第 33 卷 第 6 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.6 Computer Engineering March 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)06—0191—04 文献标识码：A 中图分类号：TP391

汉语词语语义相似度计算研究

夏 天
(中国人民大学信息资源管理学院，北京 100872)

摘 要：汉语词语的语义相似度计算是中文信息处理中的一个关键问题。该文提出了一种基于知网、面向语义、可扩展的相似度计算新方法，该方法从信息论的角度出发，定义了知网义原间的相似度计算公式，通过对未登录词进行概念切分和语义自动生成，解决了未登录词无法参与语义计算的难题，实现了任意词语在语义层面上的相似度计算。针对同义词词林的实验结果表明，该方法的准确率比现有方法高出近 15 个百分点。

关键词：词语相似度；知网；概念；义原

Study on Chinese Words Semantic Similarity Computation

XIA Tian
(School of Information Resource Management, Renmin University of China, Beijing 100872)

【Abstract】 Similarity computation of Chinese words is a key problem in Chinese information processing. This paper proposes a new method on similarity computation which is based on Hownet, geared to semantic and could be expanded. The new method defines a similarity computation formula among Hownet's sememes according to information theory, finds a way out of the difficulty that OOV words cannot participate in semantic computation by implementing concept segmentation and automatic semantic production to OOV words, and realizes the similarity computation on the semantic level among arbitrary words finally. Experimental result of CILIN indicates that the accuracy rate of the new method is nearly 15% higher than present ones.

【Key words】 Words similarity; Hownet; Concept; Sememe

汉语词汇相似度计算在自动问答、情报检索、文本聚类等应用中都是一个非常关键的问题[1,2]。针对这一问题，人们已经做了大量研究，并提出了一些定量计算方法，如字面相似度算法和以词素为处理单位的相似度测定[3]、基于某一语义分类体系的相似度计算[4]等，其中，文献[4]以知网为基础的词汇相似度计算是当前较好的方法之一，在业内有着一定程度的应用[5]。

语义词典本身有一个完备性问题：它不可能收录现实应用中的所有词语，底层词法分析的结果也不能保证与词典完全一致。传统的语义方法必然会有部分词语不在词典中，无法计算相似度，以LCMC[6]语料在知网中的出现情况统计表明，这一比例高达 7.67%。为解决以上问题，本文基于知网[7]，提出了一种面向语义、可扩展的相似度计算方法，取得了较好的效果。

1 实现过程

为便于处理，本文针对登录词和未登录词分别进行计算，登录词和未登录词的界定以知网为标准，即知网中出现的词语作为登录词，否则为未登录词。

1.1 义原相似度计算

知网中的概念由多个义原表示，要计算概念之间的相似度，首先要计算义原之间的相似度。

从信息论的角度来说，两个事物的相似度不仅与其个性有关，更应与其共性有关[8]。在图 1 中，义原“鱼”和“水果”的相似度一方面取决于它们不同的语义距离，另一方面还与它们所包含的共同部分密切相关，即应由D_1、D_2和D_3 3 个参数共同决定最终取值。为便于说明，做如下定义：

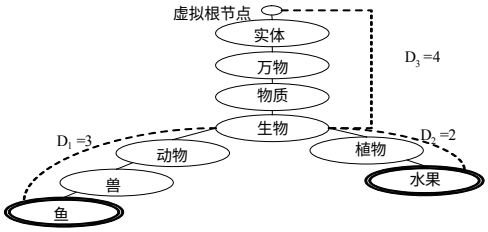

图 1 知网义原树的一个子片段

定义 1 义原深度。指义原 p 在整体义原层次体系中所处的层数位置，记为 $Depth(p)$。

为方便逻辑处理，把知网中各类不同的义原体系用一个虚拟根节点统一起来，构成一个相互联系的有机整体，并规定根节点的义原深度为 0，它的子节点深度为 1，其它依次类推。

例 1 在图 1 中，$Depth$(“鱼”)=7。

定义 2 重合度(Superposed Degree)。指两个义原p_1和p_2在义原层次体系中所拥有的相同父节点的路径长度，记为$Spd(p_1, p_2)$。

