文章编号:1001 - 9081(2010)06 - 1668 - 03

计算术语间语义相似度的混合方法

魏 韡[1,2],向 阳[1],陈 千[1]
(1. 同济大学 电子与信息工程学院,上海 201804;
2. 井冈山大学 信息科学与传媒学院,江西 吉安 343009)
(weiweihzkd@163.com)

摘 要:提出一种基于有向无环图和内在信息量的计算语义相似度的方法。首先计算出两个术语基于所在有向无环图的子图,再分别计算两个子图的交集和并集。用内在信息量方法计算出两个子图的交集和并集包含的节点的内在信息量,再计算出交集的节点内在信息量之和以及并集的节点内在信息量之和,将两者的比值作为两个术语的语义相似度。实验结果表明,该方法具有较高的准确度。

关键词:语义相似度;内在信息量;有向无环图
中图分类号: TP391 **文献标志码**:A

Combined measurement approach for semantic similarity of terms

WEI Wei[1,2], XIANG Yang[1], CHEN Qian[1]
(1. *School of Electronics and Information, Tongji University, Shanghai* 201804, *China*;
2. *School of Information Science and Communication, Jinggangshan University, Ji'an Jiangxi* 343009, *China*)

Abstract: Measuring semantic similarities of terms is a key issue in many research fields. This paper proposed a method based on the Directed Acyclic Graphs (DAG) of terms and the intrinsic information content of terms to measure the semantic similarities of terms. It first calculated the sub-graphs of two terms based on the directed acyclic graph, and then calculated the intersection and union of the sub-graphs. The semantic similarity of two terms is the ratio of the total intrinsic information content of terms in the intersection to the total intrinsic information content of terms in the union. The experimental results show that the method has a higher degree of accuracy.

Key words: semantic similarity; intrinsic information content; Directed Acyclic Graph (DAG)

0 引言

语义相似度在心理学、语言学、认知科学以及人工智能等学科中都有广泛的应用。各种不同语义相似度方法的实现为语义检索、语义服务发现、语义歧义消解等研究提供了基础条件。基于本体的术语间的语义相似度是指某个本体领域内两个术语在该本体中的相似程度。文献[1-3]主要利用术语包含的信息量来计算语义相似度,但是忽略了术语所在本体的结构。文献[4-5]利用术语间的距离来衡量语义相似度,但术语分布的不均匀会影响结果的准确性。近年来,文献[6-7]提出的方法由于综合考虑多种因素,使语义相似度的计算准确度有进一步地提高。WordNet 是目前被广泛使用的基于认知语言学的英语词典,也可以看成是有语义联系的词汇本体。本文提出了一种综合了有向无环图结构和内在信息量的计算语义相似度的方法,并且在 WordNet 上的实验取得了不错的效果。

1 相关的计算方法

目前计算语义相似度的方法主要可以分为 3 种:基于信息量的方法、基于距离的方法和混合方法。

1.1 基于信息量的方法

基于信息量的方法是利用术语共享的信息量多少来衡量它们之间的语义相似度。如果两个术语共享的信息量越大,则表明它们之间的语义相似度越大。根据信息量理论,术语本身的信息量用公式表示为:

$$IC(c) = -\log(p(c)) \tag{1}$$

其中:$p(c)$ 代表术语 c 在给定的语料库中出现的概率。Resnik[1] 提出把两个术语的共同祖先中信息量的最大值来表示两个术语的语义相似度。Resnik 方法用公式可表示为:

$$sim(c_1,c_2) = \max_{c\in S(c_1,c_2)}(IC(C)) \tag{2}$$

其中 $S(c_1,c_2)$ 表示术语 c_1 和 c_2 的公共祖先集合。Resnik 方法只考虑了术语公共祖先的信息量,若术语对的公共祖先相同,用 Resnik 方法计算任何子层的术语对的语义相似性就相同,这是不准确的。Lin[2] 提出不仅考虑两个术语的公共祖先的信息量,而且考虑两个术语本身所含的信息量。Lin 方法用公式可表示为:

$$sim(c_1,c_2) = 2\times\frac{\max_{c\in S(c_1,c_2)}(IC(C))}{IC(C_1)+IC(C_2)} \tag{3}$$

Jiang 等人[3] 提出的基于语义距离的方法,将两个术语的信息量的差作为术语间的语义距离。根据语义距离越短,术语的语义相似度越大的原则。Jiang 方法用公式可表示为:

收稿日期:2010 - 01 - 04;**修回日期**:2010 - 03 - 11。 **基金项目**:国家自然科学基金资助项目(70771077);国家 863 计划项目(2008AA04Z106);上海市科委制造业信息化专项基金资助项目(08DZ1122303)。

