基于 HowNet 的中文本体学习方法研究

贾文娟,何 丰
(北方民族大学 计算机科学与工程学院,宁夏 银川 750021)

摘 要:目前针对国内在中文环境下本体学习的研究才刚刚起步的现状,对本体学习和 HowNet 进行了简单介绍,提出了基于 HowNet 的中文本体学习的主要思路。当前,本体学习的研究重点在于概念及概念间关系抽取。采用文本语料作为输入,首先对文本进行预处理,然后基于 HowNet 生成了一个领域语义词典,在本体学习中加入领域核心概念本体,在概念关系抽取阶段,采用基于 HowNet 的语义相似度计算方法。实验证明,提出的本体学习方法能够有效改进概念和概念间关系抽取的准确度。

关键词:本体学习;HowNet;概念抽取;概念关系抽取
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2011)06-0077-04

Research of Chinese Ontology Learning Based on HowNet

JIA Wen-juan, HE Feng
(Department of Computer Science and Engineering, Beifang University of Nationality, Yinchuan 750021, China)

Abstract: At present based on ontology learning just started in the environment of Chinese, give a brief introduction about HowNet and ontology learning, and propose the main ideas based on the Chinese HowNet. Ontology learning is focused on the concept extraction and relation extraction recently. The text corpus as input, the first of the pre-text, then it generates a field dictionary based on HowNet semantic, and adds the core concepts ontology into the field of ontology learning, in the concept relation extraction stage, the similarity calculation is based on HowNet methods. Experimental results show that the proposed method can effectively improve the accuracy of ontology concept extraction and relation extraction.

Key words: ontology learning; HowNet; concept extraction; relation extraction

0 引 言

近年来,关于本体的研究在计算机科学中越来越多。本体的狭义定义,被普遍引用的最出名的是由 Gruber 提出的"本体是概念模型的说明,并且这种说明是规范的明确的"[1]。它是用来记录某个特定领域的概念和这些概念之间的关系,使这些概念之间在共享的情况下具有明确的、形式化的定义,以方便人机之间以及机器之间的交流。目前,本体被广泛应用于计算机软件方向的知识工程、机器翻译、信息检索及人工智能等领域中。

由于本体在各个研究领域中的应用不断改进,人们对本体的需求越来越多。目前具有代表性的本体有英文版的 WordNet、中文的 HowNet 等,它们大都叙述的太广泛,缺乏与具体的领域相关知识的结合,同时本体的应用总是易于局限于某一特定的领域或具体的任务。然而通过手工方式构建的领域本体的最大缺点是需要耗费大量的人力、物力和财力,并且开发周期长导致易在构建本体的时候出错,并且很难做到及时更新。因此,构建领域本体是本体研究和应用的基础。为了推动本体的发展和应用,通过自动或半自动的构建本体以及本体学习等相关技术已被相继提出,同时成为当前研究和引用的热点[2],其目的是利用机器学习和统计等技术自动或半自动地从现有的数据资源中获取期望的本体。目前该方向在国外的研究特别活跃,而国内的发展却相对滞后,表现在对中文环境下本体学习的研究才刚刚起步。

收稿日期:2010-11-12;**修回日期**:2011-03-06
基金项目:国家自然科学基金资助项目(71061001)
作者简介:贾文娟(1986-),女,硕士研究生,主要研究方向为语义网和本体;何 丰,教授,博士,硕士生导师,研究方向为语义 Web 和本体、数据挖掘、知识工程等。

1 基于 HowNet 的本体学习简介

1.1 本体学习

目前本体研究方面的共识更多的是表现为计算机服务方面,但是计算机作为机器,不能像人脑那样精确地理解人类交往中的诸多语义,计算机也只能把文本

转换成字符串进行识别。在计算机领域中的本体及其应用问题,归根到底就是对本体是如何表达共识的研究,可以延伸到本体的形式化定义问题。显式的、形式化的本体描述方式,以一种结构化的方式共享和重用领域知识为人机结合提供方便。

