Computational Linguistics and Chinese Language Processing
Vol. 7, No. 2 , August 2002, pp. 59-76

基於《知網》的辭彙語義相似度計算[1]

Word Similarity Computing Based on How-net

劉群[*] 、李素建[+]

Qun LIU , Sujian LI

摘要

詞義相似度計算在很多領域中都有廣泛的應用，例如資訊檢索、資訊抽取、文本分類、詞義排歧、基於實例的機器翻譯等等。詞義相似度計算的兩種基本方法是基於世界知識（Ontology）或某種分類體系（Taxonomy）的方法和基於統計的上下文向量空間模型方法。這兩種方法各有優缺點。

《知網》是一部比較詳盡的語義知識詞典，受到了人們普遍的重視。不過，由於《知網》中對於一個詞的語義採用的是一種多維的知識表示形式，這給詞語相似度的計算帶來了麻煩。這一點與 WordNet 和《同義詞詞林》不同。在 WordNet 和《同義詞詞林》中，所有同類的語義項（WordNet 的 synset 或《同義詞詞林》的詞群）構成一個樹狀結構，要計算語義項之間的距離，只要計算樹狀結構中相應結點的距離即可。而在《知網》中辭彙語義相似度的計算存在以下問題：

1. 每一個詞的語義描述由多個義原組成；
2. 詞語的語義描述中各個義原並不是平等的，它們之間有著複雜的關係，通過一種專門的知識描述語言來表示。

我們的工作主要包括：

1. 研究《知網》中知識描述語言的語法，瞭解其描述一個詞義所用的多個義原之間的關係，區分其在詞語相似度計算中所起的作用；我們採用一種更

[1] 本項研究受國家重點基礎研究計畫（973）支持，項目編號是 G1998030507-4 和 G1998030510。
[*] 北京大學計算語言學研究所 & 中國科學院計算技術研究所 E-mail: liuqun@ict.ac.cn
Institute of Computational Linguistics, Peking University &
Institute of Computing Technology, Chinese Academy of Science
[+] 中國科學院計算技術研究所 E-mail: lisujian@ict.ac.cn
Institute of Computing Technology, Chinese Academy of Sciences

爲結構化的方式改寫了《知網》中詞的定義（DEF），其中採用了“集合”和“特徵結構”這兩種抽象資料結構。

2. 研究了義原的相似度計算方法、集合和特徵結構的相似度計算方法，並在此基礎上提出了利用《知網》進行詞語相似度計算的演算法；

3. 通過實驗驗證該演算法的有效性，並與其他演算法進行比較。

關鍵字：《知網》 辭彙語義相似度計算 自然語言處理

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|拍攝, %tool|用具, body|身”, and the Chinese word “寫信(write a letter)” is described as: “write|寫, 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

自然語言的詞語之間有著非常複雜的關係，在實際的應用中，有時需要把這種複雜的關係用一種簡單的數量來度量，而詞義相似度就是其中的一種。

詞義相似度計算在很多領域中都有廣泛的應用，例如資訊檢索、資訊抽取、文本分類、詞義排歧、基於實例的機器翻譯等等[Gauch&Chong 1995，LI, Szpakowicz & Matwin 1995，王斌，1999，李涓子，1999]。本文的研究背景是基於實例的機器翻譯。在基於實例的機器翻譯中，詞語相似度的計算有著重要的作用。例如要翻譯“張三寫的小說”這個短語，通過語料庫檢索得到譯例：

1）李四寫的小說／the novel written by Li Si

2）去年寫的小說／the novel written last year

通過相似度計算我們發現，“張三”和“李四”都是具體的人，語義上非常相似，而“去年”的語義是時間，和“張三”相似度較低，因此我們選用“李四寫的小說”這個實例進行類比翻譯，就可以得到正確的譯文：

the novel written by Zhang San

如果選用後者作爲實例，那麼得到的錯誤譯文將是：

* the novel written Zhang San

通過這個例子可以看出相似度計算在基於實例的機器翻譯中所起的作用。

在基於實例的翻譯中另一個重要的工作是雙語對齊。在雙語對齊過程中要用到兩種語言的詞義相似度計算，這不在本文所考慮的範圍之內。

2. 詞語相似度及其計算的方法

2.1 詞語相似度的含義

詞語相似度是一個主觀性相當強的概念，沒有明確的客觀標準可以衡量。脫離具體的應用去談論詞語相似度，很難得到一個統一的定義。

本文的研究主要以基於實例的機器翻譯爲背景，因此在本文中我們所理解的詞語相似度就是兩個詞語在不同的上下文中可以互相替換使用而不改變文本的句法語義結構的程度。兩個詞語，如果在不同的上下文中可以互相替換且不改變文本的句法語義結構的可能性越大，二者的相似度就越高，否則相似度就越低。