例 2 图 1 中，Spd(“鱼”,“水果”)=4。

定义 3 相异度(Dissimilitude Degree)。指 2 个义原p_1和p_2在义原层次体系中沿父节点逐步上移，直到二者达到第一个

基金项目：中国人民大学科学研究青年基金资助项目；数据工程与知识工程教育部重点实验室(中国人民大学)开放课题基金资助项目
作者简介：夏 天(1978－)，男，博士、讲师，主研方向：自然语言处理，知识工程
收稿日期：2006-04-08 **E-mail：**cnxiatian@163.com

共同节点，所走过的最短路径长度，记为$Dsd(p_1, p_2)$。相异度与语义距离等价。

例 3 图 1 中，Dsd(“物质”,“植物”)=2。

根据以上分析，结合义原深度、重合度和相异度对义原的不同度量方式，定义义原相似度计算公式如下：

$$\mathrm{Sim}(p_1, p_2) = \frac{2 \times \mathrm{Spd}(p_1, p_2)}{\mathrm{Dsd}(p_1, p_2) + 2 \times \mathrm{Spd}(p_1, p_2)} = \frac{2 \times \mathrm{Spd}(p_1, p_2)}{\mathrm{Depth}(p_1) + \mathrm{Depth}(p_2)} \quad (1)$$

例 4 $\mathrm{Sim}(\text{“鱼”},\text{“水果”}) = \frac{2 \times 4}{7 + 6} = \frac{8}{13}$

知网在对概念描述时，有时会出现具体词，并用圆括号括起来。鉴于具体词在语义表达式中所占比重较低，特作如下规定：(1)具体词与义原的相似度均为一个较小的常数 γ；(2)具体词与具体词相似度在两词相同时为 1，否则为 0；(3)义原与空值相似度均为一个较小的常数 δ。

1.2 登录词的相似度计算

对于知网而言，登录词分为实词和虚词两类。其中，实词和虚词差别很大，可直接令实词和虚词的概念相似度为 0。对于虚词而言，因为知网总是用“{句法义原}”或“{关系义原}”进行描述，所以只需计算去掉花括弧后的句法义原或关系义原的相似度即可。

在知网中，实词概念可分成 4 个部分：(1)第一基本义原描述式，DEF项中的第一个义原；(2)其他基本义原描述式，DEF项中除第一义原外的其他独立义原或具体词；(3)关系义原描述式，DEF项中用“关系义原=基本义原”或“关系义原=(具体词)”或“(关系义原=具体词)”描述概念的部分；(4)符号义原描述式，DEF项中用“关系符号基本义原”或者“关系符号(具体词)”描述概念的部分。此处把任意两概念C_1和C_2各部分的相似度分别记为$\mathrm{Sim}_1(C_1, C_2)$、$\mathrm{Sim}_2(C_1, C_2)$、$\mathrm{Sim}_3(C_1, C_2)$和$\mathrm{Sim}_4(C_1, C_2)$。并令整体相似度为

$$\mathrm{Sim}(C_1, C_2) = \beta_1 \mathrm{Sim}_1(C_1, C_2) + \sum_{i=2}^{4} \beta_1 \beta_i \mathrm{Sim}_i(C_1, C_2) \quad (2)$$

其中，β_i ($1 \le i \le 4$)是一个可调节的参数，各部分的重要程度通过β_i进行限定，并满足：$\beta_1+\beta_2+\beta_3+\beta_4 = 1$，$\beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4 > 0$。式中$\beta_1$与$\beta_i$相乘的意义在于，概念中主要义原起了决定性作用，其相似度将对次要部分的相似度起较强的制约作用，其它 3 部分则相对独立，。

下面讨论每一部分的相似度：

(1)第一基本义原描述式：根据式(1)计算。

(2)其他基本义原描述式：假设两个概念C_1和C_2的其他基本义原分别为

$\mathrm{Set}_1=\{p_{11}, p_{12}, \ldots, p_{1m}\}$

$\mathrm{Set}_2=\{p_{21}, p_{22}, \ldots, p_{2n}\}$

则该部分的计算过程如下：

令SIZE=max{|Set_1|, |Set_2|}，|Set_1|和|Set_2|分别表示 2 个集合当前拥有的义原数量，即

```
score=0.0;
while(|Set1| > 0 or |Set2| > 0){
求出 2 个集合所有组合中相似度最大的一组义原pI∈Set1和pJ∈Set2；
    score = score + Sim(pI, pJ);
    Set1 = Set1 – {pI};
    Set2 = Set2 – {pJ};
}
Sim2 (C1, C2) = score/SIZE;
```

