基于语义树的中文词语相似度计算与分析*

张亮[1,2] 尹存燕[1] 陈家骏[1]

1.南京大学计算机软件新技术国家重点实验室 江苏南京 210093

2.江苏警官学院公安科技系 南京 210000 E-mail:zhangliangg@163.com

摘 要：基于语义资源 Hownet 的词语相似度计算是近年来的研究热点，但大多数研究都是对中科院计算所刘群提出的计算方法[1]的改进和完善。本文充分分析和利用新版 Hownet(2007)的概念架构和语义多维表达形式，从概念的主类义原、主类义原框架以及概念特性描述三个方面综合分析词语相似度，并在计算中区分语义特征相似度和句法特征相似度。实验结果理想，与人的直观判断基本一致。

关键词：语义树、词语相似度、知网 2007、语义距离

Chinese Word Similarity Computing Based on Semantic Tree

Zhang Liang[1,2] ,Yin Cunyan[1], Chen Jiajun[1]

1. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093

2. Jiangsu Police Institute, Nanjing 210000 E-mail:zhangliangg@163.com

Abstract: Chinese words similarity computing based on Hownet has become a hot point at present, but most research is the improvement and refinement of Liu-Qun's method[1]. Based on the new Hownet(2007), this paper make the best use of Hownet concept frame and semantic multi-dimension expression form, and proposes a new method which analyzes and processes Chinese words similarity from the three dimensions: the main sememe, the main sememe frame and concept characteristic description. Furthermore and the method distinguishes semantic similarity and syntax similarity. The result of experiment shows that the method has a good performance.

Keywords: Semantic Tree, Words Similarity, Hownet2007, Distance of Semantic

1. 引言

词是研究语句的基本语义单元和句法单位，词之间的相似度与相关度的分析是词语研究的核心内容之一。常见的词语相似度计算方法有两类[1]，一种是根据世界知识(ontology)或分类体系(taxonomy)计算，一种是利用大规模语料库进行统计。它们各有特点：前者简单有效，较直观易理解，与人的学习理解方式类似，但需要有较完备的知识库的支撑，这种方法较准确地反映了词语语义相似性和差异，而对句法和语用特点考虑较少；后者比较客观，综合反映了词语在句法、语义、语用等方面的相似性和差异，但较依赖训练语料，计算量大，计算方法复杂，受数据稀疏和数据噪声的干扰较大，有时会出现明显的错误。

本文基于新版《知网》进行词语相似度计算，从功用角度将词语相似度细化为语义特征相似度和句法特征相似度，改进了传统的基于知网的相似度计算方法，通过构建多维语义树，将词相似分析转化为树的相似分析，设计了一个新的词语相似度计算模型。

2. 语义资源平台《知网》

知网(HowNet)[2]是目前使用较为广泛的语义资源平台，它以概念为描述对象，揭示概

*本文承国家 863 高技术发展研究计划(编号:2006AA010109);国家自然科学基金(编号:60673043) 的资助。

念与概念之间以及概念所具有的属性之间的关系。知网定义了一套释义元语言即义原（sememe）对概念进行刻画和描述，义原是意义最小单位，具有唯一性和确定性。由义原构筑起来的知网概念层次体系（Taxonomy）是一棵概念分类树，如图 1 所示，将所有的概念从 top-down 的视角划分为事件 Event、实体 Entity、属性 Attribute、属性值 Attribute Value、次要特征 Secondary Feature 等几个特征类别。知网描述了概念之间和概念属性之间的各种关系,主要包括上下位关系、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[3]。知网具有如下特点：