相似度這個概念，涉及到詞語的詞法、句法、語義甚至語用等方方面面的特點。其中，對詞語相似度影響最大的應該是詞的語義。

在本文中，相似度被定義爲一個 0 到 1 之間的實數。

詞語距離與詞語相似度之間有著密切的關係。實際上，詞語距離和詞語相似度是一對詞語的相同關係特徵的不同表現形式，二者之間可以建立一種簡單的對應關係。對於兩個詞語 W_1 和 W_2，我們記其相似度爲 $Sim(W_1,W_2)$，其詞語距離爲 $Dis(W_1,W_2)$，那麼我們可以定義一個滿足以上條件的簡單轉換關係：

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

其中 α 是一個可調節的參數。α 的含義是：當相似度爲 0.5 時的詞語距離值。

這種轉換關係並不是唯一的，我們這裏只是給出了其中的一種可能。

在很多情況下，直接計算詞語的相似度比較困難，通常可以先計算詞語的距離，然後再轉換成詞語的相似度。

詞語相關性反映的是兩個詞語互相關聯的程度。可以用這兩個詞語在同一個語境中共現的可能性來衡量。詞語相關性和詞語相似性是兩個不同的概念，二者沒有直接的對應關係。

2.2 詞語相似度的計算方法

詞語距離有兩類常見的計算方法，一種是根據某種世界知識（Ontology）或分類體系（Taxonomy）來計算，一種利用大規模的語料庫進行統計。

根據世界知識（Ontology）或分類體系（Taxonomy）計算詞語語義距離的方法，一般是利用一部同義詞詞典（Thesaurus）。一般同義詞詞典都是將所有的詞組織在一棵或幾棵樹狀的層次結構中。我們知道，在一棵樹狀圖中，任何兩個結點之間有且只有一條路徑。於是，這條路徑的長度就可以作爲這兩個概念的語義距離的一種度量。

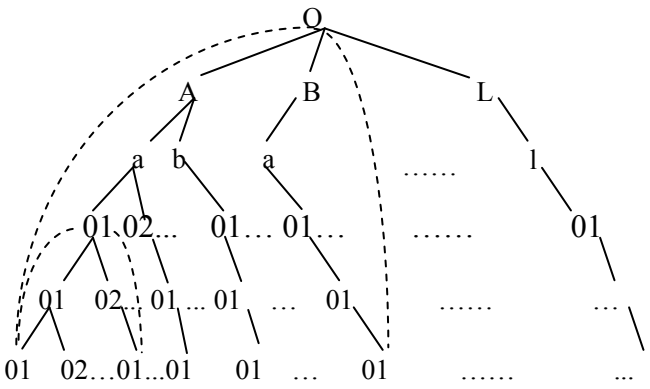

圖 1 《同義詞詞林》語義分類樹狀圖

[王斌，1999]採用這種方法利用《同義詞詞林》來計算漢語詞語之間的相似度（如圖 1 所示）。有些研究者考慮的情況**更**複雜。[Agirre & Rigau 1995]在利用 Wordnet 計算詞語的語義相似度時，除了結點間的路徑長度外，還考慮到了其他一些因素。例如：

概念層次樹的深度：路徑長度相同的兩個結點，如果位於概念層次的越高層，其語義距離較大；比如說：“動物”和“植物”、“哺乳動物”和“爬行動物”，這兩對概念間的路徑長度都是 2，但前一對詞處於語義樹的較高層，因此認爲其語義距離較大，後一對詞處於語義樹的較低層，其語義距離較小；

概念層次樹的區域密度：路徑長度相同的兩對結點，如果一對位於概念層次樹中低密度區域，另一對位元於高密度區域，那麼前者的語義距離應大於後者。引入區域密度的原因在於，有些概念層次樹中概念描述的粗細程度不均，例如在 Wordnet 中，動植物分類的描述極其詳盡，而有些區域的概念描述又比較粗疏，這會導致語義距離計算的不合理。

另一種詞語相似度的計算方法是用大規模的語料來統計。例如，利用詞語的相關性來計算詞語的相似度。事先選擇一組特徵詞，然後計算這一組特徵詞與每一個詞的相關性（一般用這組特徵詞在實際的大規模語料中在該詞的上下文中出現的頻率來度量），於是，對於每一個詞都可以得到一個相關性的特徵詞向量，然後利用這些向量之間的相似度（一般用向量的夾角餘弦來計算）作爲這兩個詞的相似度。這種做法的假設是，凡是語義相近的詞，他們的上下文也應該相似。[李涓子，1999]利用這種思想來實現語義的自動排歧；[魯松，2001]研究了如何利用詞語的相關性來計算詞語的相似度。[Dagan *et al.* 1995,1999]使用了更爲複雜的概率模型來計算詞語的距離。

