基于《知网》的词汇语义相似度计算[1]

Word Similarity Computing Based on How-net

刘群[*] **、李素建**[+]

Qun LIU , Sujian LI

摘要

词义相似度计算在很多领域中都有广泛的应用，例如信息检索、信息抽取、文本分类、词义排歧、基于实例的机器翻译等等。词义相似度计算的两种基本方法是基于世界知识（Ontology）或某种分类体系（Taxonomy）的方法和基于统计的上下文向量空间模型方法。这两种方法各有优缺点。

《知网》是一部比较详尽的语义知识词典，受到了人们普遍的重视。不过，由于《知网》中对于一个词的语义采用的是一种多维的知识表示形式，这给词语相似度的计算带来了麻烦。这一点与 WordNet 和《同义词词林》不同。在 WordNet 和《同义词词林》中，所有同类的语义项（WordNet 的 synset 或《同义词词林》的词群）构成一个树状结构，要计算语义项之间的距离，只要计算树状结构中相应结点的距离即可。而在《知网》中词汇语义相似度的计算存在以下问题：

1. 每一个词的语义描述由多个义原组成；
2. 词语的语义描述中各个义原并不是平等的，它们之间有着复杂的关系，通过一种专门的知识描述语言来表示。

我们的工作主要包括：

1. 研究《知网》中知识描述语言的语法，了解其描述一个词义所用的多个义原之间的关系，区分其在词语相似度计算中所起的作用；我们采用一种更

[1] 本项研究受国家重点基础研究计划（973）支持，项目编号是 G1998030507-4 和 G1998030510。

[*] 北京大学计算语言学研究所 & 中国科学院计算技术研究所 E-mail: liuqun@ict.ac.cn
Institute of Computational Linguistics, Peking University &
Institute of Computing Technology, Chinese Academy of Science

[+] 中国科学院计算技术研究所 E-mail: lisujian@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences

为结构化的方式改写了《知网》中词的定义（DEF），其中采用了“集合”和“特征结构”这两种抽象数据结构。

2. 研究了义原的相似度计算方法、集合和特征结构的相似度计算方法，并在此基础上提出了利用《知网》进行词语相似度计算的算法；

3. 通过实验验证该算法的有效性，并与其它算法进行比较。

关键词：《知网》 词汇语义相似度计算 自然语言处理

Abstract

Word similarity is broadly used in many applications, such as information retrieval, information extraction, text classification, word sense disambiguation, example-based machine translation, etc. There are two different methods used to compute similarity: one is based on ontology or a semantic taxonomy; the other is based on collocations of words in a corpus.

As a lexical knowledgebase with rich semantic information, How-net has been employed in various researches. Unlike other thesauri, such as WordNet and Tongyici Cilin, in which word similarity is defined based on the distance between words in a semantic taxonomy tree, How-net defines a word in a complicated multi-dimensional knowledge description language. As a result, a series of problems arise in the process of word similarity computation using How-net. The difficulties are outlined below:

1. The description of each word consists of a group of sememes. For example, the Chinese word “暗箱(camera obscura)” is described as: “part|部件, #TakePicture|拍摄, %tool|用具, body|身”, and the Chinese word “写信(write a letter)” is described as: “write|写, ContentProduct=letter|信件”;

2. The meaning of a word is not a simple combination of these sememes. Sememes are organized using a specific knowledge description language.

To meet these challenges, our work includes:

1. A study on the How-net knowledge description language. We rewrite the How-net definition of a word in a more structural format, using the abstract data structure of *set* and *feature structure*.

2. A study on the algorithm used to compute word similarity based on How-net. The similarity between sememes, that between *sets*, and that between *feature structures* are given. To compute the similarity between two sememes, we

use the distance between the sememes in the semantic taxonomy, as is done in Wordnet and Tongyici Cilin. To compute the similarity between two *sets* or two *feature structures*, we first establish a one-to-one mapping between the elements of the *sets* or the *feature structures*. Then, the similarity between the *sets* or *feature structures* is defined as the weighted average of the similarity between their elements. For *feature structures*, a one-to-one mapping is established according to the attributes. For *sets*, a one-to-one mapping is established according to the similarity between their elements.

3. Finally, we give experiment results to show the validity of the algorithm and compare them with results obtained using other algorithms. Our results for word similarity agree with people's intuition to a large extent, and they are better than the results of two comparative experiments.

Keywords: How-net, Word Similarity Computing, Natural Language Processing

1. 引言

自然语言的词语之间有着非常复杂的关系，在实际的应用中，有时需要把这种复杂的关系用一种简单的数量来度量，而词义相似度就是其中的一种。

词义相似度计算在很多领域中都有广泛的应用，例如信息检索、信息抽取、文本分类、词义排歧、基于实例的机器翻译等等[Gauch&Chong 1995，LI，Szpakowicz & Matwin 1995，王斌，1999，李涓子，1999]。本文的研究背景是基于实例的机器翻译。在基于实例的机器翻译中，词语相似度的计算有着重要的作用。例如要翻译“张三写的小说”这个短语，通过语料库检索得到译例：