本体学习(ontology learning)即是从现有知识源(如文本、词典、知识库等)方式来得到领域知识,以通用的自动或半自动的方式构造本体。作为本体学习的数据源的有半结构化文档,如XML,HTML,DTD以及纯文本。根据不同的输入数据源类型,本体学习分为基于文本、字典、结构化数据和知识库的本体学习等。从形式的观点来看,本体的结构(ontology structure)是一个五元组 O:= {C,R,HC,Rel,AO},其中集合C与集合R是不相交的,C中的元素称为概念,R中元素称为关系;HC表示概念间的分类关系;Rel表示概念间的非分类关系;AO表示本体公理。从本体的结构可以看出,本体学习的任务包括概念的获取、概念间的关系的获取(包括分类关系和非分类关系)以及公理的获取。目前本体的研究也主要集中在概念的获取和概念间的关系获取,公理的研究则相对较少。文中以非结构化的文本语料作为数据源。

1.2 HowNet

HowNet(中文名称 HowNet)是由中国科学院董振东先生开发的,按照董先生的说法(杜飞龙,1999),"《HowNet》是一个常识知识库,它的基本内容是以英汉两种词语所代表的概念为描述对象,以揭示各个概念之间以及该概念下所属的属性之间的关系[3]。"

HowNet是一个不可多得的以双语为代表的常识知识库,其基本组织单位是概念。概念使用义原定义。概念与概念之间的关系、概念与义原之间的关系以及义原与义原之间的关系构成了HowNet的网状知识体系,这些关系主要体现在HowNet的词典和各个特征文件中。

HowNet系统的哲学是[4]:"凡事都在特定的时间和空间内不停地运动和变化。通常是从一种表现形式变化到另一种表现形式,结果由其属性值的改变来体现。"所以可以得出万物是HowNet的运算和描述的基本单位,它包括部件、属性、属性值、时间、空间以及事件。其中最重要的两个因素是部件和属性[5],这两个基本单位在知网的哲学体系中的地位举足轻重。每一个事物可能是另一个事物的部件,与此同时也可能正好是其他事物的整体,这就是所谓的部件。眼睛和眉毛是一个人的部件,但与此同时,人又是他所属的家庭或社会的部件。所以一个事物到底属于整体还是属于部件,应该是具体情况具体对待的,通常也是事物所在的环境和系统来决定的。

对于部件在整体中的部位和它的功能来讲,知网遵循这样的规律:比照人类来认识事物的部件在它整体中的部位以及功能的描述。例如建筑物的门和窗比照人类的眼睛和眉毛等等。这是人类认识事物方法的共性的体现。

对属性来说,每个事物都包含着多个不同的属性,是不是同一个事物是由属性决定的,即没有属性就没有事物。人有吃东西、喝水、繁殖、对生的渴望等自然属性以及吃饭要吃的熟食、要使用碗,走路的时候要走的马路、要上学、要工作等社会属性。在某些特定的情况下可以说属性比事物更重要,即宿主决定属性。同理,属性与宿主之间的关系并不等同于部件与整体之间的关系。这体现在知网涉及属性的标注规范上,因此知网做出在标注属性时必须标注它可能的宿主的类型这一硬性规定,还必须标注它所指向的属性。

1.3 基于HowNet的中文本体学习的主要思路

通用的本体学习流程主要包括5个模块[6]:数据输入模块、预处理模块、本体抽取模块、本体评价与编辑模块。1.本体学习工具的输入可以是各种类型的数据源,预处理模块对数据源进行预处理;2.学习模块通过各种本体学习算法从预处理的结果中获取本体;3.把产生的候选本体呈现给用户;4.用户在评价/编辑模块的帮助下对产生的候选结果进行评价,并将最终的结果保存到本体库中。从以上步骤可以看出,整个过程是在用户参与下的半自动的过程。另外,需要注意一点,学习模块在获取本体的过程中需要参照已有的或现存的本体。文中对本体评价与编辑模块不做太多说明。

领域本体的最基本、最重要的元素是领域概念[7]。概念获取是本体学习的起点。中文领域概念获取难度之所以要大于英文领域概念获取归根到底还是由汉语的复杂性所决定。领域概念获取在中文环境下存在多方面困难,文本在利用中文分词系统进行标注时,专业领域术语大多被作为未登录的词被切分成散串;如何有效识别表达同一个概念的多个不同术语,这就造成了同义词识别困难;另外,判定一个概念是否为某个专业领域内的概念,这需要领域专家的参与与研讨,在一定程度上增加了概念领域性判定的困难。一个概念的所有对象所共有的属性集被认为是概念内涵,因此属性在对于概念的理解中显得至关重要。但属性关系相对于继承关系来说鲜受关注,不支持属性关系学习是现有的本体学习工具的通病。根据知网的属性——宿主关系、词汇相似度和知网中属性间上下位关系等特性,采用基于知网的属性关系学习方法来获取属性关系、对候选属性集进行过滤,将有助于提高概念关系抽取的准确率。