這兩種方法各有特點。基於世界知識的方法簡單有效，無需用語料庫進行訓練，也比較直觀，易於理解，但這種方法得到的結果受人的主觀意識影響較大，有時並不能準確反映客觀事實。另外，這種方法比較準確地反映了詞語之間語義方面的相似性和差異，

而對於詞語之間的句法和語用特點考慮得比較少。基於語料庫的方法比較客觀，綜合反映了詞語在句法、語義、語用等方面的相似性和差異。但是，這種方法比較依賴於訓練所用的語料庫，計算量大，計算方法複雜，另外，受資料稀疏和資料雜訊的幹擾較大。

本文主要研究基於《知網（Hownet）》的詞語相似度計算方法，這是一種基於世界知識的方法。

3. 《知網（Hownet）》簡介

按照《知網》的創造者——董振東先生自己的說法[杜飛龍，1999]：

《知網》是一個以漢語和英語的詞語所代表的概念爲描述物件，以揭示概念與概念之間以及概念所具有的屬性之間的關係爲基本內容的常識知識庫。

《知網》中含有豐富的辭彙語義知識和世界知識，爲自然語言處理和機器翻譯等方面的研究提供了寶貴的資源。不過，儘管《知網》提供了詳細的檔案[董振東，董強，1999]，但《知網》檔案的形式化和規範化程度都不高。

本節中，我們將主要通過對《知網》的知識描述語言的分析，利用集合、特徵結構等抽象資料形式，將《知網》的知識描述語言表示成一種更爲直觀、更爲結構化的形式，以便於後面的相似度計算。

3.1 《知網》的結構

《知網》中有兩個主要的概念：“概念”與“義原”。

“概念”是對辭彙語義的一種描述。每一個詞可以表達爲幾個概念。

“概念”是用一種“知識表示語言”來描述的，這種“知識表示語言”所用的“辭彙”叫做“義原”。

“義原”是用於描述一個“概念”的最小意義單位。

與一般的語義詞典[如《同義詞詞林》或 Wordnet]不同，《知網》並不是簡單地將所有的“概念”歸結到一個樹狀的概念層次體系中，而是試圖用一系列的“義原”來對每一個“概念”進行描述。

《知網》一共採用了個 1500 義原，這些義原分爲以下幾個大類：

1) Event|事件

2) entity|實體

3) attribute|屬性值

4) aValue|屬性值

5) quantity|數量

6) qValue|數量值

7) SecondaryFeature|次要特徵

8) syntax|語法

9) EventRole|動態角色

10) EventFeatures|動態屬性

對於這些義原，我們把它們歸爲三組：第一組，包括第 1 到第 7 類的義原，我們稱之爲“**基本義原**”，用來描述單個概念的語義特徵；第二組，只包括第 8 類義原，我們稱之爲“**語法義原**”，用於描述詞語的語法特徵，主要是詞性（Part of Speech）；第三組，包括第 9 和第 10 類的義原，我們稱之爲“**關係義原**”，用於描述概念和概念之間的關係（類似於深層格語法中的格關係）。

除了義原以外，《知網》中還用了一些符號來對概念的語義進行描述，如下表所示：

表 1: 《知網》知識描述語言中的符號及其含義

,	多個屬性之間，表示“和”的關係
#	表示“與其相關”
%	表示“是其部分”
$	表示“可以被該‘V’處置，或是該“V”的受事，物件，領有物，或者內容
*	表示“會‘V’或主要用於‘V’，即施事或工具
+	對 V 類，它表示它所標記的角色是一種隱性的，幾乎在實際語言中不會出現
&	表示指向
~	表示多半是，多半有，很可能的
@	表示可以做“V”的空間或時間
?	表示可以是“N”的材料，如對於布匹，我們標以“?衣服”表示布匹可以是“衣服”的材料
{}	(1) 對於 V 類，置於 [ ] 中的是該類 V 所有的“必備角色”。如對於“購買”類，一旦它發生了，必然會在實際上有如下角色參與：施事，佔有物，來源，工具。儘管在多數情況下，一個句子並不把全部的角色都交代出來 (2) 表示動態角色，如介詞的定義
()	置於其中的應該是一個詞表記，例如，(Chinal中國)
^	表示不存在，或沒有，或不能
!	表示某一屬性爲一種敏感的屬性，例如：“味道”對於“食物”，“高度”對於“山脈”，“溫度” 對於“天象”等
[]	標識概念的共性屬性

我們把這些符號又分爲幾類：一類是用來表示語義描述式之間的邏輯關係，我們稱之爲“**邏輯符號**”，包括以下幾個符號：, ~ ^ ；另一類用來表示概念之間的關係，我們稱之爲“**關係符號**”，包括以下幾個符號： # % $ * + & @ ? ! ；第三類包括幾個無法歸入以上兩類的“**特殊符號**”：{} () [] 。