（1）释义元语言——义原的定义和使用，使得概念描述具有较好的概括性和确定性；

（2）设计了一种知识词典描述语言(KDML)对知识形式化，提高了概念的可计算性。

（3）概念定义时纵向归类与横向关联相结合，描述结构清晰，便于计算机处理。

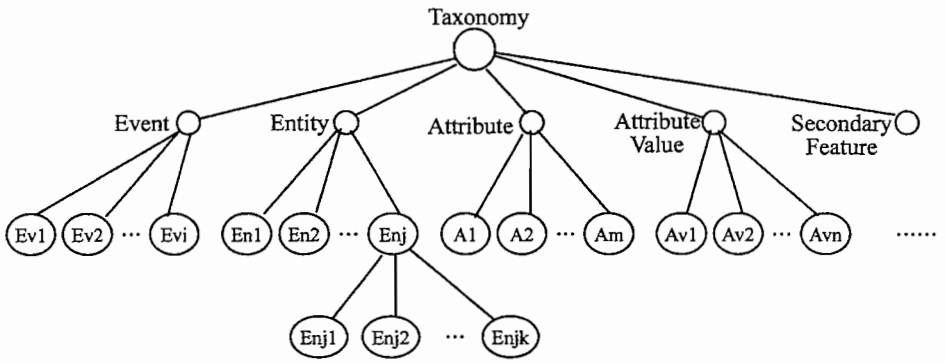

图 1　HowNet 的概念层次结构

3．词语相似度讨论

Dekang Lin 认为任何两个事物的相似度取决于它们的共性(Commonality)和个性(Differentces)[4]，他从信息理论的角度给出任意两个事物相似度的通用公式:

$$Sim(A,B)=\frac{\log P(common(A,B))}{\log P(description(A,B))} \qquad (1)$$

其中分子是描述 A、B 共性的信息量，分母是完整的描述出 A、B 的信息量大小。

1）Dekang Lin 的这一理论是目前绝大多数中文词语语义相似度计算模型的基本思想，尽管刘群等提出两个词语的相似度是它们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大小，但在其计算模型中并没有突出可替换性这一特征。词语相似度主要从语义分析的角度出发，通过比较词的义项，计算共同部分的大小，比较抽象的讨论和计算两个词之间的语义距离，这虽然可以为信息检索、机器翻译等其他自然语言处理提供一定的帮助，但是功用性不是很强，语义距离与可替换性有时并不一致，如 Similarity(盗墓人，盗墓) > Similarity(盗墓人，小偷)，即语义相似度与可替换性并不一定正相关。

2）我们认为词语相似度是一个比较粗泛的概念，根据应用需求，可以细化为语义特征相似度和句法特征相似度。语义特征相似度，也就是在同一个语境中，两个词相互替换，而不改变整个语境的语义；句法特征相似度，也就是两个词互换，而不改变原有的句法结构或依存关系，这对基于语料库的句法结构排岐有很好的帮助作用。语义特征相似度高则句法特征相似度高，反之不一定。如“他认为这是一个好主意。”其中的“好”，被“馊”替换，语义相反，但是句法关系不变，因此在基于语料的句法分析中，这类语义相反，但

句法结构一样的语料，同样具有很好的参考价值。

4. 基于知网的词语相似度分析

从相关文献看，目前基于知网的语言分析与处理绝大多数还是以旧版本（知网 2000）作为平台，其实新版（知网 2007）的概念描述架构已经有了质的不同，概念的定义由主类义原及其特性描述组成：1）主类义原相当于旧版中的第一义原，是所定义概念的最基本的意义；2）特性描述利用动态角色和特征标注复杂概念，内容上体现概念之间的关联，形式上可以为嵌套结构。整个概念的定义可以转化为一棵语义树，如图 2 所示。

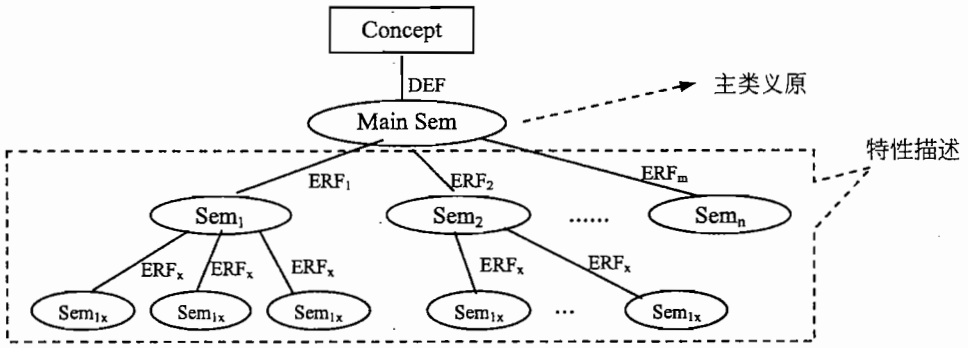

图 2　HowNet 中概念的描述框架

在知网中词用概念来描述的,一个词可以表达为几个概念，而概念用义原来描述。假设词 W_1 有 n 个概念 $C_{11},C_{12},\cdots,C_{1n}$,词 W_2 有 m 个概念 $C_{21},C_{22},\cdots,C_{2m}$，则词语 W_1 和 W_2 的相似度是其所有概念之间相似度的最大值，如公式（2）所示，其符号取该对概念相似度的符号。