2　基于 HowNet 的本体学习方法的设计

2.1　基本框架

以文本语料作为输入,运用基于 HowNet 概念获取和概念关系获取并用的方法,最终得到各个概念之间的关系,从而进行领域本体的格式化输出是本方法的重要目标(见图 1)。

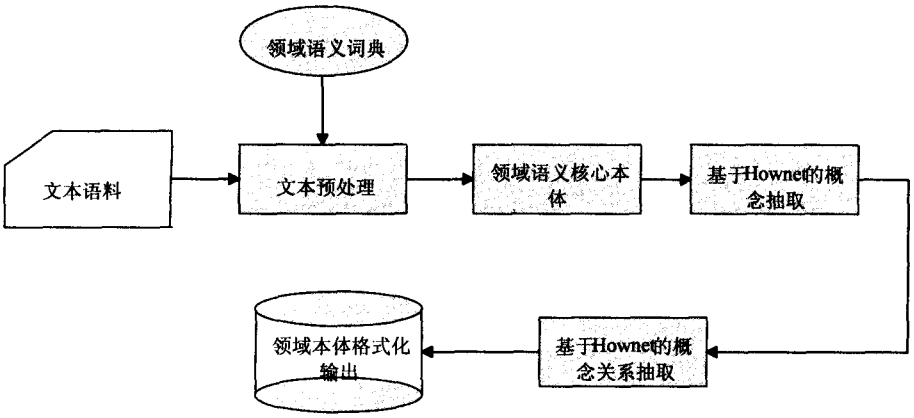

图 1　基于 HowNet 的本体学习方法基本原理

从图 1 可以看出,基于 HowNet 的本体学习方法的基本步骤包括以下几个方面:文本预处理;领域语义词典生成,输入领域核心本体;概念抽取;概念关系抽取;领域本体的格式化输出。

2.2　基于 HowNet 的领域语义词典生成

目前,获取领域知识的方法主要有二种[8]:一是通过专业词典,二是对语料库进行统计分析。而以词典作为获取概念的方式有很多缺点和局限性,很多领域没有专业词典是其局限性的表现之一,而且需要首先构建出整个领域的语料库,以及对每个词语在特定领域里的分布情况计算量大、效率低是采用统计方法的不足之处。

特定领域词典[9]是本体学习中进行领域概念抽取的基础。由于 HowNet 等语义词典大部分都属于顶层本体,具有很大的通用性,所以不能够完全满足领域语义词典的需要,所以文中在 HowNet 的基础上,结合特定领域知识建立与该领域相关的语义词典,比如建立计算机领域语义词典、法律领域语义词典等。有关命题、术语和相关文档等都属于特定领域的领域知识,在建立领域语义词典的过程中大都需要领域专家的参与,之后对其中的关键词进行分离,语义相关词检索,领域不相关词剪集。为了避免在分词过程中将领域概念切分成无意义的词,文中基于 HowNet 生成了领域语义词典,将已经知道的领域概念加入到领域词典中,提高分词的准确率。

2.3　文本预处理

对文本进行处理,包括去分词和去停用词等。中文文本的分词方法很多,文中采用中科院开发的分词工具 ICTCLAS 对文本进行分词和词性标注,这样可以在处理阶段就去掉那些对文本作用不大的虚词、介词等词语,只对一些关键的如名词、动词、形容词等重要词语进行处理。分词后将文本中出现的领域无关的高频词汇进行过滤,这样可以在很大程度上提高程序的运行效率。

