_2\times\cdots\times c_{n-1}\rightarrow c_n$；

A 为概念或者概念之间的关系所满足的公理，是一些永真式；

I 为领域内概念实例的集合。

1.2 领域本体模型

在实际的领域本体中，由于概念之间不仅仅存在着上下位关系，概念之间还通过其他各种关系可以连接，尤其在多专业构成的领域本体中还有许多自定义的关系，这使得概念的组织形式并不完全是一个树型结构，而是一个网状结构。因此，根据多专业领域本体的特点，在本体模型的基础上重新构建了领域本体模型。

定义 2 领域本体模型是一个八元组：$DO=\{C,P,H^c,R^s,R^{ud},I,F,A\}$。其中：$DO$ 表示领域本体；C 表示概念(或称为类)；P 表示领域本体中 Datatype 类型属性；H^c 表示类间的上下位(subclass-of)二元关系；R^s 表示类间的同义(synonymy)关系；R^{ud} 表示类间的用户自定义(user-defined)关系(包括 part-of 关系也用自定义关系来描述)，也就是类的 ObjectProperty；I 表示领域内概念实例的集合；F 表示概念间一种特殊的关系，可以表示为 $c_1\times c_2\times\cdots\times c_{n-1}\rightarrow c_n$；$A$ 表示领域本体中概念或者概念之间的关系所满足的公理，是一些永真式。

定义 3 概念 C 的模型是一个九元组：$C=\{P,C_{sc},C_{uc},C_s,C_r,H_c,R_s,R_{ud},I_c\}$。其中：$P$ 表示概念 C 的 Datatype 类型属性；C_{sc} 表示概念 C 的子概念(subclass)；C_{uc} 表示概念 C 的父概念(upperclass)；C_s 表示概念 C 的同义概念(equivalentclass)；C_r 表示与概念 C 有关系的概念；这里主要指通过用户自定义关系联系起来的概念；H_c 表示概念 C 的上下位关系；R_s 表示概念 C 的同义(synonymy)关系；R_{ud} 表示概念 C 的用户自定义(user-defined)关系；I_c 描述概念 C 的实例。

概念之间的关系主要分为 3 类：1)上下位关系，用 C_{sc}，C_{uc} 和 H_c 表示；2)同义关系，用 C_s 和 R_s 表示；3)用户自定义关系，用 R_{ud} 表示。

1.3 概念语义相似度计算模型

1.3.1 模型组成描述

传统本体概念间相似度计算的不足在于其语义关系只考虑了层次语义关系，没有考虑语义关系中非层次关系的影响，同时对象实例对于概念的影响也没有考虑。笔者在定义 3 的基础上，提出了计算概念之间相

似度的模型，该模型全面考虑了本体概念模型中各种元素对相似度的影响，考虑的元素主要包括属性（Datatype 类型属性）、上下位语义关系、其他语义关系（自定义关系）和实例特征。

1.3.2　MD4 模型概念语义相似度算法

在同一本体中，概念相似度计算首先需要检查 2 个概念是否同义。如果 2 个概念同义，那么 2 个概念是完全相似的，其相似度为 1。

1）上下位关系语义相似度计算

在领域本体中，只考虑上下位关系时的本体模型为树型结构。计算上下位关系语义相似度时采用基于距离的概念相似度计算方法。笔者参考陈杰等人的算法，综合考虑概念距离和层次对概念相似度的影响[4]，算法公式如下：

$$\mathrm{Sim_h}(C_i,C_j)=\frac{\alpha\times(\mathrm{dl}(C_1)+\mathrm{dl}(C_2))}{(\mathrm{Dist}(C_1,C_2)+\alpha)\times 2\times \mathrm{Max dl}\times\max(|\mathrm{dl}(C_1)-\mathrm{dl}(C_2)|,1)}\quad 。\tag{1}$$