作者简介:魏韡(1983 -),男,江西吉安人,讲师,博士研究生,主要研究方向:语义网、信息检索; 向阳(1962 -),男,重庆人,教授,博士生导师,博士,主要研究方向:决策支持系统、人工智能; 陈千(1983 -),男,湖北蕲春人,博士研究生,主要研究方向:语义网、本体论、数据挖掘、图像视频检索。

$$sim(c_1,c_2) = 1 - \frac{IC(C_1) + IC(C_2) - 2 \times \max\limits_{c \in S(c_1,c_2)}(IC(C))}{2} \quad (4)$$

1.2　基于距离的方法

基于距离的方法是根据术语在本体中的位置来确定术语之间的语义相似度。两个术语在本体中的距离越大,则其相似度越小。Norman 等人[4]提出的方法可表示为:

$$sim(c_1,c_2) = \frac{d}{d + dis(c_1,c_2)} \quad (5)$$

其中:$dis(c_1,c_2)$ 代表术语 c_1 和 c_2 之间的距离,d 是可调节的参数。这种方法比较直观,容易理解,计算也方便。这种方法假设节点的分布是均匀的,但是在本体中术语对应的节点具有层次性,一般在本体中上层的节点之间的差异比下层节点之间差异要大,所以在计算术语的语义相似度时要考虑术语对应节点所在层次的权重。Zhong 等人[5]提出的方法在上述方法的基础上做了改进,考虑到节点分布的层次的影响。在 Zhong 方法中,每个节点的层次所对应的权重都有一个值来表示,记为 $milestone$。$milestone$ 可以用式(6)来计算:

$$milestone(n) = \frac{1}{2k^{l(n)}} \quad (6)$$

其中:k 是大于 1 的预定义的参数,表示 $milestone$ 的值沿层次下降的速率(一般 k 取值为 2);$l(n)$ 为节点在分层结构中的深度(一般取节点到根的最长路径的长度)。对于根节点 $root$,$l(root) = 0$。对于层次结构中任意两个节点,它们都有一个最近的共同祖先。两个节点之间的距离由它们的值和它们共同祖先的值来决定。

$$\begin{cases} d_c(c_1,c_2) = d_c(c_1,ccp) + d_c(c_2,ccp) \\ d_c(c,ccp) = milestone(ccp) - milestone(c) \end{cases} \quad (7)$$

其中:ccp 是 c_1 和 c_2 最近的共同祖先,$d_c(c_1,c_2)$ 表示 c_1 和 c_2 之间的距离。计算出 $d_c(c_1,c_2)$ 以后,c_1 和 c_2 之间的语义相似度用式(8)表示:

$$sim(c_1,c_2) = 1 - d_c(c_1,c_2) \quad (8)$$

1.3　混合方法

混合方法综合考虑了各种因素,比基于信息量的方法和基于距离的方法提高了准确度。Li 等人[6]提出了一个组合了不同影响语义相似度因素的非线性方法。该方法可以用式(9)表示:

$$sim(c_1,c_2) = e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \quad (9)$$

其中:l 代表术语 c_1 和 c_2 之间的最短路径的长度,h 代表 c_1 和 c_2 的最近公共祖先在本体结构的高度,参数 α 和 β 分别代表最短路径的长度 l 和最近公共祖先高度 h 的影响值。OSS 方法[7]通过 3 个步骤来计算术语间的相似度。首先计算两个术语的 a-priori 分数,a-priori 分数是指针对某个特定文档的术语偏好值,可以通过分析本体结构计算出来;接着计算有多少 a-priori 分数在两个术语之间转移;最后将转移的 a-priori 分数用来表示为术语间的距离,得到语义相似度的值。OSS 方法可以用式(10)表示:

$$sim(c_1,c_2) = 1 - \frac{\log(T(c_1,c_2))}{max_D} \quad (10)$$

其中:$T(c_1,c_2)$ 表示术语 c_1 和 c_2 之间转移的 a-priori 分数,max_D 表示术语 c_1 和 c_2 之间最大的距离值。

2　基于有向无环图和内在信息量的混合方法

本文提出了一个基于有向无环图和内在信息量的混合方法。首先给出几个相关的定义。

定义 1　有向无环图。有向无环图记为 $G = (V,E)$,其中 V 表示图中节点的集合,E 表示链接图中节点的边的集合。这里本文定义的图中边是有方向的,且任意一个节点不能经过边到达该节点自身,故称为有向无环图。