(3)关系义原描述式：与(2)类似，该部分也是一个集合运算问题，不同点在于，集合在两两配对分组计算关系义原的相似度时，需要首先判断关系类型是否相同，不同时相似度为 0，否则，取出具体的义原名称计算其语义相似度。

(4)符号义原描述式：计算方式与(3)相同。

1.3 未登录词的相似度计算

未登录词可分解为两个或多个登录词，如能根据组成登录词的相关概念获得其可能的语义表达式，就可以进行相似度计算。即，未登录词的相似度可通过对应的组合概念获得，涉及 3 个过程：未登录词的概念切分，组合概念的语义确定和组合概念的相似度计算，下面逐一论述。

1.3.1 未登录词的概念切分

每个未登录词均由多个概念组合而成，利用逆向最大匹配法将其切分成多个概念，如

“信息化”→“信息”＋“化”

“歌舞厅”→“歌”＋“舞厅”

“快餐店”→“快餐”＋“店”

未登录词的概念切分也存在切分歧义问题，该问题本文不展开讨论。

1.3.2 组合概念的语义确定

假设未登录词切分为$C_1, C_2, \ldots, C_n$共n个概念，如何确定这n个概念的组合意义呢？汉语在词语组合时具有“重心后移”的特点，切分结果集当中的第i个概念C_i可以作为对部分概念$C_{i+1}, C_{i+2}, \ldots, C_n$组合意义的一种限制或补充。

定义 4 原子概念(Primitive Concept)。知网中本身存在的概念，称为原子概念。每一个登录词都对应于一个原子概念，原子概念是未登录词的切分依据，它具有确定的语义表达式。

定义 5 基本组合概念(Basic Combined Concept)。由两个原子概念组合在一起所构成的新词语，称为这两个原子概念的基本组合概念。基本组合概念的语义表达式可以通过对应的原子概念的语义表达式推理获取。

定义 6 扩展组合概念(Extended Combined Concept)。由两个以上原子概念或者原子概念与基本组合概念组合在一起所构成的新词语，称为扩展组合概念。扩展组合概念的语义表达式可以由它所辖的原子概念和基本组合概念的语义表达式推理获取。

定义 7 组合概念(Combined Concept)。基本组合概念和扩展组合概念统称为组合概念。组合概念的具体含义可根据上下文环境予以确定。

假设基本组合概念BCC由原子概念PC_1和PC_2组成，即BCC= PC_1 PC_2，并且PC_1和PC_2的语义表达式分别为Def(PC_1)和Def(PC_2)，则BCC的语义表达式可以简单地表示为

$$\mathrm{Def(BCC)=Def}(PC_2) \cup \mathrm{Def}(PC_1) \quad (3)$$

同时，根据“重心后移”原则，令PC_2的第一基本义原作为BCC的第一基本义原。令PS(C)表示概念C的第一基本义原，则有PS(BCC)=PS(PC_2)。

例 5 基本组合概念“娱乐场”根据式(3)推理得到的语义表达式：

因为 Def(“娱乐”)={recreation|娱乐,entertainment|艺}

Def(“场”)={ InstitutePlace|场所,*sell|卖,@buy|买,commercial|商}

所以 Def(“娱乐场”)=Def(“娱乐”)∪Def(“场”)={InstitutePlace|

场所, recreation|娱乐,*sell|卖,@buy|买,commercial|商,entertainment|艺}

知网在对概念进行描述时，除了独立义原外，还存在关系义原和符号义原两种描述形式，而式(3)直接合并原子概念义原的方法，在许多情况下并不能准确地反映两个义原之间的相互关系。如例 5，简单的义原合并未能把组合概念“娱乐场”中“InstitutePlace|场所”和“recreation|娱乐”两义原所应具有的时空关系表示出来。而意义相同的原子概念“游乐场”则描述得较为准确。