式中 $\mathrm{dl}(C_1)$ 和 $\mathrm{dl}(C_2)$ 分别是 C_1 和 C_2 所处的层次；$\mathrm{Dist}(C_1,C_2)$ 是概念 C_1 和 C_2 之间的本体树中的最短路径；Max 样 dl 是指本体树的最大深度，在这里除以该参数是便于计算结果的归一化处理；α 是一个可调节参数，一般 $\alpha\geqslant 0$。

2）自定义关系语义相似度计算

假设有 2 个非同义概念 C_i 和 C_j，根据定义 3 中的概念模型表示方法，可得到概念 C_i 对应的 p 个自定义关系集 $R_{\mathrm{ud}i}$ 和 p 个自定义关系对应的 m 个概念集 C_{ri}，概念 C_j 对应的 q 个自定义关系集 $R_{\mathrm{ud}j}$ 和 q 个自定义关系对应的 n 个概念集 $C_{\mathrm{r}j}$。这里，每个集合中不存在相同的元素。

当 2 个自定义关系进行比较时，如果 2 个关系是相同的，那么相似度为 1，否则相似度为 0。

自定义关系相似度计算公式如下：

$$\mathrm{Sim_{ud\_r}}(C_i,C_j)=\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}\mathrm{Sim}(r_{\mathrm{ud}i},r_{\mathrm{ud}j})}{\max(p,q)},\tag{2}$$

其中 p 和 q 分别是概念 C_i 和 C_j 对应的自定义关系的个数。

自定义关系对应的概念之间的相似度计算使用式（1），综合相似度计算公式为

$$\mathrm{Sim_{ud\_c}}(C_i,C_j)=\frac{\sum_{i=1}^{m}\max(\sum_{j=1}^{n}\mathrm{Sim}_h(C_{ri},C_{\mathrm{r}j}))}{m},\tag{3}$$

其中 m 是 C_i 的 p 个自定义关系对应的概念个数，n 是 C_j 的 q 个自定义关系对应的概念个数。

在领域本体中，C_i 和 C_j 通过自定义关系体现出的相似度 $\mathrm{Sim_{ud}}(C_i,C_j)$ 为

$$\mathrm{Sim_{ud}}(C_i,C_j)=\beta\mathrm{Sim_{ud\_r}}(C_i,C_j)+\gamma\mathrm{Sim_{ud\_c}}(C_i,C_j),\tag{4}$$

其中 β,γ 分别表示 2 种相似度的权重（可简单设定 $\beta=\gamma=0.5$），$0\leqslant\beta\leqslant 1$，$0\leqslant\gamma\leqslant 1$，$\beta+\gamma=1$。

3）概念 Datatype 类型属性相似度计算

当 2 个 Datatype 型的属性进行比较时，如果 2 个属性是相同的，那么相似度为 1，否则相似度为 0。首先确定 C_i 和 C_j 的属性集 P_i 和 P_j，概念 C_i 和 C_j 分别对应 m 和 n 个 Datatype 类型的属性（DatatypeProperty），然后对属性集合 P_i 和 P_j 进行笛卡尔乘积 $P_i\times P_j$，得到配对集，再计算 C_i 和 C_j 的属性相似度 Sim_p，得到 C_i 和 C_j 的属性相似度计算公式为

$$\mathrm{Sim}_p(C_i,C_j)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Sim}(p_i,p_j)}{\max(m,n)},\tag{5}$$

其中，m 和 n 分别是概念 C_i 和 C_j 的 Datatype 类型属性的个数。

4）实例语义相似度计算

实例语义相似度的计算采用和概念 Datatype 类型属性相似度计算相同的算法。C_i 和 C_j 的实例语义相似度计算公式为

$$\mathrm{Sim}_i(C_i,C_j)=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Sim}(i_{C_i},i_{C_j})}{\max(m,n)},\tag{6}$$