1）李四写的小说／the novel written by Li Si

2）去年写的小说／the novel written last year

通过相似度计算我们发现，“张三”和“李四”都是具体的人，语义上非常相似，而“去年”的语义是时间，和“张三”相似度较低，因此我们选用“李四写的小说”这个实例进行模拟翻译，就可以得到正确的译文：

the novel written by Zhang San

如果选用后者作为实例，那么得到的错误译文将是：

* the novel written Zhang San

通过这个例子可以看出相似度计算在基于实例的机器翻译中所起的作用。

在基于实例的翻译中另一个重要的工作是双语对齐。在双语对齐过程中要用到两种语言的词义相似度计算，这不在本文所考虑的范围之内。

2. 词语相似度及其计算的方法

2.1 词语相似度的含义

词语相似度是一个主观性相当强的概念，没有明确的客观标准可以衡量。脱离具体的应用去谈论词语相似度，很难得到一个统一的定义。

本文的研究主要以基于实例的机器翻译为背景，因此在本文中我们所理解的词语相似度就是两个词语在不同的上下文中可以互相替换使用而不改变文本的句法语义结构的程度。两个词语，如果在不同的上下文中可以互相替换且不改变文本的句法语义结构的可能性越大，二者的相似度就越高，否则相似度就越低。

相似度这个概念，涉及到词语的词法、句法、语义甚至语用等方方面面的特点。其中，对词语相似度影响最大的应该是词的语义。

在本文中，相似度被定义为一个 0 到 1 之间的实数。

词语距离与词语相似度之间有着密切的关系。实际上，词语距离和词语相似度是一对词语的相同关系特征的不同表现形式，二者之间可以建立一种简单的对应关系。对于两个词语 W_1 和 W_2，我们记其相似度为 $Sim(W_1,W_2)$，其词语距离为 $Dis(W_1,W_2)$，那么我们可以定义一个满足以上条件的简单转换关系：

$$Sim(W_1,W_2)=\frac{\alpha}{Dis(W_1,W_2)+\alpha} \qquad \cdots\cdots(1)$$

其中 α 是一个可调节的参数。 α 的含义是：当相似度为 0.5 时的词语距离值。

这种转换关系并不是唯一的，我们这里只是给出了其中的一种可能。

在很多情况下，直接计算词语的相似度比较困难，通常可以先计算词语的距离，然后再转换成词语的相似度。

词语相关性反映的是两个词语互相关联的程度。可以用这两个词语在同一个语境中共现的可能性来衡量。词语相关性和词语相似性是两个不同的概念，二者没有直接的对应关系。

2.2 词语相似度的计算方法

词语距离有两类常见的计算方法，一种是根据某种世界知识（Ontology）或分类体系（Taxonomy）来计算，一种利用大规模的语料库进行统计。

根据世界知识（Ontology）或分类体系（Taxonomy）计算词语语义距离的方法，一般是利用一部同义词词典（Thesaurus）。一般同义词词典都是将所有的词组织在一棵或几棵树状的层次结构中。我们知道，在一棵树状图中，任何两个结点之间有且只有一条路径。于是，这条路径的长度就可以作为这两个概念的语义距离的一种度量。

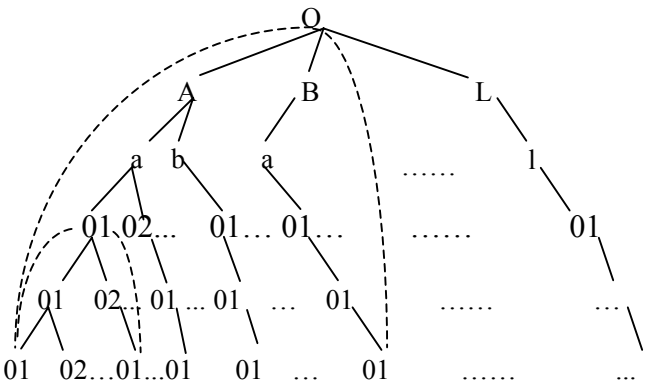

图 1 《同义词词林》语义分类树状图

[王斌，1999]采用这种方法利用《同义词词林》来计算汉语词语之间的相似度（如图 1 所示）。有些研究者考虑的情况更复杂。[Agirre & Rigau 1995]在利用 Wordnet 计算词语的语义相似度时，除了结点间的路径长度外，还考虑到了其它一些因素。例如：

概念层次树的深度：路径长度相同的两个结点，如果位于概念层次的越高层，其语义距离较大；比如说：“ 动物 ”和“ 植物 ”、“ 哺乳动物 ”和“ 爬行动物 ”，这两对概念间的路径长度都是 2，但前一对词处于语义树的较高层，因此认为其语义距离较大，后一对词处于语义树的较低层，其语义距离较小；

概念层次树的区域密度：路径长度相同的两对结点，如果一对位于概念层次树中低密度区域，另一对位于高密度区域，那么前者的语义距离应大于后者。引入区域密度的原因在于，有些概念层次树中概念描述的粗细程度不均，例如在 Wordnet 中，动植物分类的描述极其详尽，而有些区域的概念描述又比较粗疏，这会导致语义距离计算的不合理。

