第35卷 第14期 计 算 机 工 程 2009年7月
Vol.35 No.14 Computer Engineering July 2009
·软件技术与数据库· 文章编号：1000—3428(2009)14—0069—03 文献标识码：A 中图分类号：TP311

基于潜语义分析的概念名称相似度算法

黄广君，孙建国，罗俊丽
(河南科技大学电子信息工程学院，洛阳 471003)

摘 要：概念名称是本体映射中的一个重要要素。针对目前概念名称相似度计算中存在的概念名称多义性问题，提出一种改进的算法。该算法结合概念注释和义项解释，利用潜语义分析，明确概念在Wordnet中对应的义项，在一定程度上提高了相似度计算的准确度。实验表明该方法是切实可行的。

关键词：相似度；潜语义分析；义项

Concept Name Similarity Algorithm Based on Latent Semantic Analysis

HUANG Guang-jun, SUN Jian-guo, LUO Jun-li
(College of Electronic Information Engineering, Henan University of Science and Technology, Luoyang 471003)

【Abstract】 Concept name is important in the process of ontology mapping. This paper proposes a modified concept name similarity algorithm against polysemy of concept name in the process of concept name similarity. Combined with comment and gloss of synset by Latent Semantic Analysis(LSA), it determines the corresponding synset of concept in the Wordnet. So it improves the accuracy of similarity to a certain extent. Experimental result shows that the algorithm is viable.

【Key words】 similarity; Latent Semantic Analysis(LSA); synset

1 概述

在本体映射过程中，概念名称是用于评价概念相似度的一个重要信息。目前概念名称相似度计算主要是从语义角度出发，通过计算2个概念在词典中的语义距离[1]，进而得到语义相似度。基于语义的方法依赖概念含义的准确表达，而在自然语言中存在着大量的多义词现象，一个词语通常具有多个含义。如果脱离一定的语境，离开上下文对词义的限定，仅仅通过概念名称很难确定概念的含义。例如，bank如果离开取钱这个语境单独出现，很难确定它是指银行还是河堤。如果概念含义不明确，得到的语义距离是不准确的。由此可见，自然语言中的一词多义性较大地干扰了基于语义方法的计算效果。

基于语义的相似度计算方法主要分为2类[2]：一类基于概念间的路径长度，这类方法有Leacock-Chodorow, Hirst-St-Onge, path；另一类基于最近上层概念的负平均信息量(information content)，包括Resnik, Jiang-Conrath, Lin 3种方法。这6种方法都提到了对多义词的处理。Hirst-St-Onge将2个多义词之间的多条路径根据不同的连接关系进行分类，不同连接关系的权值不同。由权值最大的路径计算得到这2个词的相似度。

对于其他5种关系，均是先按照自己的方法计算2个多义词任意2个义项间的相似度，将其中的最大值作为最终的相似度。这6种方法的实质是计算语义上最相关的2个义项的相似度，但是本体概念是由多方面因素决定的。脱离了本体这个环境，仅依赖于词典中的语义得到的义项未必能代表原概念，从而计算的相似度也是不准确的。

针对上述问题，本文提出了一种改进的算法——基于潜语义分析的概念名称相似度算法。该方法综合概念名称和概念注释(comment)两方面因素，利用潜语义分析，结合概念注释和义项解释明确概念的含义，确定概念名称在语义词典Wordnet中的义项，然后通过计算概念名称对应义项的语义距离得到概念名称相似度。

2 理论基础

(1)Wordnet语义词典

Wordnet是美国Princeton大学认知科学实验室开发的一种通用的语义词典[3]。它通过词语的语义属性组织词典，其中基本的构建单位是同义词集合，也称为义项(synset)。每个义项通过一段简短的解释明确含义，代表一个意义明确唯一的概念。义项之间通过指针进行连接，指针的类型决定了它们之间的不同关系。典型的关系包括上下位关系、整体部分关系、继承关系等，同义关系蕴涵在义项中。