式中：m 和 n 分别是概念 C_i 和 C_j 的实例的个数；i_{C_i} 和 i_{C_j} 表示概念 C_i 和 C_j 的某个实例。

5）领域本体中非同义概念实际相似度计算

将上述4种相似度加权综合，得到非同义概念 C_i 和 C_j 的实际相似度计算公式为

$$\mathrm{Sim}(C_i,C_j)=\omega\mathrm{Sim}_{\mathrm{h}}(C_i,C_j)+\theta(\mathrm{Sim}_{\mathrm{p}}(C_i,C_j)+\mathrm{Sim}_{\mathrm{ud}}(C_i,C_j)+\mathrm{Sim}_{\mathrm{i}}(C_i,C_j)), \tag{7}$$

其中 ω，θ 分别表示权重，$0<\omega<1$，$0<\theta<1$，$\omega+\theta=1$，一般 ω 较大。

2 语义检索系统结构

本文设计的语义检索系统[5]分为4个大的功能模块：本体查询、文献语义预处理与概念语义相似度预计算、语义扩展检索和推理检索，具体结构如图1所示。

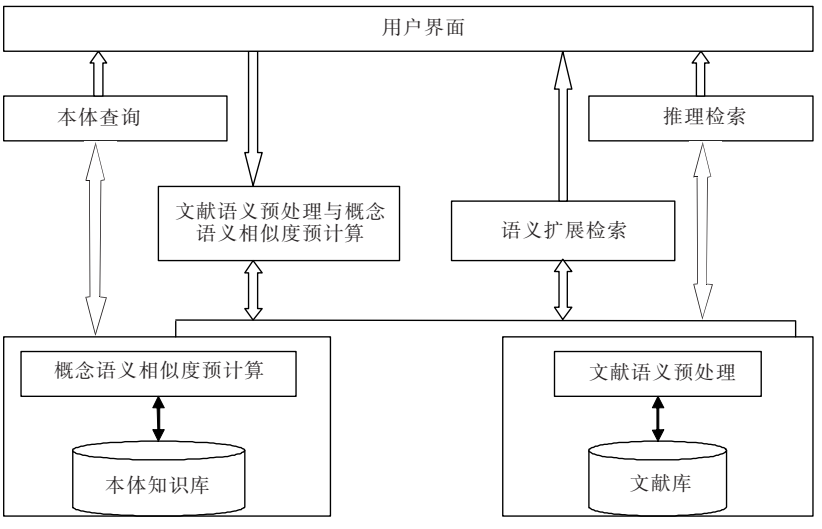

图1 系统结构图

Fig. 1 Architecture of semantic retrieval system

用户界面主要和用户进行交互，系统提供4种功能：第1种是本体查询，可以查询所建本体的概念、概念属性、概念实例等；第2种是语义扩展检索，用户以关键字形式表达查询意图；第3种是推理检索，使用本体规则和公理进行推理检索；第4种是系统给用户提供对文献语义预处理与概念语义相似度预计算进行操作的界面。

本体知识库以OWL文件的形式存储领域本体知识，文献库存储进行语义标注过的领域文献知识。

2.1 本体查询

这一模块的主要功能是使用户可以方便查询本体知识库中所建本体的概念、概念属性、概念实例等。

2.2 文献语义预处理与概念语义相似度预计算

文献语义预处理与概念语义相似度预计算的主要结构如图2所示。该模块主要包括2个部分：文献语义预处理和概念语义相似度预计算。文献语义预处理主要是对文献事先进行语义标注，按照用本体库中定义好的概念对文献进行标引。语义相似度预计算事先对本体库中的概念进行语义相似度计算，根据本文的式(7)进行相似度值的计算，并在本体库中保留相似度值，方便语义扩展检索模块进行查询关键字的扩展。

文献语义标引的最终目的是获得文档的语义向量，对本体解析后可以遍历本体中的概念对一篇文档进行标引，关键是如何确定标引概念对应的权重，即这个概念相对于这篇文档的重要性。过去的研究表明，词频和位置在反映标引词和文献主题的关系上起着重要的作用，笔者采用山西大学郑家恒等人提出的非线性函数和“成对比较法”相结合的方法，综合考虑位置和词频2个因素[6]，最终给出标引概念的权重。对于标引文档的概念和其对应权重，采用一维向量的形式来表示，文献的语义特征向量就由这2个一维向量来表现。

