

# 基于领域本体的语义合成研究

林培金<sup>1</sup>, 曹苏燕<sup>2</sup>, 应捷<sup>1</sup>

(1. 南京邮电大学, 江苏 南京 210003;

2. 南京交通职业技术学院 电子信息工程系, 江苏 南京 211188)

**摘要:**为了在检索过程中全面挖掘用户查询信息,文中提出了一种基于领域本体的语义合成技术,该方法以文本为数据源,引用数据源和领域本体之间的映射关系来表达数据文本的语义。文章提出了一个语义合成模型,该模型由领域本体、关键词语义抽取、概念语义相似度计算及语义推理等相关技术模型组成。文中对该模型进行了实验验证,通过对实验结果进行分析推理可知,文中提出的基于领域本体的语义合成模型提高了检索系统的查准率和计算机处理信息的能力,从而也提高了用户的满意度。

**关键词:**领域本体;语义合成;语义相似度;概念映射

**中图分类号:**TP31

**文献标识码:**A

**文章编号:**1673-629X(2013)07-0044-04

doi:10.3969/j.issn.1673-629X.2013.07.011

## Research on Semantic Synthesis Based on Domain Ontology

LIN Pei-jin<sup>1</sup>, CAO Su-yan<sup>2</sup>, YING Jie<sup>1</sup>

(1. Nanjing University of Posts & Telecommunications, Nanjing 210003, China;

2. Department of Electronics and Information Engineering, Nanjing Communications Institute of Technology, Nanjing 211188, China)

**Abstract:** To overall mining the user's query information in the retrieval system, a study of semantic synthesis is proposed based on domain ontology. The method uses text as data source, and the mapping relation between the data source and the domain ontology is referred to express semantic of the data text. The semantic synthesis model has been verified by experiments, which includes the relative technology models such as domain ontology, key words semantic extraction, concept semantic similarity calculation and semantic reasoning. The experiment results show that the precision rate of retrieval system, the process ability of the computer and the user satisfaction are all improved through studying on semantic synthesis based on domain ontology in this paper.

**Key words:** domain ontology; semantic synthesis; semantic similarity; concept mapping

## 0 引言

随着互联网络和数字化信息的不断发展,以及网络资源的不断丰富,如何在万维网上海量的信息中发现并获取有价值的信息将越来越成为一大难题。互联网已经发展成为一个庞大的全球化信息资源库,在这些信息资源库基础上,人们可以进行在线查询获取信息,浏览网页,下载所需资源,网购商品等等,而这些活动的进行大部分都是利用各种网络搜索引擎来实现,如 Yahoo, Google, Baidu 等这些搜索引擎可以帮助人们进行有效的信息检索和分类,以实现各种有效信息资源的查询和获取。然而这些搜索工具大都只是基于关键词的搜索引擎,利用关键词检索到的信息大都缺

乏语义层面上的考虑,不能很好地揭示信息的内涵和关联,搜索到的资源仍然是用户所不感兴趣的内容,对用户来说甚至是一些毫不相关的“垃圾信息”。因此,面对海量的、多种多样的数据,必须理解数据资源的语义才能进行更好地融合,文中通过基于领域本体的语义合成技术来提高检索系统的检索质量,从而提高计算机处理信息的能力。

本体技术应用于网络资源管理模型中可以解决不同网络管理模型相互操作时的语义信息互换,本体技术广泛应用于各种研究领域中,如在人工智能、信息检索、Web 服务、软件工程等领域有着广泛的应用。本体一词来源于哲学,之后随着人工智能和信息技术的

收稿日期:2012-09-25

修回日期:2012-12-28

网络出版时间:2013-03-05

基金项目:国家“973”重点基础研究发展计划项目(2006AA01Z201)

作者简介:林培金(1987-),女,硕士研究生,研究方向为数据仓库和决策支持系统;曹苏燕,副教授,研究方向为数据分析和决策支持系统。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130305.0829.064.html>

进步不断得到发展。Gruber<sup>[1,2]</sup>对本体给出了一个定义“本体是概念化的明确规范。”后来 Studer 等人总结前人的相关定义描述,将本体的描述概括为“共享概念化的形式的明确规范”<sup>[3]</sup>。在基于本体的知识检索领域,概念的语义相似度计算和语义的推理是进行概念语义扩展的重要步骤,因此,基于领域本体的语义合成技术就成为提高检索质量的关键技术之一。