我們看到，概念之間的關係有兩種表示方式：一種是用“**關係義原**”來表示，一種是用表示概念關係的“**關係符號**”來表示。按照我們的理解，前者類似於一種深層格關係，後者大部分是一種深層格關係的“反關係”，例如“$”我們就可以理解爲“施事、物件、領有、內容”的反關係，也就是說，該詞可以充當另一個詞的“施事、物件、領有、內容”。

義原一方面作爲描述概念的最基本單位，另一方面，義原之間又存在複雜的關係。在《知網》中，一共描述了義原之間的 8 種關係：上下位關係、同義關係、反義關係、對義關係、屬性-宿主關係、部件-整體關係、材料-成品關係、事件-角色關係。可以看出，義原之間組成的是一個複雜的網狀結構，而不是一個單純的樹狀結構。不過，義原關係中最重要的還是上下位關係。根據義原的上下位關係，所有的“基本義原”組成了一個義原層次體系（如圖 2）。這個義原層次體系是一個樹狀結構，這也是我們進行語義相似度計算的基礎。

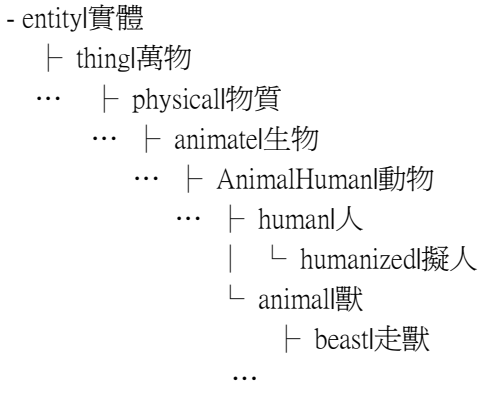

圖 2 樹狀的義原層次結構

雖然《知網》和其他的語義詞典（如《同義詞詞林》和 Wordnet）一樣，也有一個反映知識結構的樹狀層次體系，但實際上有著本質的不同。在《同義詞詞林》和 Wordnet 中，概念是描寫詞義的最小單位，所以，每一個概念都是這個層次體系中的一個結點。而在《知網》中，每一個概念是通過一組義原來表示的，概念本身並不是這個層次體系中的一個結點，義原才是這個層次體系中的一個結點。而且，一個概念並不是簡單的描述爲一個義原的集合，而是要描述爲使用某種專門的“知識描述語言”來表達的一個語義運算式。也就是說，在描述一個概念的多個義原中，每個義原所起到的作用是不同的，這就給我們的相似度計算帶來了很大的困難。下面我們就對這個描述概念的知識描述語

言進行一些考察。

3.2 《知網》的知識描述語言

《知網》通過一種知識描述語言對詞語的語義進行描述。在《知網》的文檔中，對知識描述語言做了詳盡的介紹。不過，由於該文檔過於偏重細節，不易從總體上把握。本節中我們試圖對於這種知識描述語言給出一個簡單的概括。

我們看幾個例子：

表 2：《知網》知識描述語言實例

詞	概念編號	描述語言
打	017144	exercise|鍛練,sport|體育
男人	059349	human|人,family|家,male|男
高興	029542	aValue|屬性值,circumstances|境況,happy|福,desired|良
生日	072280	time|時間,day|日,@ComeToWorld|問世,$congratulate|祝賀
寫信	089834	write|寫,ContentProduct=letter|信件
北京	003815	place|地方,capital|國都,ProperName|專,(China|中國)
愛好者	000363	human|人,*FondOf|喜歡,#WhileAway|消閒
必須	004932	{modality|語氣}
串	015204	NounUnit|名量,&(grape|葡萄),&(key|鑰匙)
從良	016251	cease|停做,content=(prostitution|賣淫)
打對折	017317	subtract|削減,patient=price|價格,commercial|商,(range|幅度=50%)
兒童基金會	024083	part|部件,%institution|機構,politics|政,#young|幼,#fund|資金,(institution|機構=UN|聯合國)

