# 基于WordNet的语义相似度计算方法

基于 WordNet 本体的语义相似度计算模型可分为三大类: 第一类是基于路径距离计算的方法，这类方法一般通过计算边的个数来计算两个概念在分类树中的语义距离，然后将语义距离转换成相似度值，该方法主要考虑的是连接两个概念的边的数量，边的个数越多表示距离越远，表示两个概念的相似度越小。第二类是基于IC （Information Content）的方法，一般通过计算两个概念共有信息量的多少来确定概念间的相似度。这些方法依据概念c 在语义词典中出现的频率给每一个概念关联一个概率p（c） ，然后将每个概念的概率的负对数似然值作为这个概念的信息量，即 IC （c） = -log p（c） 。第三类是基于属性特征的方法，该方法认为事物间的关联程度与它们所共有的属性数成正比。两个概念的公共属性数量越多，相似度越大。基于属性的语义相似度计算从词汇的特征出发来计算概念语义相似度，如利用WordNet中的同义关系，利用概念节点的祖先节点的交集和并集。在不作具体说明情况下，本文介绍的这 3 类算法都是建立在 WordNet“is\_a”关系树状分类体系基础上的。

## 1基于路径距离的语义相似度算法

基于路径距离的语义相似度算法以 WordNet is\_a 关系分类树为基础，通过两个概念在关系树中的最短距离来表示它们之间的语义相似性，该类方法认为距离越近的概念间语义相似程度越高，该类算法中典型的有**Rada**等提出的根据链接两个概念的路径中的最小长度路径来计算两个概念的相似度，也就是利用边的个数来计算。相应计算公式如式（15）所示：

(15)



后来**Wu&Palmer**提出了新的改进方法，主要思想是在别的因素不变的情况下，在分类树中层次越低的概念之间相似度越小。具体计算时除了考虑两个概念间的路径外，还考虑了两个概念的公共父节点及其在分类树中的相关深度，相应计算公式如式（16）所示：

(16)



其中，depth（ci） 表示概念 ci 在WordNet“is\_a”关系树中的深度。len（c1，c2）是指 WordNet 中两个概念（c1，c2）最短的路径距离。lso（c1，c2）表示 c1 和 c2 处于 WordNet“is\_a”树中最深层的公共父节点。

Leacock&Chodorow也提出了一种非线性的计算方法，该方法考虑了两个概念间的节点数量（包含本身）和所处的分类树的最大深度，相应计算公式如式（17）所示：

(17)



这种方法考虑了两个概念间的路径，对于一个固定的分类树来说，两个概念间的路径越多，语义相似度越小。从信息论的角度看，该算法求解得到的语义相似度值是概念间路径所提供的信息量。 Hirst-St-Onge 法认为存在较短路径的概念词 c1 和 c2 ，在遍历路径过程中方向改变的次数越少，则两个概念词相关性越强。具体计算公式如式（18）所示：

(18)



其中，C 和 k 为常量，L 代表 c1 和 c2 间最短路径长度 ，d 表 示遍历路径中方向改变的次数 。虽然Hirst-St-Onge的方法考虑了路径“方向”问题，但实验效果不佳。

通过以上典型方法可以看出，基于距离的方法的优点主要是只依靠概念所处分类树的几何模型，相对于文本语料库方法来说算法简单，计算量较小。不足之处在于因为路径是一个离散量，如果单纯基于路径去求相似度势必很难改善相似度算法的性能，精确性较差，例如在计算时本体中别的路径都被忽略，只考虑最小路径，这样求出的语义相似度值准确度不好。

## 2基于IC的语义相似度算法

基于IC的语义相似度算法的核心是算法中IC值的计算，因此研究人员在基于 IC信息容量的语义相似度算法研究主要集中与两个方向，一是语义相似度计算模型，二 是IC值计算模型。例如，Resnik，Jiang-Conrath，Lin等人都提出了各自的语义相似度算法模型。Seco，Zhou，Meng，Sánchez，Sebti等人提出了自己的 IC 计算模型。

在提出语义相似度计算模型的人中，Resnik是第一个引入本体和词典来计算相似度的人。他从语义判断的目的出发，提出在在分类树中，概念出现的频率可以用Brown语料库的名词频率来估计。Resnik根据该理论判断一对概念的相似度就应该是这对概念共享信息的数量，因此他将两个概念 c1 和 c2 的共同祖先的最大信息量作为两个概念的语义相似度。其提出的语义相似度算法模型如式（19）所示：



（19）

其中，lso（c1，c2） 指概念 c1 和 c2 位于该分类树中的最小公共父节点。

Jiang-Conrath 的语义相似度算法是在 Resnik 算法的基础上提出的，与Resnik相同的是Jiang-Conrath也利用了信息容量的概念，只不过作者通过结合一些词对的最大词义性来计算相似性。

(20)



这里 s1i和s2j 是 w1和w2 的意义（在本体中的概念）。他们利用两个概念的信息量和它们的最大信息量之差作为语义距离进行语义相似度的计算，这种方法被认为是计算两个概念的相异度。其算法模型如下：

(21)