文中在本体模型的基础上,提出基于领域本体的语义合成模型<sup>[4,5]</sup>,该模型能够比较准确地反映出源文本内容的概念语义信息,文章利用领域本体中的语义推理和语义相似度来表达数据源所表达的语义信息,并给出了数据源文本预处理模型、基于领域本体的语义合成模型以及这两个模型框架的执行过程和算法,为使用者提供了具有语义支持、服务质量保障的数据访问服务,可以有效提高检索系统的查准率和检索质量。

## 1 领域本体

### 1.1 本体模型

“本体是对概念体系明确的、形式化的、可共享的规范说明”这是 R. Studer 对本体概念认识的总结描述,也是关于本体概念获得广大研究人员认同的一种描述。R. Studer 所支持的关于本体的定义描述包含了四层含义:概念化、明确化、形式化和共享性四个特性。本体的四个特征的具体阐述如下:

“概念化”是抽象出客观世界中的一些现象,再整理成一些概念而得到的模型,其表示的含义与具体的应用环境无直接关系;

“明确化”是指概念及这些概念之间的关系都具有明确的定义;

“形式化”是指概念是能够被机器所认识的,即具有一定的机器可读性;

“共享性”指本体中的概念体现的是领域团体认可而不是个人独享的。

假设已经建立一个本体论,其中某一部分表示由图1所示。

本体描述的是团体公认的知识模型,这种知识模型通过概念及概念之间的关系来描述概念的语义。它的目的是使不同开发工具和应用平台的信息和应用能够进行通信、共享和重用,节省大量的人力、物力和财力等各种资源。在具体的实际应用中,并不一定要严格按照规定的元语来构造本体,可以根据具体需要

进行概念扩展或剪裁<sup>[6]</sup>。

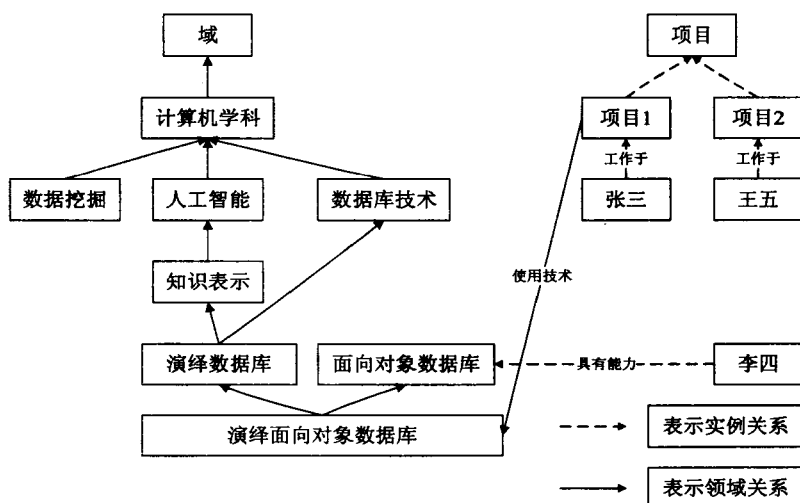


图1 一部分本体论

### 1.2 领域本体模型

在上述对本体的描述中,可以总结出使用本体的好处:

(1)使用本体可以澄清领域知识的结构,为知识表示奠定一定的模型基础;

(2)本体提供了统一的术语和概念,能够实现不同领域和不同应用之间的知识共享;

(3)本体是一种的公认的知识概念模型,易于重用,可以有效避免重复的领域知识分享。

领域本体是对具体领域中的概念和关系的抽象描述,反映的是一个对给定领域的通用观点,其通过定义概念与概念之间的关系来描述概念的语义信息,是相关领域信息资源的组织框架。由于事物的复杂性和多样性以及人们认识的局限性、对专业知识理解的局限性,以及难以确定本体的描述粒度等原因,使得构造领域本体出现较大的难度。但基于前人对本体的研究已经有了一定的基础,因此,对于领域本体的建立,不必从头开始,只需在本体模型的基础上构建领域本体模型即可。