$$Sim(W_1,W_2)=\pm\max_{i=1..n,j=1..m}\left|Sim(C_{1i},C_{2j})\right| \tag{2}$$

根据知网的概念定义，两个概念之间的相似度计算可以从以下几个方面进行：

4.1 两个概念的主类义原相似度计算

主类义原确定了概念的最基本的意义，实际上是给概念尽可能细地分类，主类义原相似度计算核心是如何计算两个义原的语义距离。义原相似度的计算一般依据义原的层次体系(上下位关系) 来计算,这种基于树状层次结构计算语义相似度的研究已经十分成熟。Eneko Agirre[5]、Dekang Lin、刘群等都提出了自己的公式，BUDAN-ITSKY 对基于 WordNet 的几种计算方法进行了比较[6]。他们的方法可以分为两大类：一种是基于两个节点之间的路径长度，一种是基于两个节点所含的共有信息大小。

我们认为，义原相似度应当同时反映出两个义原在树中的距离和两个义原公共信息的大小，同时由于 Taxonomy 中的义原树具有语义分类内涵以及结点上下位关系，处于下位的结点与上位结点同类，并且是在上位结点的语义基础上，加入更多的语义成分。结点承载的语义信息量与其到根结点的距离正相关，结点语义信息的重要程度与其到根结点的距离负相关，即离根结点越近，对语义区分的贡献值越大。

公式（3）是义原相似度计算公式，其中 α 是一个可调节的参数，取值越大则层次的区分越小，m、h、n 分别为结点 1、结点 2 和结点 1 与结点 2 的最近的共同祖先的层次数。

$$Sim\ (C_1,C_2)=\frac{2\times\sum_{i=1}^{n}1/(\alpha+i)}{\sum_{j=1}^{m}1/(\alpha+j)+\sum_{k=1}^{h}1/(\alpha+k)} \tag{3}$$

4.2 两个概念的语义树相似度计算

新版知网中概念的描述是一棵以主类义原为根的语义树，树中每一个结点都是一个义原，除根结点外，每个结点与其父结点之间的关系用一个动态角色和特征加以标识。概念的相似度，是指概念类型相似程度以及概念中相同的特性描述的广度与深度。这样概念描述的相似度计算就转化为对应的两个语义树的最大匹配。图 3 所示，词语“儿科医生”与“患儿”的概念描述中，主类义原都是“人”，虚线部分勾勒出两棵语义树中最大相似部分。

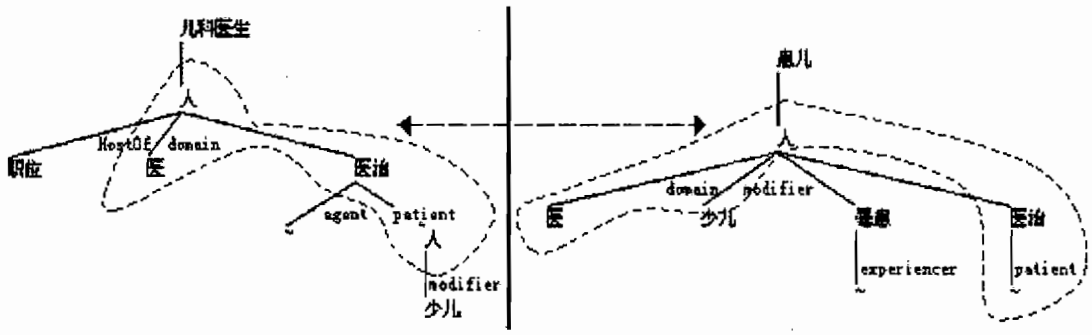

图 3　词语“儿科医生”与“患儿”的概念语义树对照图（虚线勾勒出最大相似部分）

计算两棵语义树的最大相似度算法：

1）将两个概念描述分别转换为两棵 Tree1,Tree2（根结点为主类义原，除根结点外，其他结点除包含本身的义原外，还有与父结点的关系值，即动态角色和特征）；

2）广度优先遍历 Tree1，将遍历结点存入队列 Queue 中；

3）如果 Queue≠∅，取出 Queue 中第一个元素，赋值给变量 Tree1_x；否则转 7）

4）广度优先遍历 Tree2，若存在某结点与 Tree1_x 相等，则将其赋值给 Tree2_y，并转 5）；若遍历完毕，则转 3）；

5）分别在 Tree1 和 Tree2 中，检查是否存在 Tree1_x 的子结点与 Tree2_y 的子结点相等，并且对应的动态角色和特征相等，若存在分别将其存入队列 Queue1 和 Queue2，转 6）；若不存在，则转 4）；