另一种词语相似度的计算方法是用大规模的语料来统计。例如，利用词语的相关性来计算词语的相似度。事先选择一组特征词，然后计算这一组特征词与每一个词的相关性（一般用这组特征词在实际的大规模语料中在该词的上下文中出现的频率来度量），于是，对于每一个词都可以得到一个相关性的特征词向量，然后利用这些向量之间的相似度（一般用向量的夹角余弦来计算）作为这两个词的相似度。这种做法的假设是，凡是语义相近的词，他们的上下文也应该相似。[李涓子，1999]利用这种思想来实现语义的自动排歧；[鲁松，2001]研究了如何利用词语的相关性来计算词语的相似度。[Dagan et al. 1995,1999]使用了更为复杂的概率模型来计算词语的距离。

这两种方法各有特点。基于世界知识的方法简单有效，无需用语料库进行训练，也比较直观，易于理解，但这种方法得到的结果受人的主观意识影响较大，有时并不能准

确反映客观事实。另外，这种方法比较准确地反映了词语之间语义方面的相似性和差异，而对于词语之间的句法和语用特点考虑得比较少。基于语料库的方法比较客观，综合反映了词语在句法、语义、语用等方面的相似性和差异。但是，这种方法比较依赖于训练所用的语料库，计算量大，计算方法复杂，另外，受资料稀疏和资料噪声的干扰较大。

本文主要研究基于《知网（Hownet）》的词语相似度计算方法，这是一种基于世界知识的方法。

3. 《知网（Hownet）》简介

按照《知网》的创造者——董振东先生自己的说法[杜飞龙，1999]：

《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

《知网》中含有丰富的词汇语义知识和世界知识，为自然语言处理和机器翻译等方面的研究提供了宝贵的资源。不过，尽管《知网》提供了详细的档案[董振东，董强，1999]，但《知网》档案的形式化和规范化程度都不高。

本节中，我们将主要通过对《知网》的知识描述语言的分析，利用集合、特征结构等抽象资料形式，将《知网》的知识描述语言表示成一种更为直观、更为结构化的形式，以便于后面的相似度计算。

3.1 《知网》的结构

《知网》中有两个主要的概念：“ 概念 ”与“ 义原 ”。

“ 概念 ”是对词汇语义的一种描述。每一个词可以表达为几个概念。

“ 概念 ”是用一种“ 知识表示语言 ”来描述的，这种“ 知识表示语言 ”所用的“ 词汇 ”叫做“ 义原 ”。

“ 义原 ”是用于描述一个“ 概念 ”的最小意义单位。

与一般的语义词典[如《同义词词林》或 Wordnet]不同，《知网》并不是简单地将所有的“ 概念 ”归结到一个树状的概念层次体系中，而是试图用一系列的“ 义原 ”来对每一个“ 概念 ”进行描述。

《知网》一共采用了个 1500 义原，这些义原分为以下几个大类：

1) Event|事件

2) entity|实体

3) attribute|属性值

4) aValue|属性值

5) quantity|数量

6) qValue|数量值

7) SecondaryFeature|次要特征

8) syntax|语法

9) EventRole|动态角色

10) EventFeatures|动态属性

对于这些义原，我们把它们归为三组：第一组，包括第 1 到第 7 类的义原，我们称之为“**基本义原**”，用来描述单个概念的语义特征；第二组，只包括第 8 类义原，我们称之为“**语法义原**”，用于描述词语的语法特征，主要是词性（Part of Speech）；第三组，包括第 9 和第 10 类的义原，我们称之为“**关系义原**”，用于描述概念和概念之间的关系（类似于深层格语法中的格关系）。

除了义原以外，《知网》中还用了一些符号来对概念的语义进行描述，如下表所示：

表 1: 《知网》知识描述语言中的符号及其含义

,	多个属性之间，表示“和”的关系
#	表示“与其相关”
%	表示“是其部分”
$	表示“可以被该‘V’处置，或是该“V”的受事，对象，领有物，或者内容
*	表示“会‘V’或主要用于‘V’，即施事或工具
+	对 V 类，它表示它所标记的角色是一种隐性的，几乎在实际语言中不会出现
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“V”的空间或时间
?	表示可以是“N”的材料，如对于布匹，我们标以“?衣服”表示布匹可以是“衣服”的材料
{}	(1) 对于 V 类，置于 [ ] 中的是该类 V 所有的“必备角色”。如对于“购买”类，一旦它发生了，必然会在实际上有如下角色参与：施事，占有物，来源，工具。尽管在多数情况下，一个句子并不把全部的角色都交代出来 (2) 表示动态角色，如介词的定义
()	置于其中的应该是一个词表记，例如，(China|中国)
^	表示不存在，或没有，或不能
!	表示某一属性为一种敏感的属性，例如：“味道”对于“食物”，“高度”对于“山脉”，“温度” 对于“天象”等

[]	标识概念的共性属性

我们把这些符号又分为几类：一类是用来表示语义描述式之间的逻辑关系，我们称之为“**逻辑符号**”，包括以下几个符号：, ~ ^ ；另一类用来表示概念之间的关系，我们称之为“**关系符号**”，包括以下几个符号： # % $ * + & @ ? ! ；第三类包括几个无法归入以上两类的“**特殊符号**”：{} () [] 。