在Wordnet的多种关系中，上下位关系为概念提供了很好的层次结构，处于更高层次的概念代表更抽象的意义；反之，处于较低层次的概念代表的意义就较具体。通常利用Wordnet计算名称相似度使用的是概念之间的上下位关系。2个概念在Wordnet层次结构上越接近，它们的相似度就越高，反之越低。

在Wordnet中查找某个概念时，词典并不知道概念的含义，它只是按照这个概念的词形在词典数据库中检索，返回

基金项目：教育部科学技术基金资助重点项目(03081)
作者简介：黄广君(1963－)，男，副教授、博士，主研方向：语义Web，Web Service；孙建国、罗俊丽，硕士研究生
收稿日期：2008-12-23 **E-mail：**ly_sjg@126.com

和这个概念拼写相匹配的所有义项。对于一个概念来说，它的含义是明确的。但是这个概念对应的词形往往具有多个含义。因此，Wordnet 返回的查询结果通常是多个义项。Wordnet 本身不能判断概念对应的义项，概念义项的确定依赖于用户或者概念所处的上下文。

(2)潜语义分析

潜语义分析(Latent Semantic Analysis, LSA)是一种用于知识获取和展示的计算理论和方法[4]。其基本思想是文档中的词与词之间存在某种联系，即存在某种潜在的语义结构。同义词之间具有基本相同的语义结构，多义词的使用具有多种不同的语义结构。词汇之间的这种语义结构与其在文档中的出现频率有关。因此，可以通过统计方法提取并量化这些潜在的语义结构，进而消除同义词、多义词的影响，提高文档表示的准确性。LSA 方法在信息检索、答案提取、自动判分等方面都取得了很好的应用效果。

潜语义分析的数学基础是奇异值分解(SVD)。其实现过程如下：首先根据预处理后的文档，构造一个 $m\times n$ 的词-文档矩阵：

$$\boldsymbol{A}=(f_{ij})_{m\times n} \tag{1}$$

其中，f_{ij} 表示第 i 个词在在第 j 个文档中的频率。对矩阵 $\boldsymbol{A}$ 进行奇异值分解：

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{2}$$

其中，$\boldsymbol{U}$ 是 $\boldsymbol{A}$ 的左奇异值向量，是一个 $m\times m$ 的正交矩阵，表示词汇的语义空间；$\boldsymbol{V}$ 是 $\boldsymbol{A}$ 的右奇异值向量，是一个 $n\times n$ 的正交矩阵，决定了文档在语义空间的位置；$\boldsymbol{\Sigma}$ 是对角矩阵，对角元素是矩阵 $\boldsymbol{A}$ 的奇异值。它们按照从大到小的顺序排列。为了简化计算，取前 k 个奇异值，对矩阵 $\boldsymbol{A}$ 降维处理，得到矩阵 $\boldsymbol{A}$ 的近似矩阵 $\boldsymbol{A}_k$ ：

$$\boldsymbol{A}_k=\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}} \tag{3}$$

经过变换的 $\boldsymbol{A}_k$ 不仅保持了原来的词和文档之间的语义关系，而且去除了大量因同义或多义产生的“噪声”。变换后的 $\boldsymbol{U}_k$ 仍表示词汇的语义空间，$\boldsymbol{V}_k$ 表示文档的语义空间。将检索向量 $\boldsymbol{Q}$ 向该语义空间投影，得到它在语义空间的坐标向量 $\boldsymbol{Q}'$ ：

$$\boldsymbol{Q}'=\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{U}_k\boldsymbol{\Sigma}_k^{-1} \tag{4}$$

利用式(5)求它与文本向量 $\boldsymbol{U}_j=(u_{1j},u_{2j},\cdots,u_{kj})^{\mathrm{T}}$ 的夹角余弦。

$$C_j=\sum_{i=1}^{k}(q_i'u_{ij})\Big/\Big(\sqrt{\sum_{i=1}^{k}(q_i')^2}\sqrt{\sum_{i=1}^{k}(u_{ij})^2}\Big) \tag{5}$$