Def(“娱乐场”)= {InstitutePlace|场所, @recreation|娱乐}

导致上述差别的原因在于：知网中概念的编纂是一个有人工参与的过程，每个概念的定义都凝聚了人的智慧，定义较为准确；而让计算机接近或者达到人的智能程度，在短期内还不太现实，如没有其他信息加以参照，让机器确定组合概念各义原之间具有何种关系非常困难。为解决该问题，本文引入了“参照概念”。

定义 8 参照概念。为确定基本组合概念各义原之间所具有的关系而提供的一个原子概念称为参照概念。

例如，把“游乐场”看作参照概念，就可以确定“娱乐场”中的义原“recreation|娱乐”应加上时空关系符号“@”，即变为“@recreation|娱乐”。同时，原子概念“场”与参照概念的第一基本义原一致，因此，对于“场”而言，可以只保留在参照概念的义原集合中出现的部分，即去掉“*sell|卖，@buy|买，commercial|商”。实际上，对“场”的语义限制或扩充的最大可能是通过另外一个原子概念“娱乐”的语义予以实现。假设基本组合概念BCC= PC_1 PC_2，在以原子概念PC_{ref}作为参照概念时，生成语义表达式的算法如下：

令Def(BCC)=Def(PC_2),且PS(BCC)=PS(PC2);
REF_SET= Def(PC_{ref})－{ PS(PC_{ref})};
t=Sim(PS(BCC), PS(PC_{ref}));
IF t≥ΘTHEN
Def(BCC)=PS(BCC)∪(Def(BCC)∩REF_SET);
ENDIF
FOR EACH p_i IN Def (PC_1) DO
max-sim=0, pos=0;
FOR EACH q_j IN REF_SET DO
如p_i或q_j为关系或符号义原，滤掉符号以普通义原参与运算
IF Sim(p_i, q_j)>max-sim THEN
max-sim= Sim(p_i, q_j); pos=j;
ENDIF
ENDFOR
IF (pos>0) AND (t×max-sim >Ω) THEN
p←p_i, 且p的附加符号与q_{pos}相同;
ELSE
p←p_i ;
ENDIF
Def(BCC)= Def(BCC)∪{p};
ENDFOR

其中，0≤Θ≤1, 0≤Ω≤1 表示参照概念起作用的阈值。在知网中，符号义原或关系义原一般是针对第一基本义原而言的，因此，如果组合概念与参照概念的第一基本义原相似度差别较大，则参照能力也应随之降低。

为此，算法首先从参照义原集合当中找出和待加入义原最相似的一个，计算其相似度，再乘上参照概念与组合概念的第一基本义原的相似度，以达到主义原对当前义原的制约作用。

算法中Def(BCC)∩REF_SET部分的具体运算与通常的集合相交稍有不同，过程如下：(1)每次从两个集合中各任意抽取一个义原，计算其相似度，设p_1∈Def(BCC)和p_2∈REF_SET为两两计算之后的最大者；(2)令Ξ为一阈值，如Sim(p_1,p_2)≥Ξ，则保留p_1作为相交结果，同时删除p_1和p_2，否则，直接删除p_1；(3)循环执行步骤(1)、步骤(2)，直到Def(BCC)为空时或者REF_SET为空时停止。

例 6 基本组合概念“娱乐场”以“游乐场”为参照概念时得到的语义表达式，即

因为 Def(“游乐场”)={InstitutePlace 场所,@recreation|娱乐}
Def(“娱乐”)={recreation|娱乐,entertainment|艺}
Def(“场”)={InstitutePlace|场所,*sell|卖,@buy|买,commercial|商}
所以 Def(“娱乐场”)={InstitutePlace|场所, @recreation|娱乐, entertainment|艺}