我们看到，概念之间的关系有两种表示方式：一种是用“**关系义原**”来表示，一种是用表示概念关系的“**关系符号**”来表示。按照我们的理解，前者类似于一种深层格关系，后者大部分是一种深层格关系的“反关系”，例如“$”我们就可以理解为“施事、对象、领有、内容”的反关系，也就是说，该词可以充当另一个词的“施事、对象、领有、内容”。

义原一方面作为描述概念的最基本单位，另一方面，义原之间又存在复杂的关系。在《知网》中，一共描述了义原之间的 8 种关系：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。可以看出，义原之间组成的是一个复杂的网状结构，而不是一个单纯的树状结构。不过，义原关系中最重要的还是上下位关系。根据义原的上下位关系，所有的“基本义原”组成了一个义原层次体系（如图 2）。这个义原层次体系是一个树状结构，这也是我们进行语义相似度计算的基础。

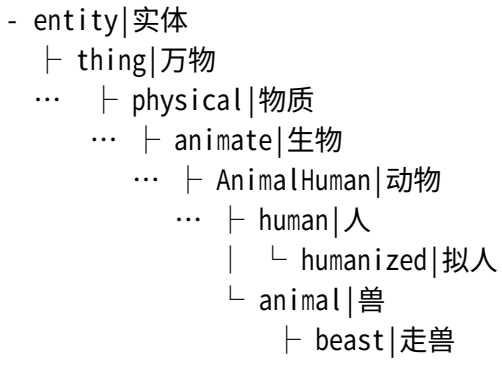

图 2 树状的义原层次结构

虽然《知网》和其它的语义词典（如《同义词词林》和 Wordnet）一样，也有一个反映知识结构的树状层次体系，但实际上有着本质的不同。在《同义词词林》和 Wordnet 中，概念是描写词义的最小单位，所以，每一个概念都是这个层次体系中的一个结点。而在《知网》中，每一个概念是通过一组义原来表示的，概念本身并不是这个层次体系中的一个结点，义原才是这个层次体系中的一个结点。而且，一个概念并不是简单的描

述为一个义原的集合，而是要描述为使用某种专门的“知识描述语言”来表达的一个语义表达式。也就是说，在描述一个概念的多个义原中，每个义原所起到的作用是不同的，这就给我们的相似度计算带来了很大的困难。下面我们就对这个描述概念的知识描述语言进行一些考察。

3.2 《知网》的知识描述语言

《知网》通过一种知识描述语言对词语的语义进行描述。在《知网》的文文件中，对知识描述语言做了详尽的介绍。不过，由于该文档过于偏重细节，不易从总体上把握。本节中我们试图对于这种知识描述语言给出一个简单的概括。

我们看几个例子：

表 2：《知网》知识描述语言实例

词	概念编号	描述语言
打	017144	exercise|锻练,sport|体育
男人	059349	human|人,family|家,male|男
高兴	029542	aValue|属性值,circumstances|境况,happy|福,desired|良
生日	072280	time|时间,day|日,@ComeToWorld|问世,$congratulate|祝贺
写信	089834	write|写,ContentProduct=letter|信件
北京	003815	place|地方,capital|国都,ProperName|专,(China|中国)
爱好者	000363	human|人,*FondOf|喜欢,#WhileAway|消闲
必须	004932	{modality|语气}
串	015204	NounUnit|名量,&(grape|葡萄),&(key|钥匙)
从良	016251	cease|停做,content=(prostitution|卖淫)
打对折	017317	subtract|削减,patient=price|价格,commercial|商,(range|幅度=50%)
儿童基金会	024083	part|部件,%institution|机构,politics|政,#young|幼,#fund|资金,(institution|机构=UN|联合国)

我们将这种知识描述语言归纳为以下几条：

1) 《知网》收入的词语主要归为两类，一类是实词，一类是虚词；

2) 虚词的描述比较简单，用“**{句法义原}**”或“**{关系义原}**”进行描述；

3) 实词的描述比较复杂，由一系列用逗号隔开的“语义描述式”组成，这些“语义描述式”又有以下三种形式：

基本义原描述式：用“基本义原”进行描述；

关系义原描述式：用“关系义原=基本义原”或者“关系义原=(具体词)”或者“(关系义原=具体词)”来描述；

关系符号描述式：用“ 关系符号 基本义原 ”或者“ 关系符号(具体词) ”加以描述，我们还注意到，可以有多个关系符号描述式采用同一个关系符号；

4) 在实词的描述中，第一个描述式总是一个**基本义原描述式**，这也是对该实词最重要的一个描述式，这个**基本义原**描述了该实词的最基本的语义特征。

根据以上分析，我们将《知网》对一个实词的义项描述重新表示如下：

$$\text{实词概念}:\begin{bmatrix}\text{第一基本义原描述}=\text{基本义原}_a\\ \text{其他基本义原描述}=\{\text{基本义原}_b,\text{基本义原}_c,\ldots\ldots\}\\ \text{关系义原描述}=\begin{bmatrix}\text{关系义原}_1=\text{基本义原}_x\mid\text{具体词}_x\\ \text{关系义原}_2=\text{基本义原}_y\mid\text{具体词}_y\\ \ldots\ldots\end{bmatrix}\\ \text{关系符号描述}=\begin{bmatrix}\text{关系符号}_1=\{\text{义原}_u\mid\text{具体词}_u,\text{义原}_v\mid\text{具体词}_v,\ldots\ldots\}\\ \text{关系符号}_2=\{\text{义原}_s\mid\text{具体词}_s,\text{义原}_t\mid\text{具体词}_t,\ldots\ldots\}\\ \ldots\ldots\end{bmatrix}\end{bmatrix}$$