该值表示两者之间的相似度。求这些值中的最大值，与之对应的文档就是需要检索的文档。

本文将 LSA 理论应用于概念语义的相似判断上，结合本体概念的注释和义项的解释，确定概念在 Wordnet 中对应的义项。通过统计方法构造一个潜在的特征词-义项语义空间，消除了词之间的相关性，能够有效地解决词匹配方式所面临的同义和多义现象，提高概念含义判断的准确度。

3 方法设计

基于潜语义分析的概念名称相似度算法是在传统相似度计算方法的基础上，通过引入概念注释这个上下文信息，利用潜语义分析明确概念在 Wordnet 中对应的义项，使用基于路径长度的方法求得相应义项的相似度，进而得到概念名称相似度。

该算法包括 3 个模块：概念预处理模块，概念语义确定模块，相似度计算模块。具体算法流程如图 1 所示。

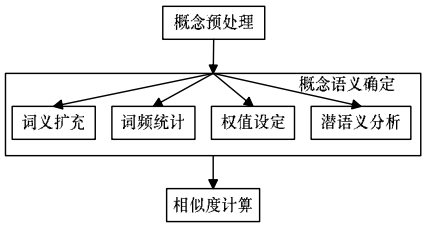

图 1　基于潜语义分析的概念名称相似度算法流程

(1)概念预处理模块

利用词典计算相似度使用的是语义信息。因此，在该模块中，首先在 Wordnet 中查找概念名称对应的义项。词典返回表示该义项的同义词集合和义项解释。通常义项解释是一个短语句，本体概念的注释也是一个短语句，而潜语义分析是通过对多个特征词在文档中的词频进行处理的。因此，需要提取相关语句的特征词。在英文语句中单词通过空格自然分开，不需要对分词进行特殊处理，利用空格可以提取不同的词。在这多个词中，部分辅助性词对于语句的意义表达贡献不大，如 is, a, this, there。将这些词作为停用词从提取的词中去除。经过以上处理，概念注释及每个义项解释都得到了相应的特征词组。

(2)概念语义确定模块

在以往基于语义的计算方法中，基本都是通过寻找词的多个义项间的最相近义项而得到相似度。但是由于缺少一定的语境，很难确定概念的含义，所找到的义项可能与概念不对应，因此通过这种方法计算的相似度在某些情况下是不准确的。本文引入本体定义中概念名称的注释，在一定的上下文中明确了概念的含义。另外，在 Wordnet 中，利用义项的解释可以确定该义项代表的含义，弥补了在义项中几个同义词不能准确表达词义的不足。将这 2 种信息相结合，可以确定概念在 Wordnet 中对应的义项。

(3)词义扩充

概念语义确定模块主要使用的是潜语义分析。而在潜语义分析过程中，抽取特征词的数目对于结果的正确性影响很大。特征词太少，不能准确表达文档的含义。特征词太多，无疑会增加算法的复杂度。在 Wordnet 中，义项的解释一般在 7 个词左右，显然不利于潜语义分析。为了提高潜语义分析的准确性，需要扩充特征词。本文将义项、该义项的上位词、下位词的义项、这些义项的解释经过概念预处理抽取的特征词并入原义项的特征词。义项 i 经过扩充后的特征词具体形式如下：

$$kw_i=\left\{\begin{matrix} w_{ij}\mid w_{ij}\in synset_i,gloss_i,hype_i, \\ hypegloss_i,hypo_i,hypogloss_i \end{matrix}\right\} \tag{6}$$

其中，$synset_i$ 代表义项 i 的同义词集合；$gloss_i$ 代表义项解释；$hype_i$ 代表该义项的上位义项的同义词集合；$hypegloss_i$ 代表上位义项解释；$hypo_i$ 代表下位义项的同义词集合；$hypogloss_i$ 代表下位义项解释；$1\le i<m$，m 表示概念在 Wordnet 中义项的个数；$1\le j<n$，n 是义项 i 相对应的特征词的总个数；经过以上的扩充，每个义项对应的特征词数目不仅有所增加，而且对义项的表达更加完整。