我們將這種知識描述語言歸納爲以下幾條：

1) 《知網》收入的詞語主要歸爲兩類，一類是實詞，一類是虛詞；

2) 虛詞的描述比較簡單，用“**{句法義原}**”或“**{關係義原}**”進行描述；

3) 實詞的描述比較複雜，由一系列用逗號隔開的“語義描述式”組成，這些“語義描述式”又有以下三種形式：

基本義原描述式：用“基本義原”進行描述；

關係義原描述式：用“關係義原=基本義原”或者“關係義原=(具體詞)”或者“(關係義原=具體詞)”來描述；

關係符號描述式：用“關係符號 基本義原”或者“關係符號(具體詞)”加以描述，我們還注意到，可以有多個關係符號描述式採用同一個關係符號；

4) 在實詞的描述中，第一個描述式總是一個**基本義原描述式**，這也是對該實詞最重

要的一個描述式，這個**基本義原**描述了該實詞的最基本的語義特徵。

根據以上分析，我們將《知網》對一個實詞的義項描述重新表示如下：

$$
\text{實詞概念}:\left[\begin{array}{l}
\text{第一基本義原描述}=\text{基本義原}_a \\
\text{其他基本義原描述}=\{\text{基本義原}_b,\text{基本義原}_c,\ldots\} \\
\text{關係義原描述}=\left[\begin{array}{l}
\text{關係義原}_1=\text{基本義原}_x\,|\,\text{具體詞}_x \\
\text{關係義原}_2=\text{基本義原}_y\,|\,\text{具體詞}_y \\
\ldots\ldots
\end{array}\right] \\
\text{關係符號描述}=\left[\begin{array}{l}
\text{關係符號}_1=\{\text{義原}_u\,|\,\text{具体詞}_u,\text{義原}_v\,|\,\text{具体詞}_v,\ldots\} \\
\text{關係符號}_2=\{\text{義原}_s\,|\,\text{具体詞}_s,\text{義原}_t\,|\,\text{具体詞}_t,\ldots\} \\
\ldots\ldots
\end{array}\right]
\end{array}\right]
$$

在上面的運算式中，“〔……〕”表示特徵結構，“{……}”表示集合，“|”表示“或”。特徵結構和集合是這個運算式中使用的兩種抽象資料結構，也是下面我們進行相似度計算時面對的主要問題。

4. 基於《知網》的語義相似度計算方法

從上面的介紹我們看到，與傳統的語義詞典不同，在《知網》中，並不是將每一個概念對應於一個樹狀概念層次體系中的一個結點，而是通過用一系列的義原，利用某種知識描述語言來描述一個概念。而這些義原通過上下位關係組織成一個樹狀義原層次體系。我們的目標是要找到一種方法，對用這種知識描述語言表示的兩個語義運算式進行相似度計算。

利用《知網》計算語義相似度，一個最簡單的方法就是直接使用詞語語義運算式中的第一基本義原描述式，把詞語相似度等價於第一基本義原的相似度。這種方法好處是計算簡單，但沒有利用知網語義運算式中其他部分豐富的語義資訊。

[Li Sujian *et al.* 2002]中提出了一種詞語語義相似度的計算方法，計算過程綜合利用了《知網》和《同義詞詞林》。在義原相似度的計算過程中，不僅考慮了義原之間的上下位關係，還考慮了義原之間的其他關係。在計算詞語相似度時，加權合併了《同義詞詞林》的詞義相似度、《知網》語義運算式的義原相似度和義原關聯度。由於《同義詞詞林》和《知網》採用完全不同的語義體系和表達方式，詞表也相差較大，因此這種演算法中把它們合併計算的合理性值得懷疑。另外，我們前面介紹過，詞語相關度和相似度是兩個不同的概念，把語義關聯度加權合併計入義原相似度中，是不合適的。

4.1 詞語相似度計算

對於兩個漢語詞語 W_1 和 W_2，如果 W_1 有 n 個義項（概念）：S_{11}，S_{12}，……，S_{1n}，W_2 有 m 個義項（概念）：S_{21}，S_{22}，……，S_{2m}，我們規定，W_1 和 W_2 的相似度是各個概念的相似度之最大值，也就是說：

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

這樣，我們就把兩個詞語之間的相似度問題歸結到了兩個概念之間的相似度問題。當然，我們這裏考慮的是孤立的兩個詞語的相似度。如果是在一定上下文之中的兩個詞語，最好是先進行詞義排歧，將詞語標注爲概念，然後再對概念計算相似度。

4.2 義原相似度計算

由於所有的概念都最終歸結于用義原（個別地方用具體詞）來表示，所以義原的相似度計算是概念相似度計算的基礎。

由於所有的義原根據上下位關係構成了一個樹狀的義原層次體系，我們這裏採用簡單的通過語義距離計算相似度的辦法。假設兩個義原在這個層次體系中的路徑距離爲 d，根據公式(1)，我們可以得到這兩個義原之間的語義距離：

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

其中 p_1 和 p_2 表示兩個義原（primitive），d 是 p_1 和 p_2 在義原層次體系中的路徑長度，是一個正整數。α 是一個可調節的參數。

用這種方法計算義原相似度的時候，我們只利用了義原的上下位關係。實際上，在《知網》中，義原之間除了上下位關係外，還有很多種其他的關係，如果在計算時考慮進來，可能會得到更精細的義原相似度度量，例如，我們可以認爲，具有反義或者對義關係的兩個義原比較相似，因爲它們在實際的語料中互相可以替換的可能性很大。對於這個問題這裏我們不展開討論。