在上面的表达式中，“ [……] ”表示特征结构，“ {……} ”表示集合，“ | ”表示“ 或 ”。特征结构和集合是这个表达式中使用的两种抽象数据结构，也是下面我们进行相似度计算时面对的主要问题。

4. 基于《知网》的语义相似度计算方法

从上面的介绍我们看到，与传统的语义词典不同，在《知网》中，并不是将每一个概念对应于一个树状概念层次体系中的一个结点，而是通过用一系列的义原，利用某种知识描述语言来描述一个概念。而这些义原通过上下位关系组织成一个树状义原层次体系。我们的目标是要找到一种方法，对用这种知识描述语言表示的两个语义表达式进行相似度计算。

利用《知网》计算语义相似度，一个最简单的方法就是直接使用词语语义表达式中的第一基本义原描述式，把词语相似度等价于第一基本义原的相似度。这种方法好处是计算简单，但没有利用知网语义表达式中其它部分丰富的语义信息。

[Li Sujian *et al.* 2002]中提出了一种词语语义相似度的计算方法，计算过程综合利用了《知网》和《同义词词林》。在义原相似度的计算过程中，不仅考虑了义原之间的上下位关系，还考虑了义原之间的其它关系。在计算词语相似度时，加权合并了《同义词词林》的词义相似度、《知网》语义表达式的义原相似度和义原关联度。由于《同义词词林》和《知网》采用完全不同的语义体系和表达方式，词表也相差较大，因此这种算

法中把它们合并计算的合理性值得怀疑。另外，我们前面介绍过，词语相关度和相似度是两个不同的概念，把语义关联度加权合并计入义原相似度中，是不合适的。

4.1 词语相似度计算

对于两个汉语词语 W1 和 W2，如果 W1 有 n 个义项（概念）：S11，S12，……，S1n，W2 有 m 个义项（概念）：S21，S22，……，S2m，我们规定，W1 和 W2 的相似度是各个概念的相似度之最大值，也就是说：

$$Sim(W_1,W_2)=\max_{i=1..n,\,j=1...m} Sim(S_{1i},S_{2j}) \qquad \cdots\cdots(2)$$

这样，我们就把两个词语之间的相似度问题归结到了两个概念之间的相似度问题。当然，我们这里考虑的是孤立的两个词语的相似度。如果是在一定上下文之中的两个词语，最好是先进行词义排歧，将词语标注为概念，然后再对概念计算相似度。

4.2 义原相似度计算

由于所有的概念都最终归结于用义原（个别地方用具体词）来表示，所以义原的相似度计算是概念相似度计算的基础。

由于所有的义原根据上下位关系构成了一个树状的义原层次体系，我们这里采用简单的通过语义距离计算相似度的办法。假设两个义原在这个层次体系中的路径距离为 d，根据公式(1)，我们可以得到这两个义原之间的语义距离：

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \qquad \cdots\cdots(3)$$

其中 p_1 和 p_2 表示两个义原（primitive），d 是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数。 α 是一个可调节的参数。

用这种方法计算义原相似度的时候，我们只利用了义原的上下位关系。实际上，在《知网》中，义原之间除了上下位关系外，还有很多种其它的关系，如果在计算时考虑进来，可能会得到更精细的义原相似度度量，例如，我们可以认为，具有反义或者对义关系的两个义原比较相似，因为它们在实际的语料中互相可以替换的可能性很大。对于这个问题这里我们不展开讨论。

另外，在知网的知识描述语言中，在一些义原出现的位置可能出现一个具体词（概念），并用圆括号( )括起来。所以我们在计算相似度时还要考虑到具体词和具体词、具体词和义原之间的相似度计算。理想的做法应该是先把具体词还原成《知网》的语义表达式，然后再计算相似度。这样做将导致函数的递归调用，这会使算法变得很复杂。由于具体词在《知网》的语义表达式中只占很小的比例，因此，在我们的实验中，为了简化起见，我们做如下规定：

具体词与义原的相似度一律处理为一个比较小的常数（γ）；

具体词和具体词的相似度，如果两个词相同，则为 1，否则为 0。

4.3 虚词概念的相似度的计算

我们认为，在实际的文本中，虚词和实词总是不能互相替换的，因此，虚词概念和实词概念的相似度总是为零。

由于虚词概念总是用“**{句法义原}**”或“**{关系义原}**”这两种方式进行描述，所以，虚词概念的相似度计算非常简单，只需要计算其对应的**句法义原**或**关系义原**之间的相似度即可。