6）如果 Queue1=∅，则转 4）；否则取出 Queue1 中第一个元素，赋值给变量 Tree1_x；取出 Queue2 中第一个元素，赋值给变量 Tree2_y，转 5）；

7）结束处理。

注：结束处理是将匹配中记录下来的相似块按大小和先后顺序进行比较，大者优先、前者优先；广度优先遍历，可以保证如果存在同样的相似块，则层次高的排在前面。

4.3 两个概念的主类义原框架相似度计算

如前所述，主类义原是对一个概念的根本属性的规定，是概念的第一义原，在形式上表现为紧邻标识符“DEF”后面的义原。所谓义原框架是对在义原树上的义原的本质属性的描述，是在语义分类的基础上，对义原本身语义的更细致的刻划。如“人”作为 entity 分类树上的一个义原结点，其义原框架为：DEF={AnimalHuman|动物:HostOf={Ability|能力}{Name|姓名}{Wisdom|智慧},{speak|说:agent={~}},{think|思考:agent={~}}}。两个概念的主类义原框架相似度计算是在提取两个概念的主类义原的基础上，做义原框架的相似度计

算，是对 4.1 节中主类义原相似度计算的补充，其计算方法与概念语义树相似度计算一致。

4.4 两个概念反义和对义关系的计算

从语义树角度看，反义义原结点之间（或对义义原结点之间）的绝对距离不大，甚至很小，它们之间语义距离大，是描述对象属性或动态特征的语义极性的表现，如“喜欢”与“厌恶”、“抽象”与“具体”等，其在义原树上的垂直关系如下所示：

喜欢: 事件 → 静态 → 状态 → 精神状态 → 态度 → 好态 → 喜欢

厌恶: 事件 → 静态 → 状态 → 精神状态 → 态度 → 坏态 → 厌恶

由于对象属性或动态特征的语义极性存在，我们给出本文中概念的相似度的定义。

定义 1：概念 C1,C2 的相似度 Similarity(C1,C2)是在知网概念描述框架基础上,根据 C1,C2 的概念类别、语义特性描述和主类义原框架等几个方面的相似程度，并考虑 C1,C2 的语义极性，而计算出的一个综合值。Similarity(C1,C2)∈[-1，+1]，Similarity(C1,C2)反映 C1,C2 的语义特征相似度，Similarity(C1,C2)的绝对值反映 C1,C2 的句法特征相似度。

如果概念 C1、C2 是反义或对义关系，则 Similarity(C1,C2)=-1。如果概念 C1、C2 的上位义原是反义或对义关系，则 Similarity(C1,C2)=-1。如果概念 C1、C2 中存在反义或对义关系，则 Similarity(C1,C2)为负值。根据知网反义词表和对义词表进行反义和对义的计算。

4.5 词语相似度的综合计算

我们在 4.1---4.2 的基础上，下面给出基于知网的词语相似度的完整计算公式。

$$Similarity(C_1,C_2)=\begin{cases}-1 & \text{当且仅当}C_1,C_2\text{或其上位义原为反义词 或对义词}\\ \theta\times(\beta_1\times Sim_1(C_1,C_2)+\beta_2\times\gamma\times Sim_2(C_1,C_2) & \\ \qquad +\beta_3\times Sim_3(C_1,C_2)) & \end{cases} \tag{4}$$

其中 C1,C2 是进行相似度计算的两个概念，θ 是决定 Similarity(C1,C2)符号的系数，如果 C1,C2 概念特性描述中含有反义或对义关系，则 θ =-1，否则 θ =+1；Sim_1(C1,C2)是 C1,C2 的主类义原相似度计算，Sim_2(C1,C2)是 C1,C2 的语义树相似度计算，Sim_3(C1,C2)是 C1,C2 的主类义原框架相似度计算 β1、β2、β3 分别是对应计算的权重，β1+β2+β3=1，β3≤β1≤β2。γ 为惩罚因子，如果 C1,C2 的特性描述中存在某个共同的 Event，并同时存在依附于该 Event 的不同的动态角色与特征关系，则 γ =0.5，否则 γ =1。如在词语“儿科医生”与“患儿”的概念语义树对照图中，“儿科医生”与“患儿”具有相同的主类义原“人”，在特性描述中都有“医治”这个 Event，但在“儿科医生”中，主类义原“人”是 Event“医治”的 agent；“患儿”中，主类义原“人”是 Event“医治”的 patient，即他们的动态角色不同。