领域本体模型可以用8元组表示: $DO = \{C, P, R', H', R'', I, FA\}$ ,其中, $DO$ 表示领域本体; $C$ 表示概念的集合; $P$ 表示领域本体中的属性集合; $R'$ 表示类间的同义关系; $H'$ 表示类间的上下位关系; $R''$ 表示类间的用户自定义关系(包括 part-of 关系也用自定义关系来描述),也就是类的对象属性; $I$ 表示为领域内概念实例的集合; $F$ 表示为一种函数关系, $A$ 表示为概念或者概念之间的关系所满足的公理,是一些永真式。

领域本体成为人机之间、机器与机器之间相互理解的语义基础。基于领域本体的语义合成可以将用户通过自然语言的需求描述映射到具有公共认知性的领域本体中,在语义合成的过程中,解决语义冲突问题,

从而提高了检索的质量,返回给用户满意的结果。

## 2 文本预处理

(1) 对数据源进行分词、词性标注等预处理<sup>[7]</sup>。可以采用分词工具来实现,如 IKAnalyzer、ictclas4j、paoding、MMSEG4J 等都是中文分词系统,这些系统的主要功能有:中文分词、词性标注和未登录词识别;

(2) 消除停用词。去除如标点符号、注释等与文本无关的额外信息,形成纯文本(字符)组成的资源流。预处理过程必须消除这些停用词,这样可以消除这些无用词汇对后面信息处理过程的干扰,降低噪音,提高处理效率;

(3) 抽取关键词<sup>[8]</sup>。将文本中出现的一些名词、动词或是人名、地名等抽取出来,它们一般具有表示文本的重要意义。文本预处理框架如图 2 所示。

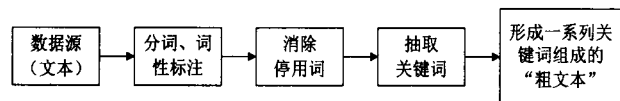


图 2 文本预处理框架

## 3 语义合成

### 3.1 概念映射

从上述经过预处理得到的粗文本可表示为:  $d = (k_1, k_2, \dots, k_n)$ , 其中  $d$  表示用户输入的文档,  $k_n$  为描述的关键词。文中的概念映射是将这些关键词映射到领域本体中,从而将文本映射形成基于领域本体的概念集合。

具体实现过程如下:

1) 首先将“粗文本”中的关键词逐一与领域本体概念树中的概念进行匹配;

2) 匹配过程中使用到语义相似度的计算和语义推理等技术,如果匹配成功,则用领域本体中的概念来替换源文本中的关键词,如果匹配失败,则认为该关键词不符合本体概念,不进行任何处理;

3) 通过挖掘文本信息中的语义信息,用领域本体概念代替文本中的关键词,形成领域本体的概念集合。

### 3.2 语义相似度

形成领域本体的概念集合后,通过计算“概念-概念”的语义相似度<sup>[9-11]</sup>,来进行相关的语义合成。假设  $C_i$  和  $C_j$  是概念集合中的两个概念,文中在进行相似度计算时,首先做如下工作:

(1) 如果  $C_i$  和  $C_j$  相似,即为同义关系,则相似度  $R(C_i, C_j) = 1$ ;

(2) 如果  $C_i$  和  $C_j$  不相似,则采用下面的相似度计算公式:

$$\text{Sim}(C_i, C_j) =$$

$$\theta * \frac{d(C_i) + d(C_j)}{\text{Dist}(C_i, C_j) * 2 * \text{Dep} * \max(|d(C_i) - d(C_j)|, 1)}$$

其中  $d(C_i)$  和  $d(C_j)$  分别是  $C_i$  和  $C_j$  所处树状结构的层,  $\text{Dist}(C_i, C_j)$  描述了概念  $C_i$  和  $C_j$  之间在树中的最短路径,  $\text{Dep}$  是该本体树的最大深度。 $\theta$  是公式中用来调节的参数,取值为大于 0 的整数。