4.4 实词概念的相似度的计算

从前面的分析可知，《知网》的知识描述语言可以通过义原和集合、特征结构这两种抽象数据结构来表达。义原之间的相似度计算问题已经解决，剩下的问题就是集合和特征结构的相似度问题了。

我们的基本设想是：整体相似要建立在部分相似的基础上。把一个复杂的整体分解成部分，通过计算部分之间的相似度得到整体的相似度。

假设两个整体 A 和 B 都可以分解成以下部分：A 分解成 A_1，A_2，……，A_n，B 分解成 B_1，B_2，……，B_m，那么这些部分之间的对应关系就有 $m \times n$ 种。问题是：这些部分之间的相似度是否都对整体的相似度发生影响？如果不是全部都发生影响，那么我们应该如何选择发生影响的那些部分之间的相似度？选择出来以后，我们又如何得到整体的相似度？

我们认为：一个整体的各个不同部分在整体中的作用是不同的，只有在整体中起相同作用的部分互相比较才有效。例如比较两个人长相是否相似，我们总是比较它们的脸型、轮廓、眼睛、鼻子等相同部分是否相似，而不会拿眼睛去和鼻子做比较。

因此，在比较两个整体的相似性时，我们首先要做的工作是对这两个整体的各个部分之间建立起一一对应的关系，然后在这些对应的部分之间进行比较。

还有一个问题：如果某一部分的对应物为空，如何计算其相似度？我们这里采用一种简单的处理办法：

将任一非空值与空值的相似度定义为一个比较小的常数（δ）；

下面我们分别考虑集合和特征结构的相似度计算问题。

4.4.1 特征结构的相似度计算

特征结构可以理解为一个“属性：值”对（Attribute-Value Pair）的集合，我们将一个“属性：值”对称为一个“特征”（Feature）。在一个特征结构中，每个“特征”的“属性”

是唯一的。

计算两个特征结构的相似度，首先要在两个特征结构的特征之间建立起一一对应的关系。由于每个特征结构的各个特征都具有不同的属性，因此这种一一对应关系通过特征的属性很容易建立起来：属性相同的特征之间一一对应，如果没有属性相同的特征，那么该特征的对应物为空。

这样，特征结构的相似度就转化为各个特征的相似度的加权平均。其中的权值反映出该属性在特征结构中的重要程度。在目前我们认为所有特征具有相同的重要性。

剩下的问题就是计算两个特征的相似度。特征由“属性”和“值”组成。由于“属性”相同，于是，两个特征的相似度可以等价于其“值”的相似度。

4.4.2 集合的相似度计算

集合的相似度计算比特征结构更为复杂，因为集合的元素是无序而且平等的，因此首要任务是要在两个集合的元素之间建立一一对应关系。

两个集合的相似度计算模型，必须满足我们对于集合相似度计算的一些直观要求。这里我们列出以下两条：

1. 一个集合和它本身的相似度为 1；

2. 假设两个集合都有 n 个元素，其中 m（m<n）个元素相同，又假设两个元素的相似度只能是 0（不同）或 1（相同），那么这两个集合的相似度应该是 m/n。

要计算两个集合的相似度，最容易想到的方法是首先计算两个集合的所有元素两两之间的相似度，然后再进行加权平均。但是这样会带来一个问题，就是一个集合和它本身的相似度可能不为 1，除非它的任意两个元素之间的相似度都为 1。这个结果当然是不合理的。这也从另一个角度说明我们先前定义的原则（首先在两个集合的元素之间建立一一对应关系）的合理性。

在本文中，我们采用以下算法来为两个集合的元素之间建立一一对应关系：

1. 首先计算两个集合的所有元素两两之间的相似度；
2. 从所有的相似度值中选择最大的一个，将这个相似度值对应的两个元素对应起来；
3. 从所有的相似度值中删去那些已经建立对应关系的元素的相似度值；
4. 重复上述第 2 步和第 3 步，直到所有的相似度值都被删除；
5. 没有建立起对应关系的元素与空元素对应。

根据上述算法建立起两个集合元素的一一对应关系后，我们就很容易计算两个集合的相似度了：集合的相似度等于其元素对的相似度的加权平均。又因为集合的元素之间都是平等的，所以我们可以将所有的权值取成相同的，于是：集合的相似度等于其元素对的相似度的算术平均。

4.4.3 实词概念相似度的计算

由前面的分析我们知道，在《知网》中对一个实词的描述可以表示为一个特征结构，该特征结构含有以下四个特征：

第一基本义原描述：其值为一个基本义原，我们将两个概念的这一部分的相似度记为 $Sim_1(S_1,S_2)$；

其它基本义原描述：对应于语义表达式中除第一基本义原描述式以外的所有基本义原描述式，其值为一个基本义原的集合，我们将两个概念的这一部分的相似度记为 $Sim_2(S_1,S_2)$；

关系义原描述：对应于语义表达式中所有的关系义原描述式，其值是一个特征结构，对于该特征结构的每一个特征，其属性是一个关系义原，其值是一个基本义原，或一个具体词。我们将两个概念的这一部分的相似度记为 $Sim_3(S_1,S_2)$；