Lin对语义相似度的理解与Resnik 类似，他认为一方面概念有共性的一面，另一方面每一个概念也有自己单独的信息容量。因此他提出的的语义相似度算法都与特定的应用、特定的领域有关。后来，Lin考虑定义一个通用的计算相似度的方法，他认为两个概念的相似性应该能被信息共性和信息总量的比值来测算，其提出的算法模型如下：

(22)



这里 hypo（c） 返回值是概念 c 的所有子节点数， max \_nodes代表的是该分类树本身所包含的概念的最大数量。该方法中默认WordNet是基础分类树，但该模型存在的问题是两个处于不同层次但下义词数量相同的概念，它们的相似度计算结果也相同。

为了避免这种情况，David Sanchez等引入了一种新的基于自身结构的模型来计算 IC的值，这种方法没有考虑深度问题，而是引入了下义词，通过一个概念的下义词和上义词来进行计算。他们认为依靠叶子来区分不同概念就已经足够了，David新提出的IC计算公式如下:



(23)

式中 leaves（a） = m ∈ C/m ∈ hyponyms（a） ，这里 C 是本体概念集合，m 是其中的叶子节点。 subsumers（a） = {c ∈ C / a ≤ c}∪{a}，这里 a ≤ c 是指 a 为 c 的直系下义词节点。所有的叶子概念不论深度都拥有相同的最大信息容量 -log （1/（max \_leaves + 1）） 。该模型存在的问题是有相同上义词概念但有不同下义词结构和下义词数会导致IC值相等。也就意味着这两个概念意义相同但包含信息不同。

为了克服这种情况，Zhou等引入了概念的下义词的相对深度，提出了一种新的计算IC值的方法：

(24)



引入深度作为权重参数虽然克服了 David 算法的问题，但又导致了新的问题的产生，例如该参数必须通过实验调试来确定具体的 K 值。

为此Meng等整合了Seco and Zhou的方法来克服这种情况，他考虑了每个概念的深度和该概念的每个下义词的深度，并改变了术语 hypo（c）而通过另一个术语max \_dep th 来体现对IC计算的贡献:



(25)

对概念 c，dep th（c） 是概念 c 在分类树中的深度， max \_dep th 是分类树的最高深度，max \_nodes 是分类树中的概念的最大数量。

David Sanchez等后来针对自己文章存在的不足，又提出了一种新的利用概念共性（commonness）来计算 IC值的模型，这种模型依靠叶子节点的 subsumer（上义词）的数量来计算IC的值。具体计算公式如下所示:

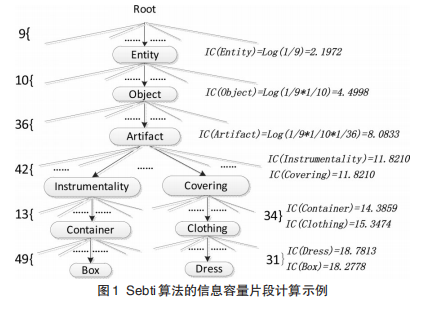
(26)



这里 commonness（ci）= ∑commonness（n）， n 为叶子节点并且 n 是概念 ci 的 subsumer（上义词），但 ci 不是叶子节点，commonness（ci） = 1/subsumers（n）。

Sebti等也提出了一种包含了深度的以下义词数量为基础的IC计算新方法来。具体计算方法如图1所示。

以Box的IC值计算为例：



(27)



通过以上分析可以看出，这些方法都是以不同形式利用了本体中的节点和结构，例如 hypo（c），leaves（c），subsumers（c）和depth（c），其中Resnik 的算法、Jiang-Conrath的算法与Lin 算法都是直接或间接的将两个概念父节点的 IC 值作为两个概念的相似度，Seco，Zhou and Sánchez等人都在一定程度上利用层次结构来计算本体中的IC，都有一定的代表性。

## 3基于属性特征的语义相似度计算方法

基于属性的语义相似度计算模型的方法依据的思想是事物之间的关联程度与其所具有的公共属性数相关，两个概念的公共属性数量越多，相似度越大。因此，这类方法的基本原理是通过判断两个概念对应的属性集的相似程度来区分概念的语义相似度。

Tversky算法模型是该类算法的典型，其语义相似度计算模型如下:

(28)



其中，θ ，α ，β ，f （c1 ∩ c2） 返回的是 c1 和 c2 的公共属性项的数量，f （c1 - c2） 返回的是 c1 有但 c2 中没有的属性数，f （c2 - c1） 返回的是 c2 有但 c1 中没有的属性数，参数分别表示 c1 和 c2 的公共属性和非公共属性对其相似度计算的影响程度，参数值的确定由具体任务决定，且由于概念词相似度的非对称性，α 和 β 值不一定相同。基于属性的语义相似度计算模型在国内典型的有荀恩东等人从WordNet中提取同义词并采取向量空间方法计算的语义相似度模型。该计算模型首先从WordNet的同义词词集（Synset）、类属信息（Class）和意义解释（Sense Explanation）三方面抽取候选同义词，进行特征提取后计算被比较概念词的各个意义（Sense），然后在这三个特征空间通过计算距离来计算各自的意义相似度，最后基于意义相似度来综合计算词语的相似度。