通常情况下,概念距离与概念语义相似度有如下关系:即两个概念的距离越大,则语义相似度越小;如果两个概念距离越小,其语义相似度越大。

文中预先设定一个阈值,用来判断概念  $C_i$  和  $C_j$  之间的关系。如果计算出来的语义相似度值  $\text{Sim}(C_i, C_j)$  高于设定的阈值,则认为概念  $C_i$  和  $C_j$  之间存在一定的关联,并提取此关联,否则,那些语义相似度低于设定阈值的概念将被忽略,不提取任何信息,由此可挖掘出文本中隐含的语义信息,而去除那些不相关或者联系不大的冗余概念。语义合成的总体框架如图 3 所示。

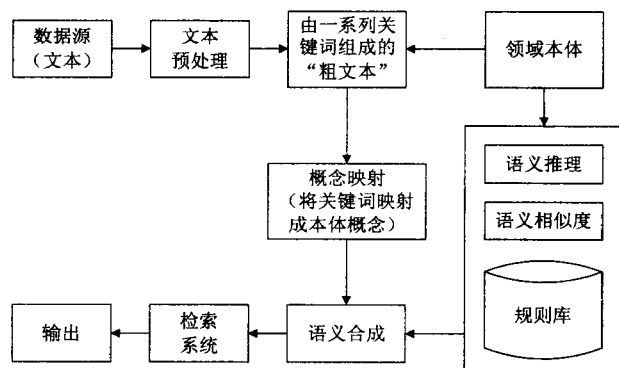


图 3 基于领域本体的语义合成框架

## 4 实验分析

针对文中提出的基于领域本体的语义合成研究,对其进行了初步的实验,首先做如下描述:假设实验系统中一共有  $s$  篇文献,现需要在检索系统中检索某课题的文献,并计算其查全率和查准率。

$s$ : 表示检索系统中的总文献数;

$sp$ : 表示检索系统中与课题相关的文献数;

$w1$ : 表示不使用语义合成所检索出的文献数;

$w2$ : 表示使用语义合成后所检索出的文献数;

$wp1$ : 表示不使用语义合成所检索出的与课题相关的文献数;

$wp2$ : 表示使用语义合成后所检索出的与课题相关的文献数;

$r1 = wp1/sp$ : 表示不使用语义合成的查全率;

$r2 = wp2/sp$ : 表示使用语义合成后的查全率;

$p1 = wp1/w1$ : 表示不使用语义合成的查准率;

$p2 = wp2/w2$ : 表示使用语义合成后的查准率。

文中进行了三次实验,分别取  $s = 100, 1000, 2000$ ,

可得三次实验的各个参数值和结果如表 1 所示:

表 1 使用语义合成技术前后的查全率和查准率

s	sp	w1	wp1	w2	wp2	r1	r2	p1	p2
100	50	60	48	40	35	0.96	0.7	0.8	0.875
1000	100	200	100	110	98	1.0	0.98	0.5	0.89
2000	200	800	198	210	190	0.99	0.95	0.25	0.9

从表中可看出,通过比较一般检索和使用语义合成技术后的检索的查全率和查准率,可得出如下分析结果:使用了语义合成技术后的查全率稍稍低于普通检索,但查准率高于不使用语义合成技术的普通检索,且查询效率明显高于普通的检索;并发现随着检索系统中的总文献数的增加,基于语义合成技术的检索优点就会显得越突出。这是因为通过语义合成技术,过滤了一些语义无关的信息,从而提高了检索质量。

5 结束语

文中提出了基于领域本体的语义合成的研究,将待查询信息与本体技术相结合,有效地解决了查询过程中的语义问题;本体为用户输入的查询描述加入了语义信息,提高了信息查询的准确度和智能化,提高了查询的质量和效率,从而使计算机处理文本信息的能力和品质也改善了,与此同时,用户信息检索的满意度也提高了。

文中下一步可以针对不同的数据源或数据对象进行语义合成,将该语义合成技术进一步深入研究,并应用到更广阔的领域。

参考文献:

[1] 刘 琼,李宝敏.一种果品领域本体库的构建方法[J]. 计算机技术与发展,2009,19(1):197-199.

[2] Gruber T R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing[J]. International Journal of Human Computer Studies,1995,43(5/6):907-928.

[3] Studer R,Benamins V R,Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering,1998,25(1-2):161-197.

[4] 王 宁,王能斌.异构数据源集成系统查询分解和优化的实现[J]. 软件学报,2000,11(2):222-228.