关系符号描述：对应于语义表达式中所有的关系符号描述式，其值也是一个特征结构，对于该特征结构的每一个特征，其属性是一个关系义原，其值是一个集合，该集合的元素是一个基本义原，或一个具体词。我们将两个概念的这一部分的相似度记为 $Sim_4(S_1,S_2)$。

于是，两个概念语义表达式的整体相似度记为：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i Sim_i(S_1,S_2) \qquad \cdots\cdots(4)$$

其中，β_i（1≤i≤4）是可调节的参数，且有：

$\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geq\beta_2\geq\beta_3\geq\beta_4$

后者反映了 Sim_1 到 Sim_4 对于总体相似度所起到的作用依次递减。由于第一基本义原描述式反映了一个概念最主要的特征，所以我们应该将其权值定义得比较大，一般应在 0.5 以上。

在实验中我们发现，如果 Sim_1 非常小，但 Sim_3 或者 Sim_4 比较大，将导致整体的相似度仍然比较大的不合理现象。因此我们对公式(4)进行了修改，得到公式如下：

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2) \qquad \cdots\cdots(5)$$

其意义在于，主要部分的相似度值对于次要部分的相似度值起到制约作用，也就是说，如果主要部分相似度比较低，那么次要部分的相似度对于整体相似度所起到的作用也要降低。且可以保证一个词和它本身的相似度仍为 1。

下面我们再分别讨论每一部分的相似度。

第一基本义原描述：就是两个义原的相似度，按照公式(3)计算即可；

其它基本义原描述：其值为一个集合，转换为两个基本义原集合的相似度计算问题；

关系义原描述：其值为一个特征结构，转换为两个特征结构的相似度计算问题。而

这个特征结构中特征的值就是基本义原或具体词，因此这两个特征结构的相似度计算也可以最终还原到基本义原或具体词的相似度计算问题。这里，由于无法区分关系义原之间的重要程度，我们将对各个特征的相似度取算术平均；

关系符号描述：其值为一个特征结构，转换为两个特征结构的相似度计算问题。而这个特征结构中特征的值又是一个集合，集合的元素才是基本义原或具体词，因此这两个特征结构的相似度计算也可以最终还原到基本义原或具体词的相似度计算问题。同样，由于无法区分关系符号之间的重要程度，我们将对各个特征的相似度取算术平均；

到此为止，我们已经讨论了基于《知网》的词语相似度计算的所有细节，具体的算法我们不再详细说明。

5. 实验及结果

根据以上方法，我们实现了一个基于《知网》的语义相似度计算程序模块。

词语相似度计算的结果评价，最好是放到实际的系统中（如基于实例的机器翻译系统），观察不同的相似度计算方法对实际系统的性能的影响。这需要一个完整的应用系统。在条件不具备的情况下，我们采用了人工判别的方法。

我们使用了三种方法来计算词语相似度，并把它们的计算结果进行比较：

方法 1：仅使用《知网》语义表达式中第一基本义原来计算词语相似度；

方法 2：Li Sujian et al. (2002) 中使用的词语语义相似度计算方法；

方法 3：本文中介绍的语义相似度计算方法；

在实验中，根据在多次尝试中取得的经验，我们将几个参数值设置如下：

$\alpha = 1.6$

$\beta_1 = 0.5$, $\beta_2 = 0.2$, $\beta_3 = 0.17$, $\beta_4 = 0.13$

$\gamma = 0.2$

$\delta = 0.2$

实验结果如下表所示：

表 3：实验结果（一）

词语 1	词语 2	词语 2 的语义	方法 1	方法 2	方法 3
男人	女人	人,家,女	1.000	0.668	0.861
男人	父亲	人,家,男	1.000	1.000	1.000
男人	母亲	人,家,女	1.000	0.668	0.861
男人	和尚	人,宗教,男	1.000	0.668	0.861
男人	经理	人,#职位,官,商	1.000	0.351	0.630
男人	高兴	属性值,境况,福,良	0.016	0.024	0.048
男人	收音机	机器,*传播	0.186	0.008	0.112

男人	鲤鱼	鱼	0.347	0.009	0.209
男人	苹果	水果	0.285	0.004	0.171
男人	工作	事务,$担任	0.186	0.035	0.112
男人	责任	责任	0.016	0.005	0.126

考察方法 3 的结果，我们可以看到，“男人”（取义项“人，家，男”）和其它各个词的相似度与人的直觉是比较相符合的。

将方法 3 和方法 1、方法 2 的结果相比较，可以看到：方法 1 的结果比较粗糙，只要是人，相似度都为 1，显然不够合理；方法 2 的结果比方法 1 更细腻一些，能够区分不同人之间的相似度，但有些相似度的结果也不太合理，比如“男人”和“工作”的相似度比“男人”和“鲤鱼”的相似度更高。从可替换性来说，这显然不合理，至少“男人”和“鲤鱼”都是有生命物体，而“工作”只可能是一个行为或者一个抽象事物。方法 2 出现这种不合理现象的原因在于其计算方法把部分语义关联度数值加权计入了相似度中。另外，方法 2 的结果中，“男人”和“和尚”的相似度比“男人”和“经理”的相似度高出近一倍，而方法 3 的结果中，这两个相似度的差距更合理一些。