扩展组合概念的语义生成可采用逆向合并的方式。假设扩展组合概念ECC=PC_1,PC_2,…,PC_n，其语义表达式的生成过程如下：(1)首先取出最后两个原子概念PC_{n-1}和PC_n，并采用基本组合概念的语义计算方式计算其组合语义，记为Def(ECC)；(2)把Def(ECC)看作是某个原子概念的语义表达式，从i=n-2~1，逐步合并Def(ECC)和PC_i，并把每次合并结果重新记为Def(ECC)；(3)Def(ECC)即为扩展组合概念ECC的语义表达式。

1.3.3 组合概念的相似度计算

组合概念相似度计算又可分为两种：(1)组合概念(未登录词)和原子概念(登录词)的相似度计算；(2)组合概念与组合概念的相似度计算。

对于第(1)种情况，可以把参与运算的原子概念作为组合概念的参照概念，求解组合概念的语义表达式，进而计算两个语义集合的相似度。对第(2)种情况本文采取了一种简化策略：首先根据式(3)计算两个组合概念各自的语义表达式，然后相互以对方为参照概念修正自己的语义关系。最后以修正后的语义集合进行相似度计算。

2 实验结果

由于对相似度计算方法的优劣评判还没有一个较为通用的标准，因此本文采用了 2 种方式：(1)采用人工判别的方法，给出了部分词语的相似度；(2)以同义词词林(以下简称词林)为测试数据，并设定一个相似度阈值，如果两个词的相似度大于这一值，则认为是同义词，这样对词林中的每一组同义词进行统计分析。

由于时间和资源限制，因此本文只实现并对比了 3 种方法：文献[3]的字面相似度算法，文献[4]的语义相似度算法，本文提出的方法。

在本实验中，各方法对应参数的取值分别如下：

方法 1 α=0.6,β=0.4；

方法 2 α=1.6,β1=0.5,β2=0.2, β3=0.17, β4=0.13, γ=0.2, δ=0.2；

方法 3 β1=0.5,β2=0.2, β3=0.17, β4=0.13,γ=0.2,δ=0.2,Θ=0.5, Ξ=0.6, Ω=0.8；

第(1)种方式的结果如表 1 所示(未登录词用下划线予以标注)。可以看到，方法 1 没有涉及词汇语义问题，大部分结果都不理想；方法 2 对于登录词的计算相对合理，但对未登录词却无法处理；方法 3 则对登录词和未登录词的计算结果都比较合理。

表 1 部分实验结果对比

词语 1	词语 2	方法 1	方法 2	方法 3
男人	女人	0.550	0.861	0.967
男人	父亲	0.000	1.000	1.000
男人	男孩	0.450	0.750	0.900
男人	工作	0.000	0.108	0.200
自行车	电动车	0.383	0.000	0.840
娱乐场	商城	0.000	0.000	0.867
娱乐场	游乐场	0.717	0.000	0.840
北京	北京市	0.733	0.000	1.000
快餐店	饭店	0.408	0.000	0.900
电子书	电子图书	0.804	0.000	0.933

第(2)种方式对词林中每组同义词两两计算其相似度，如相似度超过指定阈值，则把它们看作为同义词，结果如图 2。

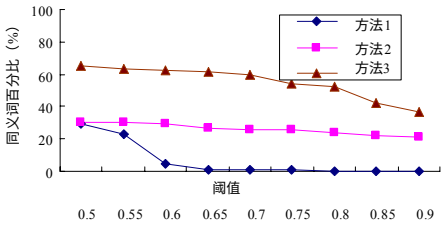

图 2 词林同义词的相似度统计

从图 2 看出，当阈值为 0.5 时，方法 1 与方法 2 的效果基本相同，随着阈值的提高，方法 1 的识别率急剧下降，方法 2 相对平稳。方法 3 的效果最好，即使阈值取到了 0.9，其识别率仍然要比方法 2 高出近 15 个百分点。

3 结论

本文提出了一种基于语义、可扩展的相似度计算方法，并解决了未登录词的相似度计算难题。根据实验结果，得出以下结论：(1)单纯的字面相似度算法以字为处理单位，不涉及语义信息，在大规模应用中效果较差；(2)基于某种语义体系的相似度算法，尤其是以文献[4]为代表的方法，与字面相似度算法相比，准确率大大提高。但由于未登录词问题，因此降低了该类方法的准确率和扩展性；(3)本文提出的计算方法，可通过原子概念并根据参照概念，自动获取未登录词对应的组合概念，进而计算其相似度，取得了更好的效果。

参考文献

1 刘亚军，徐　易. 一种基于加权语义相似度模型的自动问答系统[J]. 东南大学学报, 2004, 34(5): 609-612.