2.4　输入领域核心概念本体

领域知识是主要围绕着一些重要概念组织起来,例如在众所周知的体育领域,人们平常爱好的网球、乒乓球、足球、篮球运动等一系列运动概念都是由词“球”与其它不同的字词搭配形成。因此,为了快速从中文文本和已有资源中有效地学习领域本体,首先要创建一个领域核心概念本体,通过这些领域中的核心词汇,可以获取大量的领域术语。围绕领域核心概念本体构建语料库,在本体学习的开始就确定了目标,缩小了本体的学习范围,可以降低语料选取偏差对本体学习结果的影响,使本体学习具有一定程度的指导性和针对性,有效地避免了大规模运算,从而提高了效率,达到预期的效果。基于 HowNet 中的概念,领域核心词汇“硬盘”是 HowNet 中的“机器”,属于“电脑”的一种,它可以与之连接形成比较完整的本体,在这里也能将“硬盘”作为领域里的关键概念,从而使“硬盘”与电脑形成整体部分关系。

2.5　概念抽取

HowNet 的根本思想在于将各个词语的词义以义项表示,HowNet 中每一个词可用多个概念来表示,也成为义项,每一个概念用一个记录来表示,如下所示:

NO. =032582

W_C=锻炼

G_C=V [duan4 lian4]

S_C=

E_C= ~意志, ~耐力, ~实践能力, ~野外生存能力, ~心性

W_E=steel

G_E=V

S_E=

E_E=

DEF={cultivate|培养}

其中 NO. 为义项的编号,W_C ,G_C ,E_C 分别是汉语中的词语、词性和例子,对应的,W_ E、G_ E、E_ E 分别是英语的词语、词性和例子,DEF 是 HowNet 的核心,是 HowNet 对于概念的定义,称为一个语义表达

式,它由 KDML[10](Knowledge Database Mark-up Language) HowNet 知识系统描述语言来表示。如,爱:DEF={emotion|情感:CoEvent={love|爱恋}},词汇在 HowNet 中的首义原是指该词汇在 DEF 定义中出现的第一个义原,例如:"爱"的首义原就是"emotion|情感",它能很好地表达出词汇所对应概念的主要语义信息。

一种有效地解决 HowNet 中概念关系的直接表征问题的可视化方法,就是对 HowNet 进行形式概念分析。例:对"医"进行概念分析,会得到一些相关的概念,"医院"、"医生"等,再根据"医院"、"医生"进行上下位扩展,得到"医药"、"医治"、"疾病"等等。

2.6 概念关系抽取

HowNet 对各个概念之间和概念的具体属性之间的各种关系做了重点的描述。上下位关系、同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系组成了系统 HowNet 主要包括的 8 种关系。不难得出结论,义原之间组成的不止是一个单纯的树状结构,取而代之的是一个复杂的网状结构。在 HowNet 中,词语由义项来表示,义项通过义原进行描述,故义项之间的关系可以通过义原的关系得到,而义原之间的关系,可以通过义原之间的距离得到。

文献[11]中通过义原的语义距离的计算进行 HowNet 中概念(义项)的相似度计算。先找出路径距离,也就是两个义原在知网的层次体系中的路径距离,设为 d,便可以计算出这个义原之间的语义距离。其公式为 $\mathrm{Sim}(p1,p2)=a/(a+d)$。其中:$p1$ 和 $p2$ 表示义原;d 是一个正整数,表示 $p1$ 和 $p2$ 在义原层次体系中的路径长度;a 是一个可调节的参数。以实词为例,其基本计算公式为 $\mathrm{Sim}(A,B)=\sum_{i=1}^{4}\beta_i \mathrm{Sim}_i(A,B)$。其中:$\beta_i(1\leqslant i\leqslant 4)$ 是可调节的参数,并且满足 $\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$。$\mathrm{Sim}_1(A,B)$ 到 $\mathrm{Sim}_4(A,B)$ 分别为第一独立义原描述式、其他独立义原描述式、关系义原描述式及符号义原描述式的相似度。如图 2 所示,分别为 HowNet 中义原和义项的相似度计算。