[5] Tomasic A,Amouroux R,Bonnet P,et al. The distributed information search component (Disco) and the World Wide Web[J]. ACM SIGMOD Record,1997,26(2):546-548.

[6] 王昭龙,李 霞,许瑞芳.多关键字查询中 LCA 剪枝概念树的查询扩展技术研究[J]. 计算机科学,2010,37(4):132-135.

[7] 赵 心.一种基于关联规则的中午概念集生产算法[J]. 计算机科学,2004,31(7):175-177.

[8] 韩家炜,孟小峰,王 静,等. Web 挖掘研究[J]. 计算机研究与发展,2000,37(5):513-520.

[9] 吴 建,吴朝晖,李 莹,等.基于本体论和词汇语义相似度的 Web 服务发现[J]. 计算机学报,2005,28(4):595-602.

[10] Navigli R,Velardi P. Learning domain ontologies from document warehouses and dedicated web sites[J]. Computational Linguistics,2004,30(2):151-179.

[11] 黄 果,周竹荣.基于领域本体的概念语义相似度计算研究[J]. 计算机工程与设计,2007,28(10):2460-2463.

(上接第 43 页)

标准的分析,设计了一套 IPv6 协议测试集,利用自主开发的测试系统对 Linux 下的一种 IPv6 协议实现进行了测试,给出了测试结果,并对结果进行了分析,验证了该协议实现的正确性和完备性。目前,基于模块化测试描述模型和 MTDL 的协议测试集还不够完备,后续将继续致力于协议测试例的开发,提高可测试协议的覆盖度,并进一步完善测试系统功能。

参考文献:

[1] 田 军,张玉军,李忠诚. IPv6 协议一致性测试系统[J]. 计算机辅助设计与图形学学报,2002,14(4):296-300.

[2] 郑红霞,田 军,张玉军,等. IPv6 协议一致性测试例的设计[J]. 计算机应用,2003,23(4):62-64.

[3] Interoperability Lab:IP Consortium Test Suite, Internet Protocol Version 6[M]. Hampshire:University of New Hampshire, 2000.

[4] Tahi project. IPv6 node requirements revision[EB/OL]. 2004-10. <http://www.tahi.org/ume/testspec/ph1-host-policy.html>.

[5] Tian Jun,Li Zhongcheng. The next generation Internet protocol and its test[C]//Proc. of IEEE International Conference on Communications. [s. l.]:[s. n.],2001:210-215.

[6] Zhang Yujun,Li Zhongcheng. A new formal test suite specification language for IPv6 conformance testing[C]//Proceedings of International Conference on Communication Technology. [s. l.]:[s. n.],2003:174-177.

[7] ISO/IEC 9646:IT-OSI-Conformance testing methodology and framework[S]. [s. l.]:[s. n.],1996.

[8] RFC2460: Internet Protocol, Version 6 (IPv6) Specification [S]. [s. l.]:Network Working Group,1998.

[9] Li Qing,Jinmei T,Shima K. IPv6 详解[M]. 北京:人民邮电出版社,2009.

[10] 张 升,刘兴伟,郭 闯. IPv6 协议一致性测试[J]. 西华大学学报:自然科学版,2009,28(1):34-37.

[11] 夏启志. IPv6 协议一致性测试通用执行系统设计与实现[D]. 北京:中国科学院计算技术研究所,2005.

[12] 孙静波,张玉军,李忠诚. IPv6 中邻居发现协议及其测试[J]. 计算机工程与应用,2004,24(32):79-81.

# 基于领域本体的语义合成研究


作者:

林培金, 曹苏燕, 应捷, LIN Pei-jin, CAO Su-yan, YING Jie

作者单位:

林培金, 应捷, LIN Pei-jin, YING Jie(南京邮电大学, 江苏南京, 210003), 曹苏燕, CAO Su-yan(南京交通职业技术学院电子信息工程系, 江苏南京, 211188)

刊名:

计算机技术与发展 

英文刊名:

Computer Technology and Development

年, 卷(期):

2013, 23(7)

本文链接: [http://d.wanfangdata.com.cn/Periodical\\_wjfz201307011.aspx](http://d.wanfangdata.com.cn/Periodical_wjfz201307011.aspx)