文档语义表示之后的概念向量和权重向量如下：

$$\mathrm{Document}(1)=\{c_{11},c_{12},\cdots,c_{1m}\}, \tag{8}$$

$$\mathrm{Weight}(1)=\{w_{11},w_{12},\cdots,w_{1m}\}。 \tag{9}$$

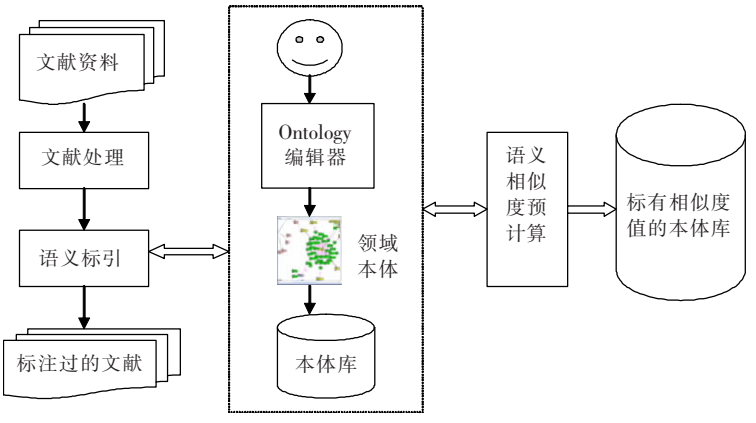

图 2　文献语义预处理与概念语义相似度预计算结构图

Fig. 2　Process structure of semantic precomputation for document and the concept similarity

2.3　语义扩展检索

语义扩展检索模块的主要功能是把用户输入的查询词进行语义扩展，然后把语义扩展向量和从文献语义预处理模块中取出的标引向量进行相似度计算，计算后得到的相似度与用户设立的阈值进行比较，如果大于阈值则文献与查询相关，返回该文档查询结果，并按照相似度大小将排序后的文献列表返回给用户界面。

用户查询词经过概念语义相似度计算扩展以后，语义扩展向量可以用扩展后的概念向量（包括用户输入的查询概念词）和对应的权重向量来表现，这 2 个向量用一维向量的形式表示。

$$\text{Document}(2)=\{c_{21},c_{22},\cdots,c_{2n}\}, \tag{10}$$

$$\text{Weight}(2)=\{w_{21},w_{22},\cdots,w_{2n}\}。 \tag{11}$$

计算文档语义特征向量和用户查询语义扩展向量的语义相似度，本文借鉴的方法，首先计算两两概念之间的语义相似度，然后计算 2 个向量之间的语义相似度[7]。

对于文档语义特征向量 Document(1)中的概念 c_{1i}，其对应的权重为 w_{1i}，用户查询语义扩展向量 Document(2)中的概念 c_{2j}，其对应的权重为 w_{2j}。那么，对于这 2 个概念(c_{1i}, c_{2j})，其相似度计算公式如下：

$$\text{Sim}(c_{1i},c_{2j})=\frac{\log\left(\frac{w_{1i}+w_{2j}}{2}\right)}{\text{distance}(c,c_{1i})+\text{distance}(c,c_{2j})+1}。 \tag{12}$$

其中 distance(c_1, c_2)是 c_1 和 c_2 之间最短路径所包含的边的条数，用于计算 c_1 和 c_2 之间的距离。

最终计算 2 个一维向量的相似度，可以用以下的方法得到：

$$\text{Sim}(\text{Document}(1),\text{Document}(2))=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\text{Sim}(c_{1i},c_{2j})}{m\times n}。 \tag{13}$$

其中，m 是文献语义特征向量的概念向量中概念的个数，n 是用户查询语义扩展向量的概念向量中概念的个数。

2.4　推理检索

本体推理检索是在构建的本体的基础上，获得本体中隐含的知识或推理出需要的知识。推理检索的目的是回答用户问题，并检索出相关文献，按文献语义标注时的相似度排序后提交给用户[8]。推理检索模块直接使用本体中的规则和公理的语义关系进行推理检索，如利用本体中的子类公理（subClassOf）、同义（equivalentClass）等，实现了实例推理查询、实例所属类推理查询和三元组推理查询。