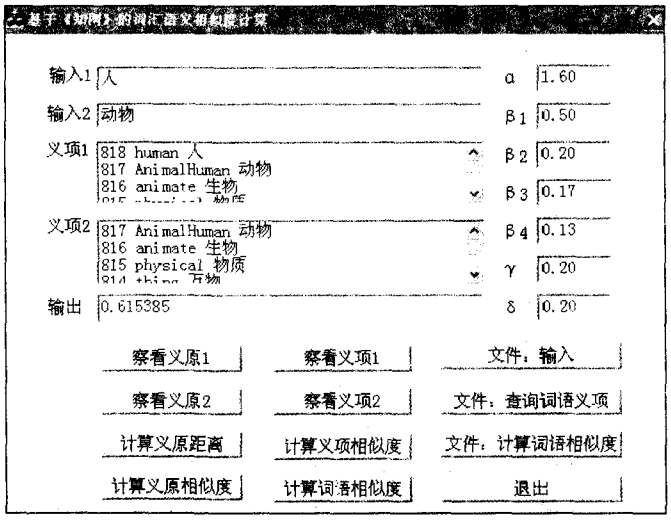

图 2 基于 HowNet 的词汇语义相似度计算

3 实验结果及分析

文中以领域核心本体的概念数的不同在同一数据集上进行对比实验来证明提出方法的有效性。实验数据集采用计算机作为领域语料背景。实验分词词性标注采用中科院开发的分词工具 ICTCLAS 对文本进行分词和词性标注,实现相似度计算采用 HowNet 计算得到的词汇语义相似度。实验结果采用查准率与查全率相结合的方式,查准率计算公式表示如下:

$$\mathrm{Pre}=\mathrm{com}\cap\mathrm{ref}/\mathrm{com}$$

查全率公式如下:

$$\mathrm{Ref}=\mathrm{com}\cap\mathrm{ref}/\mathrm{ref}$$

其中:com 为学习生成的本体,ref 为手工生成的本体。

表 1 为实验结果,可以看到领域核心概念数越多,概念和关系抽取的查准率和查全率都能得到提高,充分验证了加入领域核心本体概念和基于 HowNet 的本体学习方法的有效性。

表 1 实验结果表

	核心概念	概念抽取(%)	概念关系抽取(%)
查准率	25	76.79	34.78
	28	77.01	45.92
查全率	25	67.19	76.51
	28	68.65	79.24

4 结束语

目前,本体学习大多以统计方法为主,虽然着重地统计了词汇的词频,但却没有对词汇之间的语义关系进行很好的考虑,也不能很好地表示各个概念所表达的词义与概念之间的关系。文中对目前的本体学习方法在传统统计方法的基础上进行改进,加入领域核心概念本体,词汇之间的语义相似度通过 HowNet 来表达,并将其结合到本体学习中,以达到改进概念和关系提取的精度的效果。实验表明该方法是可行并且有效的。在下一步的工作中,将随着 HowNet 的不断升级和完善,进一步扩充领域语义词典,并对基于 HowNet 的概念关系抽取算法进行改进研究,以期望能够得到更好的效果。

参考文献

[1] Gruber T R. A translation approach to portable ontology

(下转第 84 页)

低值达到了 8%。

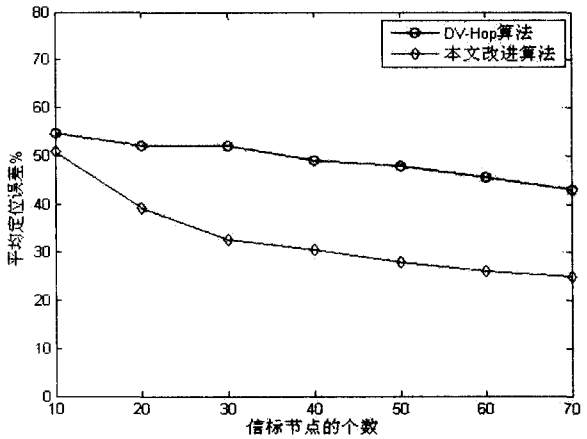

图 3 引入补偿系数后两种定位算法的平均定位误差

在图 3 中,由于进一步通过补偿系数来校正未知节点到信标节点的估计距离,使之更接近实际距离,从而更进一步地降低了未知节点的定位误差,由图中可以看出改进算法的定位误差比传统的 DV-Hop 定位算法平均减少了 15% ~ 18%,定位精度明显有所提高。当信标节点数量达到 60 时,两种算法的定位精度提高的幅度逐渐变小,平均定位误差变化趋于平稳。

总而言之,文中改进算法的平均定位误差始终小于原 DV-Hop 算法且拥有良好的稳定性。但是,在执行算法的过程中要广播大量的信标节点对增长比和计算补偿系数,文中的改进算法在通信开销和计算量方面比原算法会有所增加。