2 李有梅. 基于词义的关键词抽取方法研究[J]. 情报理论与实践, 2000, 23(2): 81-83.

3 朱毅华，侯汉清，沙印亭. 计算机识别汉语同义词的两种算法比较和测评[J]. 中国图书馆学报, 2002, 28(140): 82-85.

4 刘　群，李素建. 基于《知网》的词汇语义相似度计算[C]//第 3 届中文词汇语义学研讨会论文集. 2002-05.

5 夏　天，樊孝忠，刘　林. 基于 ALICE 的汉语自然语言接口[J]. 北京理工大学学报, 2004, 24(10): 885-889.

6 McEnery T. The Lancaster Corpus of Mandarin Chinese[EB/OL]. 2004-10. http:// www.ling.lancs.ac.uk/corplang/lcmc/.

7 董振东，董　强. 知网[Z]. 2002-12. http://www.keenage.com/.

8 Dekang L. An Information-theoretic Definition of Similarity[C]// Proceedings of the 15th International Conference on Machine Learning. 1998: 296-304.

(上接第 140 页)

输的内容进行加密和认证。SSL 协议由记录协议和握手协议组成，建立在传输层和应用层之间，提供的安全信道有私密性、确认性、可靠性 3 个特性。记录协议主要完成分组和组合、压缩和解压缩，以及消息认证和加密等功能。握手协议描述安全连接建立的过程，在客户和服务器传送应用层数据之前，完成加密算法、密钥加密密钥算法的确定，以及交换预主密钥，并最后产生相应的客户和服务器 MAC 密钥、会话加密密钥等功能。

SSL 协议已经是成熟的安全通信协议，且已经集成于很多编程工具中，在实现时很方便，因此采用 SSL 协议来进一步保证系统内部的通信安全。

2.6 其他模块

为完善主机检测部分，还研制了 2 个附加模块：文件完整性检测模块和文件实时监控模块。文件完整性检测模块采用 MD5 散列算法实现，首先对需要检测的重要文件建立编码数据库，检测时再次调用检测模块生成编码数据，通过比较两次编码数据来确定文件是否被修改。首次生成的文件编码数据发送到数据库服务器保存。文件实时监控模块利用操作系统的 API 函数，可以实时地检测所监视的文件的添加、删除、修改等情况。

此外，为了更方便地查看系统报警信息，还开发了一套报警信息管理系统，并实现了按不同的条件进行统计告警信息的功能。任意节点均可以通过浏览器查看报警信息，实现了任意时间、地点且远程监控网络的目的。

3 结束语

本文提出了一种分布式自治性入侵检测系统，采用 2 层结构框架减少了控制层次，通过结合了协议分析和模式匹配技术的自治性检测节点来实现分布式检测，每个节点均能独立完成入侵检测功能，并采用一种自定义的安全通信协议及 SSL 协议来保证通信安全，可在任意节点通过 B/S 模式浏览告警信息，实现了入侵检测系统的分布式、自治性、高效检测和良好的抗攻击能力。下一步还需要在灵活性、智能性、与其他安全部件的协同性等方面做进一步的改进，使得本系统功能更加全面，能够检测更多的攻击和更好地保障系统自身的安全。

参考文献

1 薛静锋，宁宇鹏，阎　慧. 入侵检测技术[M]. 北京：机械工业出版社, 2004.

2 Snapp S R, Brentano J, Dias G V, et al. Distributed Intrusion Detection System ——Motivation, Architecture, and an Early Prototype[C]//Proceedings of the 14th National Computer Security Conference. 1991.

3 Porras P, Schnackenberg D, Staniford-Chen S, et al. The Common Intrusion Detection Framework Architecture[EB/OL]. 2005. http:// www.isi. edu/gost/cidf/drafts/architecture.txt.