另外，在知網的知識描述語言中，在一些義原出現的位置可能出現一個具體詞（概念），並用圓括號( )括起來。所以我們在計算相似度時還要考慮到具體詞和具體詞、具體詞和義原之間的相似度計算。理想的做法應該是先把具體詞還原成《知網》的語義運算式，然後再計算相似度。這樣做將導致函數的遞迴調用，這會使演算法變得很複雜。由於具體詞在《知網》的語義運算式中只占很小的比例，因此，在我們的實驗中，爲了簡化起見，我們做如下規定：

具體詞與義原的相似度一律處理爲一個比較小的常數（γ）；

具體詞和具體詞的相似度，如果兩個詞相同，則爲 1，否則爲 0。

4.3 虛詞概念的相似度的計算

我們認爲，在實際的文本中，虛詞和實詞總是不能互相替換的，因此，虛詞概念和實詞概念的相似度總是爲零。

由於虛詞概念總是用“**{句法義原}**”或“**{關係義原}**”這兩種方式進行描述，所以，虛詞概念的相似度計算非常簡單，只需要計算其對應的**句法義原**或**關係義原**之間的相似度即可。

4.4 實詞概念的相似度的計算

從前面的分析可知，《知網》的知識描述語言可以通過義原和集合、特徵結構這兩種抽象資料結構來表達。義原之間的相似度計算問題已經解決，剩下的問題就是集合和特徵結構的相似度問題了。

我們的基本設想是：整體相似要建立在部分相似的基礎上。把一個複雜的整體分解成部分，通過計算部分之間的相似度得到整體的相似度。

假設兩個整體 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，那麼這些部分之間的對應關係就有 mxn 種。問題是：這些部分之間的相似度是否都對整體的相似度發生影響？如果不是全部都發生影響，那麼我們應該如何選擇發生影響的那些部分之間的相似度？選擇出來以後，我們又如何得到整體的相似度？

我們認爲：一個整體的各個不同部分在整體中的作用是不同的，只有在整體中起相同作用的部分互相比較才有效。例如比較兩個人長相是否相似，我們總是比較它們的臉型、輪廓、眼睛、鼻子等相同部分是否相似，而不會拿眼睛去和鼻子做比較。

因此，在比較兩個整體的相似性時，我們首先要做的工作是對這兩個整體的各個部分之間建立起一一對應的關係，然後在這些對應的部分之間進行比較。

還有一個問題：如果某一部分的對應物爲空，如何計算其相似度？我們這裏採用一種簡單的處理辦法：

將任一非空值與空值的相似度定義爲一個比較小的常數（δ）；

下面我們分別考慮集合和特徵結構的相似度計算問題。

4.4.1 特徵結構的相似度計算

特徵結構可以理解爲一個“屬性：值”對（Attribute-Value Pair）的集合，我們將一個“屬性：值”對稱爲一個“特徵”（Feature）。在一個特徵結構中，每個“特徵”的“屬性”是唯一的。

計算兩個特徵結構的相似度，首先要在兩個特徵結構的特徵之間建立起一一對應的關係。由於每個特徵結構的各個特徵都具有不同的屬性，因此這種一一對應關係通過特徵的屬性很容易建立起來：屬性相同的特徵之間一一對應，如果沒有屬性相同的特徵，那麼該特徵的對應物爲空。

這樣，特徵結構的相似度就轉化爲各個特徵的相似度的加權平均。其中的權值反映出該屬性在特徵結構中的重要程度。在目前我們認爲所有特徵具有相同的重要性。

剩下的問題就是計算兩個特徵的相似度。特徵由“屬性”和“值”組成。由於“屬性”相同，於是，兩個特徵的相似度可以等價於其“值”的相似度。

4.4.2 集合的相似度計算

集合的相似度計算比特徵結構更爲複雜，因爲集合的元素是無序而且平等的，因此首要任務是要在兩個集合的元素之間建立一一對應關係。

兩個集合的相似度計算模型，必須滿足我們對於集合相似度計算的一些直觀要求。這裏我們列出以下兩條：

1. 一個集合和它本身的相似度爲 1；

2. 假設兩個集合都有 n 個元素，其中 m（m<n）個元素相同，又假設兩個元素的相似度只能是 0（不同）或 1（相同），那麼這兩個集合的相似度應該是 m/n。

要計算兩個集合的相似度，最容易想到的方法是首先計算兩個集合的所有元素兩兩之間的相似度，然後再進行加權平均。但是這樣會帶來一個問題，就是一個集合和它本身的相似度可能不爲 1，除非它的任意兩個元素之間的相似度都爲 1。這個結果當然是不合理的。這也從另一個角度說明我們先前定義的原則（首先在兩個集合的元素之間建立一一對應關係）的合理性。