由于概念相似度的计算的功用目的，有必要突出语义分析和句法分析中的词语可替换性。深入研究知网的概念表达体系结构，我们发现概念的主类义原确定了概念的最基本的意义，是概念语义分类的依据，而无论是词语的语义特征还是句法特征，都与概念语义分类密切相关，为体现这一特性，我们将公式（4）修正为公式（5）。

$$Similarity(C_1,C_2)=\begin{cases}-1 & \text{当且仅当}C_1,C_2\text{其或上位义原为反义词 或对义词}\\ \theta\times(\beta_1\times Sim_1(C_1,C_2)+\beta_2\times\gamma\times Sim_2(C_1,C_2)\times Sim_1(C_1,C_2) & \\ \qquad +\beta_3\times Sim_3(C_1,C_2)) & \end{cases} \tag{5}$$

在公式（4）的基础上，对语义树相似度计算 Sim_2(C1,C2)乘上主类义原相似度计算 Sim_1(C1,C2)，这一修改的含义是：如果 Sim_1(C1,C2)值较大，则公式（4）的计算值接近于

公式（5），如果 $Sim_1(C1,C2)$值较小，则第二项的计算值较小，整个公式的计算值就较小。

5. 相关实验与结果分析

由于目前对中文词语相似度还没有形成统一的规范，也没有相关标注语料提供实验平台，因此中文词语相似度计算的实验设计与数据筛选困难较大，如果随机的选取一些词语，很难说明问题。我们从本文中文词语相似度的功用目的出发，在遴选实验数据时侧重那些能说明语义特征和句法特征以及语义相关性的词语。

表 1　相关参数设置

参数符号	参数组 1	参数组 2	参数组 3	参数说明
α	0	1	4	调节主类义原在义原树上的深度的影响
β1	0.3	0.2	0.2	主类义原相似度所占权重
β2	0.4	0.7	0.3	概念语义树相似度所占权重
β3	0.3	0.1	0.5	主类义原框架相似度所占权重
γ	0.5 or 1	0.5 or 1	0.5 or 1	动态角色与特征关系的惩罚因子，需惩罚取 0.5，否则取 1
θ	+1 or -1	+1 or -1	+1 or -1	决定 Similarity(C1,C2)的正负值，若存在反义或对义则取-1

表 1 是实验中的相关参数，分为 3 个不同的参数组，通过不同参数的权重的不同设置，观察实验结果的合理性。

表 2　部分实验结果

词 1	词 2	刘群	知网在线	相似度(基于参数组 1)	相似度(基于参数组 2)	相似度(基于参数组 3)	相似度(参数动态调节)
男人	女人	0.86	0.87	-0.92	-0.86	-0.94	-0.94
男人	父亲	1.00	0.86	0.90	0.83	0.93	0.94
男人	经理	0.63	0.85	0.79	0.64	0.85	0.86
男人	高兴	0.05	0.00	0.00	0.00	0.00	0.00
合算	得不偿失	0.10	0.01	-0.49	-0.24	-0.56	-0.69
北部	北	0.44	0.57	0.59	0.45	0.62	0.68
卖方	采购员	0.74	0.87	-0.83	-0.71	-0.88	-0.88
鱼类	鲤鱼	1.00	0.95	1.00	1.00	1.00	1.00
色调	蓝色	0.05	0.00	0.00	0.00	0.00	0.00
医生	患者	0.57	0.31	-0.87	-0.77	-0.90	-0.89
医生	心理医生	—	0.98	0.97	0.94	0.97	0.97
患儿	儿科医生	—	0.30	-0.83	-0.70	-0.87	-0.84
盗墓人	盗墓	—	0.46	0.32	0.56	0.24	0.35
盗墓人	小偷	—	0.62	0.86	0.76	0.90	0.88
雨伞	打伞	—	0.10	0.31	0.55	0.24	0.29
雨伞	雨衣	0.33	0.03	0.44	0.46	0.34	0.39

实验表明主类义原及其框架对概念之间的类别区别贡献较大，如在参数组 1 和参数组 3 中，由于 β1 和 β3 的值设置的相对较高，在（“盗墓人”、“盗墓”）和（“盗墓人”，“小偷”）、（“雨伞”，“打伞”）和（“雨伞”，“雨衣”）这几组数据的计算中，能较好地体现类别区别。