3　实验和结果

高速铁路领域由工务工程、牵引供电、动车组、运营管理等不同的专业领域构成，它是多专业领域的一个代表。在集成后的高速铁路多专业领域本体的基础上，以笔者提出的语义检索系统结构，开发了一个面向高速铁路文档知识的语义检索和推理系统，并对其进行了实验分析。

2个最常用的基于相关性的语义检索系统评价指标分别是准确率(precision)和召回率(recall),为了考察语义扩展检索方法的有效性,采用准确率和召回率作为评测标准。

实验目的:比较语义扩展检索方法(expand)和传统的关键字检索方法(non_expand)。

实验方法:分别用语义扩展检索方法和传统的关键字检索方法,进行10次检索,比较这2种方法的结果。结果如表1所示。

表1 不同方法的召回率和准确率

Tab.1 Recall and precision of different methods

检索项	召回率/%		准确率/%	
	传统方法	语义扩展方法	传统方法	语义扩展方法
1	40.46	49.10	30.65	40.72
2	41.01	44.75	32.33	38.94
3	33.56	52.92	40.00	50.40
4	40.86	48.96	41.60	58.68
5	43.69	59.75	45.90	65.70
6	48.30	53.50	49.56	62.45
7	47.80	54.90	50.78	69.26
8	50.10	51.10	42.34	52.90
9	42.60	50.60	39.52	54.50
10	51.50	60.30	47.56	59.50

通过实验,可以看出语义扩展检索方法在准确率和召回率上要明显优于传统的关键字检索方法。所以这在一定程度上证明了语义扩展检索方法的有效性。虽然人工选择的相关集有一定的不确定性,但这个不确定性也是人机交互系统所不可完全避免的一个问题。

4 结 语

以本体为基石的语义网的出现,克服了传统检索方法的不足,为实现信息检索提供了一种全新的方法,能够大大提高检索的效率和精确度。面向多专业领域,在本体模型的基础上提出了多专业领域本体模型,并给出了多专业领域本体概念语义相似度计算模型,此方法作为语义扩展的基础。在此基础上设计了一个基于本体的多专业领域知识语义检索系统,该系统包括4部分:本体查询、语义扩展检索、推理检索和用户界面。最后以高速铁路领域文献资料知识作为实验对象,对本文提出的语义检索系统进行实验分析,从测试结果来看,开发的语义检索原型系统可以获取较高的准确率和召回率。

参考文献:

[1] 金 芝.知识工程中的本体论研究[A].世纪之交的知识工程与知识科学[C].北京:清华大学出版社,2001.447-465.

[2] 刘紫玉,黄 磊.高速铁路领域本体构建方法研究[J].情报学报(Journal of the China Society for Scientific and Technical Information),2009,28(2):195-200.

[3] STUDER R,BENJAMINS V R,FENSEL D. Knowledge engineering:Principles and methods[J]. Data and Knowledge Engineering,1998,25(1/2):161-197.

[4] 陈 杰,蒋祖华.领域本体的概念相似度计算[J].计算机工程与应用(Computer Engineering and Applications),2006,42(33):163-166.

[5] 孔田野,李万龙,张海鸥.基于药品本体的信息检索系统研究[J].河北科技大学学报(Journal of Hebei University of Science and Technology),2008,29(3):223-226.

[6] 郑家恒,卢娇丽.关键词抽取方法的研究[J].计算机工程(Computer Engineering),2005,31(18):194-196.

[7] WU Jiang-ning,YANG Guang-fei. An ontology-based method for project and domain expert matching[A]. FSKD(2)[C].[S.l.]:[s.n.],2005.176-185.

[8] 董 慧,余传明,徐国虎,等.基于本体的数字图书馆检索模型研究——历史领域知识推理机制[J].情报学报(Journal of the China Society for Scientific and Technical Information),2006,25(6):666-678.