通过上述扩充，将概念的每个义项对应的特征词合并，得到概念的特征词向量 $\boldsymbol{KW}$。

$$\begin{aligned}\boldsymbol{KW}=\{&w_i\mid w_i\in kw_j,w_i\ne w_k,i\ne k\\ &w_i\in synset_j\ or\ hype_j\ or\ hypo_j\\ &w_k\notin synset_j\ or\ hype_j\ or\ hypo_j\}\end{aligned} \tag{7}$$

其中，特征词向量中的元素 w_i 互不相等。因为义项中的同义词集合表达的是同一个含义，而且潜语义分析在同义词处理上开销很大，所以把同义词中的多个词作为一个词来处理。相当于把义项中的所有同义词看作一个整体，作为特征词-义项矩阵中的一列，而不是把它的每个特征词单独作为一列。在这里选择同义词集合中的第 1 个词代表该同义词集合。

(4)词频统计

根据得到的特征词向量，对每个义项依据扩展项的类别分别进行特征词词频统计，并记录不同特征词在义项同义词集合、义项解释、上位义项同义词集合、上位义项解释、下位义项同义词集合和下位义项解释中的词频 $fsysset_j(w_i)$，$fgloss_j(w_i)$，$fhype_j(w_i)$，$fhypegloss_j(w_i)$，$fhypo_j(w_i)$，$fhypogloss_j(w_i)$。由于本文把同义词集合中的词作为一个整体对待，因此统计的词如果和这个同义词集合中的词相同，那么与同义词集合对应的词频将增 1。如果统计的是同义词集合本身，这时的词频只能记录 1 次。

(5)权值设定

在语义词典 Wordnet 中，不同的义项扩展项与义项关系远近不同，所以，它对义项含义表达的贡献也不同。根据扩展项的类别，本文设置了不同的权值。义项和义项解释中的特征词权值一致，而且最高，下位义项和下位义项解释中的特征词权值最低。将这个权值和上述统计的词频相结合得到特征词在概念某个义项中的词频。具体计算如下：

$$\begin{aligned} f_{ij} = {} & w_{synset} \cdot fsysset_j(w_i) + w_{gloss} \cdot fgloss_j(w_i) + \\ & w_{hype} \cdot fhype_j(w_i) + w_{hypegloss} \cdot fhypegloss_j(w_i) + \\ & w_{hypo} \cdot fhypo_j(w_i) + w_{hypogloss} \cdot fhypogloss_j(w_i) \end{aligned} \tag{8}$$

其中，$w_{synset} = w_{gloss} > w_{hype} = w_{hypegloss} > w_{hypo} = w_{hypogloss}$。

(6)潜语义分析

经过以上处理，最终形成特征词-义项矩阵。对该矩阵依照式(3)计算，得到与概念对应的潜语义空间。将概念注释按照提取的特征词进行词频统计，利用式(4)，向该潜语义空间投影变换得到概念注释向量。通过式(5)计算每个义项向量和概念注释向量的夹角余弦值，该余弦值表示义项和概念注释的相关度。从这些值中选择最大值，与之对应的义项作为概念在 Wordnet 中的确定义项。

(7)相似度计算模块

由于本文重点在概念语义的确定上，因此简化了相似度计算部分，仅考虑 2 个义项之间的最短距离，而忽略了义项在语义树中深度、密度对相似度计算的影响。具体相似度计算如下：

$$\mathrm{sim}(c1,c2) = 1/(1+d) \tag{9}$$

其中，d 表示语义距离，它是义项 $c1$,$c2$ 在 Wordnet 语义树中的最短距离。

4 实验

(1)实验数据

测试数据选自 OAEI 数据集，该数据集用于 2007 年国际本体映射大赛。它主要描述了书籍信息，包括 54 组本体数据。编号 101 的数据是它的参考本体，也是最完整的本体。其中包括 33 个概念、40 个属性、24 个关系。在这里同样将 101 作为参考本体。本文相似度计算以名称相似度为主，所以，选取了编号 205 的本体数据作为映射本体，该组数据体现了名称之间的近义关系。另外该组数据包括概念的注释信息，便于使用潜语义分析确定概念在 Wordnet 中的含义。