“色调”的 DEF={Hue|浓淡:host={Color|颜色}}，“蓝色”的 DEF={blue|蓝}，其对应的主类义原的上位关系链分别为：属性→外观→浓淡，属性值→外观值→颜色值→蓝，即看

似关系密切的词语，在知网的概念架构中分别属于“属性”和“属性值”两个不同的类别，因此相似度为 0。从能否替换的角度看，它们确实可替换性较差，但是它们同时存在一定的语义关联，针对不同的应用目标，在相似度计算中，应考虑进属性和属性值之间的关系。

实验结果中的正负值，较好地反映出语义特征相似度和句法特征相似度，如（“合算”，“得不偿失”），语义相反，但句法结构中具有替代性。“鱼类”和“鲤鱼”的相似度为 1，是因为它们的 DEF 都是{fish|鱼}。Hownet 中某些词语定义的细致程度还有待进一步的提高。

由于知网中，概念的语义是从概念特性描述、主类义原、主类义原框架 3 个方面进行定义的，具体到某些词语，在这 3 个方面的描述分量并不是很平衡，有些特性描述较细致，但义原或义原框架却较简略，而有些却正好相反。如“北”、“北部”，它们的主类义原框架描述较细致，且相似性高，因此在参数组 3 等到结果最大。固定地设定某组参数，对某些词效果叫好，可能对另外一些词，则不然。针对这一情况，进行参数的动态设定，即综合考虑概念定义的 3 个方面，动态调节参数。参数计算公式（6）所示：

$$\begin{cases} \beta_1 = \log(\sum_{i=1}^{n1} 1 + \sum_{j=1}^{m1} 1) \Big/ \Sigma_\beta \\ \beta_2 = \log(\sum_{i=1}^{n2} 1 + \sum_{j=1}^{m2} 1) \Big/ \Sigma_\beta \\ \beta_3 = \log(\sum_{i=1}^{n3} 1 + \sum_{j=1}^{m3} 1) \Big/ \Sigma_\beta \\ \Sigma_\beta = \log(\sum_{i=1}^{n1} 1 + \sum_{j=1}^{m1} 1) + \log(\sum_{i=1}^{n2} 1 + \sum_{j=1}^{m2} 1) + \log(\sum_{i=1}^{n3} 1 + \sum_{j=1}^{m3} 1) \end{cases} \quad (6)$$

其中，n1、m1 分别为参与比较的两个概念的主类义原在义原树上的深度，n2、m2 分别为两个概念的特性描述中的结点数目，n3、m3 分别为两个主类义原框架中的结点数目。

6. 结论与展望

本文分析和利用新版 Hownet 的概念架构和语义多维表达形式，从概念的主类义原、主类义原框架以及概念特性描述三个方面综合分析词语相似度，并从实际功用出发，将词语相似度细分为语义特征相似度和句法特征相似度，并在计算中体现出不同。实验结果较为理想，与人的直观判断基本一致。《知网》含有丰富的词汇语义知识，新版《知网》概念描述架构体系有了较大的改进和完善，为中文语义处理提供了很好的平台。在后继研究中，将着力于挖掘和利用 Hownet 中的动态角色与特征提供的概念之间更为细致的语义关联，分析词语相似性与相关性之间的内在联系和转换。

参 考 文 献

[1] 刘群,李素建. 基于《知网》的词汇语义相似度的计算[C] 第三届汉语词汇语义学研讨会. 台北,2002.

[2] 董振东, 董强. 知网, http://www.keenage.com

[3] 董振东,董强,郝长伶，知网的理论发现, 中文信息学报 2007 年 第 04 期

[4] Dekang Lin. An Information Theoretic Definition of Similarity Semantic distance in WordNet [C] Proceedings of the Fifteenth International Conference on Machine Learning. 1998.

[5] Eneko Agirre , German Rigau. A Proposal for Word Sense Disambiguation using Conceptual Distance[C] Proceedings of the First International Conference on Recent Advanced in NL P. 1995.

[6] BUDANITSKY, A. AND HIRST, G. Semantic distance in WordNet: An experimental, application oriented evaluation of five measures[C] Workshop on WordNet and Other Lexical Resources , Second meeting of the North American Chapter of the Association for Computational Linguistics. 2001.