表 4 中给出另外一些测试结果，供读者参考：

表 4：实验结果（二）

词语 1	词语 2	相似度	词语 1	词语 2	相似度
工人	教师	0.722	粉红	红	1
工人	科学家	0.576	粉红	红色	1
工人	农民	0.722	粉红	绿	0.861
工人	运动员	0.722	粉红	颜色	0.059
教师	科学家	0.576	绿	颜色	0.059
教师	农民	0.722	十分	非常	1
教师	运动员	0.722	十分	特别	0.624
科学家	农民	0.576	思考	考虑	1
科学家	运动员	0.6	思考	思想	0.074
农民	运动员	0.722	考虑	思想	0.074
中国	美国	0.936	跑	跳	0.444
中国	联合国	0.136	跑	跳舞	0.127
中国	安理会	0.114	跑	运行	0.444
中国	欧洲	0.733	运行	跳舞	0.151

可以看到，绝大部分结果还是比较合理的，但也有部分结果不够合理，例如“中国”和“联合国”、“中国”和“安理会”的相似度都过低，这是因为，“中国”、“联合

国”、“安理会”在《知网》中的第一基本义原分别是“地方”、“机构”、“部件”。“跑”和“跳”的相似度也较低，这是因为这两个词被简单定义为两个基本义原，而缺少其它信息。这也从一个侧面反映了知网的某些定义不合理或不一致之处。

需要声明的是，上述试验中，每个词都只取了一个最常见的义项，而不是考虑所有义项。

6. 结论

与传统的语义词典不同，《知网》采用了 1500 多个义原，通过一种知识描述语言来对每个概念进行描述。

为了计算用知识描述语言表达的两个概念的语义表达式之间的相似度，我们采用了“整体的相似度等于部分相似度加权平均”的做法。首先将一个整体分解成部分，再将两个整体的各个部分进行组合配对，通过计算每个组合对的相似度的加权平均得到整体的相似度。我们具体讨论了特征结构和集合这两种抽象数据结构中各个组成部分的组合配对方式。通过对概念的语义表达式反复使用这一方法，可以将两个语义表达式的整体相似度分解成一些义原对的相似度的组合。对于两个义原的相似度，我们采用根据上下位关系得到语义距离并进行转换的方法。

实验证明，我们的做法充分利用了《知网》中对每个概念进行描述时的丰富的语义信息，得到的结果与人的直觉比较符合，词语相似度值刻划也比较细致。

参考文献：

Agirre E. and Rigau G., “A proposal for word sense disambiguation using conceptual distance”, *Proc. of International Conference Recent Advances in Natural Language Processing (RANLP)* , 1995, pp. 258-264, Tzigov Chark, Bulgaria.

Dagan I., Marcus S., et al. , “Contextual Word Similarity and Estimation from Sparse Data”, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* , 1993, pp. 164-171

Dagan I., Lee L. and Pereira F., “Similarity-based models of word cooccurrence probabilities”, *Machine Learning, Special issue on Machine Learning and Natural Language,* 34(1-3) , 1999, pp. 43-69

Gauch S. and Chong M. K., “Automatic Word Similarity Detection for TREC 4 Query Expansion”, *Proc. of TREC-4: The 4th Annual Text REtrieval Conf.*, Nov. 1995, Gaithersburg, MD, 1995, pp. 527-536

LI Sujian, ZHANG Jian, HUANG Xiong and BAI Shuo, “Semantic Computation in Chinese Question-Answering System”, *Journal of Computer Science and Technology* 17(6) , 2002, pp. 993-999

LI Xiaobin, Szpakowicz S., and Matwin S., “A WordNet-based algorithm for word sense disambiguation”, *Proc. of the Twelth International Joint Conference on Artificial Intelligence (IJCAI)*. 1995, pp. 1368-1374

李涓子，“汉语词义排歧方法研究”，清华大学博士论文，1999

王斌，“汉英双语语料库自动对齐研究”，中国科学院计算技术研究所博士学位论文，1999

鲁松，“自然语言中词相关性知识无导获取和均衡分类器的构建”，中国科学院计算技术研究所博士论文，2001

董振东，董强（1999），“知网”, http://www.keenage.com

杜飞龙 (1999)，《知网》辟蹊径，共享新天地——董振东先生谈知网与知识共享，《微电脑世界》杂志，1999 年第 29 期

注：本文最早发表在“第三届汉语词汇语义学研讨会（台北，2002）”上，并被会议推荐给台湾主办的学术刊物“Computational Linguistics and Chinese Language Processing”（中文名称《中文计算语言学》），经审稿录用后，作者根据审稿者提出的意见做了一定的修改。这里是修改后的版本，且刊物中使用的繁体字已转换成了简体字。在此向三位不知名的审稿者给本文提出的中肯意见以及刊物的编辑为本文付出的辛苦劳动表示衷心的感谢！作者更要向《知网》的发明人董振东先生和董强先生表示感谢，他们的工作是本文工作的基础。