在本文中，我們採用以下演算法來爲兩個集合的元素之間建立一一對應關係：

1. 首先計算兩個集合的所有元素兩兩之間的相似度；
2. 從所有的相似度值中選擇最大的一個，將這個相似度值對應的兩個元素對應起來；
3. 從所有的相似度值中刪去那些已經建立對應關係的元素的相似度值；
4. 重複上述第 2 步和第 3 步，直到所有的相似度值都被刪除；
5. 沒有建立起對應關係的元素與空元素對應。

根據上述演算法建立起兩個集合元素的一一對應關係後，我們就很容易計算兩個集合的相似度了：集合的相似度等於其元素對的相似度的加權平均。又因爲集合的元素之間都是平等的，所以我們可以將所有的權值取成相同的，於是：集合的相似度等於其元素對的相似度的算術平均。

4.4.3 實詞概念相似度的計算

由前面的分析我們知道，在《知網》中對一個實詞的描述可以表示爲一個特徵結構，該特徵結構含有以下四個特徵：

第一基本義原描述：其值爲一個基本義原，我們將兩個概念的這一部分的相似度記爲 $Sim_1(S_1,S_2)$；

其他基本義原描述：對應於語義運算式中除第一基本義原描述式以外的所有基本義原描述式，其值爲一個基本義原的集合，我們將兩個概念的這一部分的相似度記爲 $Sim_2(S_1,S_2)$；

關係義原描述：對應於語義運算式中所有的關係義原描述式，其值是一個特徵結構，對於該特徵結構的每一個特徵，其屬性是一個關係義原，其值是一個基本義原，或一個具體詞。我們將兩個概念的這一部分的相似度記爲 $Sim_3(S_1,S_2)$；

關係符號描述：對應於語義運算式中所有的關係符號描述式，其值也是一個特徵結構，對於該特徵結構的每一個特徵，其屬性是一個關係義原，其值是一個集合，該集合的元素是一個基本義原，或一個具體詞。我們將兩個概念的這一部分的相似度記爲 $Sim_4(S_1,S_2)$。

於是，兩個概念語義運算式的整體相似度記爲：

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

其中，β_i（$1\le i\le 4$）是可調節的參數，且有：

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

後者反映了 Sim_1 到 Sim_4 對於總體相似度所起到的作用依次遞減。由於第一基本義原描述式反映了一個概念最主要的特徵，所以我們應該將其權值定義得比較大，一般應在 0.5 以上。

在實驗中我們發現，如果 Sim_1 非常小，但 Sim_3 或者 Sim_4 比較大，將導致整體的相似度仍然比較大的不合理現象。因此我們對公式(4)進行了修改，得到公式如下：

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

其意義在於，主要部分的相似度值對於次要部分的相似度值起到制約作用，也就是說，如果主要部分相似度比較低，那麼次要部分的相似度對於整體相似度所起到的作用也要降低。且可以保證一個詞和它本身的相似度仍爲 1。

下面我們再分別討論每一部分的相似度。

第一基本義原描述：就是兩個義原的相似度，按照公式(3)計算即可；

其他基本義原描述：其值爲一個集合，轉換爲兩個基本義原集合的相似度計算問題；

關係義原描述：其值爲一個特徵結構，轉換爲兩個特徵結構的相似度計算問題。而這個特徵結構中特徵的值就是基本義原或具體詞，因此這兩個特徵結構的相似度計算也可以最終還原到基本義原或具體詞的相似度計算問題。這裏，由於無法區分關係義原之間的重要程度，我們將對各個特徵的相似度取算術平均；

關係符號描述：其值爲一個特徵結構，轉換爲兩個特徵結構的相似度計算問題。而這個特徵結構中特徵的值又是一個集合，集合的元素才是基本義原或具體詞，因此這兩

個特徵結構的相似度計算也可以最終還原到基本義原或具體詞的相似度計算問題。同樣，由於無法區分關係符號之間的重要程度，我們將對各個特徵的相似度取算術平均；

到此爲止，我們已經討論了基於《知網》的詞語相似度計算的所有細節，具體的演算法我們不再詳細說明。

5. 實驗及結果

根據以上方法，我們實現了一個基於《知網》的語義相似度計算程式模組。

詞語相似度計算的結果評價，最好是放到實際的系統中（如基於實例的機器翻譯系統），觀察不同的相似度計算方法對實際系統的性能的影響。這需要一個完整的應用系統。在條件不具備的情況下，我們採用了人工判別的方法。