定义 2　路径。$G = \langle V,E \rangle$ 是一个有向无环图,设节点 v_a 和 v_b 之间路径 $P = (v_0,v_1,\cdots,v_n)$,其中 $v_0 = v_a, v_n = v_b, V_i$ 是 $V_{i+1}(0 \leqslant i \leqslant n-1)$ 的直接祖先,即 V_i 和 V_{i+1} 存在有向边连接。

定义 3　子图。针对节点 v_a 的子图 $G_a = \langle V_a,E_a \rangle$,其中 V_a 表示至少和 v_a 有一条路径的节点的集合,E_a 表示只链接 V_a 中节点的边的集合。

定义 4　子图的交。设针对节点 v_a 的子图 $G_a = \langle V_a,E_a \rangle$ 以及针对节点 v_b 的子图 $G_b = \langle V_b,E_b \rangle$,则子图 G_a 和 G_b 的交记为 $G_{a\cap b}$。$G_{a\cap b} = \langle V_{a\cap b},E_{a\cap b} \rangle$,其中 $V_{a\cap b}$ 表示 V_a 和 V_b 的交集,$E_{a\cap b}$ 表示链接 $V_{a\cap b}$ 的边。

定义 5　子图的并。设针对节点 v_a 的子图 $G_a = \langle V_a,E_a \rangle$ 以及针对节点 v_b 的子图 $G_b = \langle V_b,E_b \rangle$,则子图 G_a 和 G_b 的并记为 $G_{a\cup b}$。$G_{a\cup b} = \langle V_{a\cup b},E_{a\cup b} \rangle$,其中 $V_{a\cup b}$ 表示 V_a 和 V_b 的并集,$E_{a\cup b}$ 表示链接 $V_{a\cup b}$ 的边。

传统的基于信息量理论计算术语的信息量需要根据语料库来分析术语出现的概率。这种方法因为需要解析庞大的语料库使得时间花费比较大,而且计算的结果和语料库的规模和类型也相关。文献[8]提出了一种只根据术语所在的本体结果来计算术语的信息量,称为内在信息量,用式(11)表示为:

$$IC(c) = 1 - \frac{\log(hypo(c) + 1)}{\log(max_{wn})} \quad (11)$$

其中:$hypo(c)$ 表示术语 c 的子孙的个数,max_{wn} 表示术语所在的本体中所有术语的个数。

根据以上定义,设本体中所有术语构成一个有向无环图,其中任意两个术语对应的节点 v_a 和 v_b,采用有向无环图结构和内在信息量的混合方法计算术语间的语义相似度的步骤如下。

1) 分别计算出术语对应的节点 v_a 和 v_b 对应的子图 G_a 和 G_b。

2) 分别计算出子图 G_a 和 G_b 的交 $G_{a\cap b}$ 和子图 G_a 和 G_b 的并 $G_{a\cup b}$。

3) 根据式(11),计算出 $V_{a\cap b}$ 和 $V_{a\cup b}$ 中各节点的内在信息量。

4) 计算出 $V_{a\cap b}$ 中各节点的内在信息量之和:

$$IC(V_{a\cap b}) = \sum_{i \in a\cap b} IC(i) \quad (12)$$

5) 计算出 $V_{a\cup b}$ 中各节点的内在信息量之和:

$$IC(V_{a\cup b}) = \sum_{i \in a\cup b} IC(i) \quad (13)$$

6) 将 $V_{a\cap b}$ 中各节点的内在信息量之和 $IC(V_{a\cap b})$ 与 $V_{a\cup b}$

中各节点的内在信息量之和 $IC(V_{a\cup b})$ 的比值作为节点 v_a 和 v_b 代表的术语之间的语义相似度，用公式表示为：

$$sim(v_a,v_b) = \frac{IC(V_{a\cap b})}{IC(V_{a\cup b})} \tag{14}$$

根据图 1 所示，采用基于有向无环图和内在信息量的混合方法计算节点 f 和 g 代表的术语的语义相似度。图中所有节点的内在信息量已经标注。f 的子图和 g 的子图的交集包括 $\{a,b,c\}$，并集包括 $\{a,b,c,d,f,g\}$。交集中节点的内在信息量之和是 0.394，并集中节点的内在信息量之和是 2.866。最后得到节点 f 和 g 代表的术语的语义相似度是 0.137。