因为概念名称相似度是本文讨论的重点，所以只计算概念名称相似度，不再处理属性、关系相似度。观察 101 组，205 组数据不难发现，在 33 个概念中有 12 个概念是合成词，在 Wordnet 中查不到这些合成词。因此，对这些合成词进行了分词处理，选取其中的某个词替代整个词，它不仅能够在 Wordnet 查询到，而且能表达原词的基本含义。例如，ReferenceGuide 经过处理选取 Guide 代表原词。另外，还需要将名词的复数形式转换为该词的单数形式，方便在 Wordnet 中查询，例如 Directions 转换为 Direction。

(2)评估方法

在本体映射中，普遍采用信息检索中的查准率、查全率作为评估映射的重要指标。查准率表示正确发现的相似概念占发现的所有概念的比率，查全率是指正确发现的相似概念占所有相似概念的比率。本文同样使用查准率、查全率对相似度计算的实验结果进行评价。

(3)实验设计

为了验证基于潜语义分析的概念名称相似度算法，本文设计了 2 个算法。算法 1 使用了传统的基于语义距离的算法。该方法的基本思路如下：假设计算概念 $c1$, $c2$ 的相似度。通过 Wordnet 查找得到，$c1$ 有 $w1$ 个义项，$c2$ 有 $w2$ 个义项。利用式(9)分别计算 $c1$, $c2$ 任意 2 个义项的相似度，把这多个相似度中的最大值作为概念 $c1$, $c2$ 的相似度。算法 2 是本文的方法，它首先通过 Wordnet 查找得到 $c1$, $c2$ 的多个义项，利用潜语义分析，结合概念的注释，确定概念 $c1$, $c2$ 对应的义项，使用式(9)计算相应义项的相似度，所得的义项的相似度就是概念 $c1$, $c2$ 的相似度。本文在上述数据集上分别进行了这 2 种算法的实验。另外，在潜语义分析中，k 值的选择是个重要问题。k 值选得太大，达不到降维和简化运算的目的；选得太小，保留下来的重要语义结构太少，分辨义项的能力不足。因此，在算法 2 的实验中，设计了 4 个不同的 k 值来说明在该类问题中语义空间维数的选择。

5 结果及分析

实验结果见图 2。从基于潜语义分析的实验可以看出，k=1 时效果很差，当 k=2 时，查准率和查全率都很高，而在 k>2 时，效果却逐渐变弱。这说明由 k=1 计算所得语义空间基本不能表示原空间。而在 k=2 时，已经能够分辨出概念在 Wordnet 中的义项。但是随着 k 值的增大，潜语义空间逐渐逼近原词汇-文档空间，语义空间中的多义噪声对于义项识别的干扰不断增强。由此可知，二维语义空间在短句含义的表达上具有很好的效果。

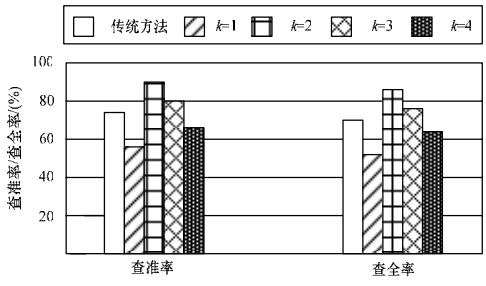

图 2　实验结果比较

与算法 1 的实验结果对比可以看出，当 k=2 时，本文方法在查准率、查全率 2 个方面都明显优于传统方法。传统方法由于没有使用上下文信息，很难确定概念名称的确切含义，

(下转第 74 页)