我們使用了三種方法來計算詞語相似度，並把它們的計算結果進行比較：

方法 1：僅使用《知網》語義運算式中第一基本義原來計算詞語相似度；

方法 2：Li Sujian et al. (2002) 中使用的詞語語義相似度計算方法；

方法 3：本文中介紹的語義相似度計算方法；

在實驗中，根據在多次嘗試中取得的經驗，我們將幾個參數值設置如下：

$\alpha = 1.6$

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

$\gamma = 0.2$

$\delta = 0.2$

實驗結果如下表所示：

表 3：實驗結果（一）

詞語 1	詞語 2	詞語 2 的語義	方法 1	方法 2	方法 3
男人	女人	人,家,女	1.000	0.668	0.861
男人	父親	人,家,男	1.000	1.000	1.000
男人	母親	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	經理	人,#職位,官,商	1.000	0.351	0.630
男人	高興	屬性值,境況,福,良	0.016	0.024	0.048
男人	收音機	機器,*傳播	0.186	0.008	0.112
男人	鯉魚	魚	0.347	0.009	0.209
男人	蘋果	水果	0.285	0.004	0.171
男人	工作	事務,$擔任	0.186	0.035	0.112
男人	責任	責任	0.016	0.005	0.126

考察方法 3 的結果，我們可以看到，“男人”（取義項“人，家，男”）和其他各個詞的相似度與人的直覺是比較相符合的。

將方法 3 和方法 1、方法 2 的結果相比較，可以看到：方法 1 的結果比較粗糙，只要是人，相似度都爲 1，顯然不夠合理；方法 2 的結果比方法 1 更細膩一些，能夠區分不同人之間的相似度，但有些相似度的結果也不太合理，比如“男人”和“工作”的相似度比“男人”和“鯉魚”的相似度更高。從可替換性來說，這顯然不合理，至少“男人”和“鯉魚”都是有生命物體，而“工作”只可能是一個行爲或者一個抽象事物。方法 2 出現這種不合理現象的原因在於其計算方法把部分語義關聯度數值加權計入了相似度中。另外，方法 2 的結果中，“男人”和“和尚”的相似度比“男人”和“經理”的相似度高出近一倍，而方法 3 的結果中，這兩個相似度的差距更合理一些。

表 4 中給出另外一些測試結果，供讀者參考：

表 4：實驗結果（二）

詞語 1	詞語 2	相似度	詞語 1	詞語 2	相似度
工人	教師	0.722	粉紅	紅	1
工人	科學家	0.576	粉紅	紅色	1
工人	農民	0.722	粉紅	綠	0.861
工人	運動員	0.722	粉紅	顏色	0.059
教師	科學家	0.576	綠	顏色	0.059
教師	農民	0.722	十分	非常	1
教師	運動員	0.722	十分	特別	0.624
科學家	農民	0.576	思考	考慮	1
科學家	運動員	0.6	思考	思想	0.074
農民	運動員	0.722	考慮	思想	0.074
中國	美國	0.936	跑	跳	0.444
中國	聯合國	0.136	跑	跳舞	0.127
中國	安理會	0.114	跑	運行	0.444
中國	歐洲	0.733	運行	跳舞	0.151

可以看到，絕大部分結果還是比較合理的，但也有部分結果不夠合理，例如“中國”和“聯合國”、“中國”和“安理會”的相似度都過低，這是因爲，“中國”、“聯合國”、“安理會”在《知網》中的第一基本義原分別是“地方”、“機構”、“部件”。“跑”和“跳”的相似度也較低，這是因爲這兩個詞被簡單定義爲兩個基本義原，而缺少其他資訊。這也從一個側面反映了知網的某些定義不合理或不一致之處。

需要聲明的是，上述試驗中，每個詞都只取了一個最常見的義項，而不是考慮所有義項。

6. 結論

與傳統的語義詞典不同，《知網》採用了 1500 多個義原，通過一種知識描述語言來對每個概念進行描述。

爲了計算用知識描述語言表達的兩個概念的語義運算式之間的相似度，我們採用了“整體的相似度等於部分相似度加權平均”的做法。首先將一個整體分解成部分，再將兩個整體的各個部分進行組合配對，通過計算每個組合對的相似度的加權平均得到整體的相似度。我們具體討論了特徵結構和集合這兩種抽象資料結構中各個組成部分的組合配對方式。通過對概念的語義運算式反復使用這一方法，可以將兩個語義運算式的整體相似度分解成一些義原對的相似度的組合。對於兩個義原的相似度，我們採用根據上下位關係得到語義距離並進行轉換的方法。

實驗證明，我們的做法充分利用了《知網》中對每個概念進行描述時的豐富的語義資訊，得到的結果與人的直覺比較符合，詞語相似度值刻劃也比較細緻。

參考文獻：

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，“漢語詞義排歧方法研究”，清華大學博士論文，1999

王斌，“漢英雙語語料庫自動對齊研究”，中國科學院計算技術研究所博士學位論文，1999

魯松，“自然語言中詞相關性知識無導獲取和均衡分類器的構建”，中國科學院計算技術研究所博士論文，2001

董振東，董強（1999），“知網”, http://www.keenage.com

杜飛龍 (1999)，《知網》辟蹊徑，共用新天地——董振東先生談知網與知識共用，《微電腦世界》雜誌，1999 年第 29 期