4 结束语

文中针对 DV-Hop 算法在随机分布网络环境中的局限性,先利用最小均方误差法对原算法中计算每跳平均距离的方法做改进,再引入补偿系数来校正未知节点到信标节点间的距离,提出一种基于补偿系数的 DV-Hop 改进算法。经过仿真验证,该改进算法的平均定位误差低于原算法并拥有良好的稳定性。但是,改进算法是通过增加通信量和计算量来提高定位精度的,算法的开销会有所增加。如何在保持较好定位精度的同时降低通信开销和计算量是将来需要进一步研究的内容。

参考文献:

[1] 张 杰,胡向东. 定位技术在无线传感器网络中的应用[J]. 电信快报,2008(8):34-36.

[2] 孙利民,李建中,陈 瑜,等. 无线传感器网络[M]. 北京:清华大学出版社,2005:149-151.

[3] Niculescu D, Nath B. DV Based Positioning in Ad hoc Networks[J]. Journal of Telecommunication Systems, 2003, 22(1-4): 267-280.

[4] 王书聪. 无线传感器网络分布式定位算法研究[J]. 计算机技术与发展,2008,18(11):62-65.

[5] Niculescu D, Nath B. Ad-hoc Positioning System (APS)[C]// Proc. of the IEEE GLOBECOM. San Antonio: [s. n.], 2001: 2926-2931.

[6] 刘少飞,赵清华,王华奎. 基于平均跳距估计和位置修正的 DV-Hop 定位算法[J]. 传感技术学报,2009,22(8):1154-1158.

[7] 徐建波,刘亚辉. 基于不同平面的无线传感器网络节点定位算法[J]. 计算机工程与应用,2008,44(24):115-117.

[8] 嵇玮玮,刘 中. DV-Hop 定位算法在随机传感器网络中的应用研究[J]. 电子与信息学报,2008,30(4):970-974.

[9] 张贤达. 现代信号处理[M]. 北京:清华大学出版社,2002:40-42.

[10] 刘 锋,张 翰,杨 骥. 一种基于加权处理的无线传感器网络平均每跳距离估计算法[J]. 电子与信息学报,2008,30(5):1222-1225.

[11] LI Jian, Zhang Jianmin, Liu Xiande. A Weighted DV-Hop Localization Scheme for Wireless Sensor Networks[C]// International Conference on Scalable Computing and Communications; The Eighth International Conference on Embedded Computing. [s. l.]: [s. n.], 2009: 269-272.

[12] 林金朝,陈晓冰,刘海波. 基于平均跳距修正的无线传感器网络节点迭代定位算法[J]. 通信学报,2009,30(10):107-113.

(上接第 80 页)

specifications[R]. [s. l.]: Knowledge System Laboratory, 1993.

[2] 强 彦,谢红薇. 基于 Web 数据的本体概念抽取[J]. 电脑开发与应用,2007,20(1):37-39.

[3] 唐旭日. WordNet 与 HowNet 之关系研究[J]. 湖北广播电视大学学报,2007,27(7):124-125.

[4] 廖 剑,冷 静,李艳燕,等. 知网的形式概念分析及概念相似度研究[J]. 计算机应用研究,2007,24(1):33-35.

[5] 辛日华. HowNet 的构成分析与研究[J]. 呼伦贝尔学院学报,2003,3(6):81-83.

[6] 杜小勇,李 曼,王 珊. 本体学习研究综述[J]. 软件学报,2006,17(9):1837-1847.

[7] 聂规划,傅 魁. 基于 Web 的中文本体学习研究[J]. 情报杂志,2008(6):13-16.

[8] 梁 健,王惠临. 基于文本的本体学习方法研究[J]. 信息系统,2007,30(1):112-116.

[9] 唐一之. 基于知网的领域概念抽取与关系分析研究[J]. 湘潭大学自然科学学报,2009,31(1):135-140.

[10] 李 佳,祝 铭,刘 辰,中文本体映射研究与实现[J]. 中文信息学报,2007,21(4):27-33.

[11] 刘 群,李素建. 基于知网的词汇语义相似度计算[C]// 第三届汉语词汇语义学研讨会论文集. 台北:[出版者不详], 2002.