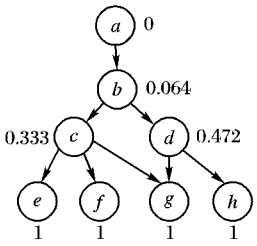

图 1 术语所在的有向无环图

3 实验与分析

评价语义相似度方法，目前被大多数研究者使用的是 Rubenstein and Goodenough 提出的 65 对单词对以及给出的相似度标准。Rubenstein and Goodenough 让 51 个测试对象根据语义的相似度对 65 对单词进行打分，分值为 0～4，语义相似度越高的单词对分值越高。本文将基于有向无环图和内在信息量的混合方法和其他方法分别计算 65 对单词对的语义相似度，和人工判断对比，通过式(15)得到 Pearson 相关系数：

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \tag{15}$$

其中：r 表示相关系数，x_i 表示计算方法得到的第 i 对单词的语义相似度，y_i 表示人工判断得到的第 i 对单词的语义相似度，n 表示单词对的数量。实验结果如表 1 所示。

实验结果表明，本文方法有比较高的准确度。

4 结语

本文提出的基于有向无环图和内在信息量的混合方法采用基于内在信息量的计算方法避免了分析语料库的问题，同时综合考虑了术语所在的有向无环图的结构，使得语义相似度的计算更符合人工的判断。该方法可以应用于和 WordNet 类似的本体的术语间语义相似度的计算。

表 1 各方法的相关系数

语义相似度计算方法	相关系数
Resnik 方法	0.846
Lin 方法	0.858
Jiang 方法	0.864
Zhong 方法	0.754
Li Yuhua 方法	0.908
OSS 方法	0.911
本文方法	0.914

参考文献：

[1] RESNIK P. Using information content to evaluate semantic similarity in a taxonomy [C]// Proceedings of the 14th International Joint Conference on Artificial Intelligence. California: Morgan Kaufmann Publishers, 1995: 448 – 453.

[2] LIN D. An information-theoretic definition of similarity [C]// Proceedings of 15th International Conference on Machine Learning. California: Morgan Kaufmann Publishers, 1998: 296 – 304.

[3] JIANG J, CONRATH D. Semantic similarity based on corpus statistics and lexical taxonomy [C]// Proceedings of International Conference on Research in Computational Linguistics. Washington, DC: IEEE, 1997: 19 – 33.

[4] NORMAN FOO, GARNER B, TSUI E, *et al.* Semantic distance in conceptual graphs [C]// Proceedings of Fourth Annual Workshop on Conceptual Structures. New York: Ellis Horwood, 1992: 149 – 154.

[5] ZHONG J W, ZHU H P, LI J M, *et al.* Conceptual graph matching for semantic search [C]// Proceedings of the 10th International Conference on Conceptual Structure. Berlin: Springer-Verlag, 2002: 92 – 106.

[6] LI Y, BANDAR A, MCLEAN D. An approach for measuring semantic similarity between words using multiple information sources [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4) 871 – 882.

[7] ZUBER V S, FALTINGS B. OSS: A semantic similarity function based on hierarchical ontologies [C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. California: Morgan Kaufmann Publishers, 2007: 551 – 556.

[8] SECO N, VEALE T, HAYES J. An intrinsic information content metric for semantic similarity in WordNet [C]// Proceedings of the 16th European Conference on Artificial Intelligence. Amsterdam: IOS Press, 2004: 1089 – 1090.

(上接第 1663 页)

[2] 张刚，刘挺，郑实福，等. 大规模网页快速去重算法[C]// 中国中文信息学学会二十周年学术会议论文集：续集. 北京：清华大学出版社，2001：18 – 25.

[3] 吴平博，陈群秀，马亮. 基于特征串的大规模中文网页快速去重算法研究[J]. 中文信息学报，2003，17(2)：28 – 35.

[4] 孔素然，黄萱菁. 基于模糊匹配思想的网页去重算法[D]. 上海：复旦大学，2006.

[5] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001，18(9)：23 – 26.

[6] 王哲. 基于特征码的网页去重算法研究[J]. 山东广播电视大学学报，2009(1)：14 – 16.

[7] 李卫，刘建毅，王枞. 基于全信息的网络文本信息去重算法研究[C]// 第十一届中国人工智能学术年会. 北京：北京邮电大学出版社，2005：1276 – 1281.

[8] 连浩，刘悦，许洪波，等. 改进的基于布尔模型的网页查重算法[J]. 计算机应用研究，2007，24(2)：36 – 39.

[9] 韩冰，林鸿飞. 大规模文本去重策略研究[D]. 大连：大连理工大学，2008.

[10] 王建勇，谢正茂，雷鸣，等. 近似镜像网页检测算法的研究与评价[J]. 电子学报，2000，28(Z1)：129 – 132.