第37卷 第10期
Vol.37 No.10

计 算 机 工 程
Computer Engineering

2011年5月
May 2011

·软件技术与数据库·

文章编号：1000—3428(2011)10—0047—03 文献标识码：A 中图分类号：TP391

基于词性及词性依存的句子结构相似度计算

蓝雁玲[1]，陈建超[2]

(1. 华南理工大学计算机科学与工程学院计算机系，广州 510006；2. 广东商学院数学与计算科学学院，广州 510320)

摘 要：为提高句子相似度的准确率，从结构相似度出发，提出基于词性及词性依存关系的句子结构相似度计算方法。该方法从正向和逆向比较句子的词性序列，获得2个句子词性及词性依存关系的最优匹配，从而计算句子结构相似度。实验结果表明，该方法能使句子结构相似度计算更合理。

关键词：句子结构；词性依存；结构相似度；自然语言处理

Chinese Sentence Structures Similarity Computation Based on POS and POS Dependency

LAN Yan-ling[1], CHEN Jian-chao[2]

(1. Dept. of Computer, Institute of Computer Science & Engineering, South China University of Technology, Guangzhou 510006, China;
2. School of Mathematics and Computing Science, Guangdong University of Business Studies, Guangzhou 510320, China)

【Abstract】In order to improve the accuracy of sentence similarity from structures similarity, this paper proposes a similarity measure method of Chinese sentence structures. It performs the optimal matching between the Part Of Speech(POS) sequences and POS dependency of two compared sentences. Experimental results show that the new method works well and it is more reasonable than the other methods.

【Key words】sentence structures; Part Of Speech(POS) dependency; structures similarity; natural language processing

DOI: 10.3969/j.issn.1000-3428.2011.10.015

1 概述

在自然语言处理领域中，随着基于实例的机器翻译、案例知识表示与检索、FAQ(Frequently Asked Question)自动回复等的提出，句子相似度计算成为核心问题。句子相似度在不同的应用领域有不同的含义。在案例知识表示与检索应用领域中，相似度反映与用户查询在语义上的匹配程度，相似度越高，表明该文本与用户请求越接近。当用户提交问题后，系统利用特征索引和相似度方法从案例库中找出与当前问题最佳匹配的案例。合理的句子相似度计算方法是自然语言处理领域中的关键步骤。现有句子相似度算法有基于语义、结构或是两者之间的结合，其中，语义是句子词的信息；结构体现词之间的修饰关系。因此，语义相似和结构相似是计算句子相似度的2个关键因素。

为提高句子相似度的准确率，本文从结构相似度角度出发，提出基于词性及其依存的句子结构相似度计算方法。该方法从表面结构、结构特征2个方面刻画句子信息。在计算句子结构相似度时，分别从正向和逆向比较2个句子的词性序列以获得2个句子词性及其依存关系的最优匹配路径，并将此最优值作为句子结构相似度。

2 相关算法

针对句子相似度问题，已有不少学者做了大量的工作，现有算法按照对语句分析深度来看，主要分为2类：

(1)基于向量空间模型的方法，常用算法有TF-IDF。

(2)对语句进行完全的句法和语义分析，这是一种深层结构分析法，对被比较的2个句子进行深层的句法分析，找出依存关系，并在依存分析结果的基础上进行相似度计算。

文献[1]提出一种基于关键词加权的汉语句子相似度计算方法，该方法在计算句子相似度时，综合考虑语法与语义2个层次的相似度，融合了它们的优点。文献[2]从表面结构、结构特征和语义特征三维对刻画句子信息。文献[3]将词串粒度和结构结合起来计算句子相似度，考虑了相同词串的数目及长度和对应的权值信息。文献[4]基于语义和单词序列信息来计算句子或短文之间的相似度，该方法考虑了词的次序。

3 基于词性及词性依存的句子结构相似度计算

3.1 词性依存的定义

一个完整的句子由主成分和修饰成分组成。主成分一般为句子中的核心动词，是句子的支配者，修饰成分用来描述语境，从属于支配者。相同的主成分可以由不同修饰成分来修饰，以达到不同的渲染效果。因此，要整体把握句义，需要了解主成分和修饰成分之间的支配关系，即连续词串之间的依存关系。现有信息研究依存关系的5条公理[5]如下：

(1)一个句子中只有一个成分是独立的；

(2)其他成分直接依存于某一成分；

(3)任何一个成分都不能依存于2个或2个以上的成分；

(4)如果A成分直接依存于B成分，而C成分在句中位于A和B之间，那么C或者直接依存于B，或直接依存处于A和B之间的某一成分；

(5)中心成分左右两边的其他成分相互不发生关系。

句子成分信息可由词性来反映，词性依存关系中各成分之间的修饰关系体现了句子的整体性，而其词之间的距离体现了句子的连续性。通过分析句子词性及其依存来计算句子

基金项目：广东省自然科学基金资助项目(07006474)；广东省科技攻关基金资助项目(2007B010200044)

作者简介：蓝雁玲(1985－)，女，硕士，主研方向：数据挖掘；陈建超，博士

收稿日期：2010-10-29 **E-mail：**widegooseblue@163.com

结构相似度，可获得句子表面结构和结构特征的相似度。

3.2 相似度计算

有些修饰词在句中没有任何的信息价值，因此，在计算句子结构相似度之前，需要进行数据预处理，包括过滤停用词和分词处理。

本文把句子结构相似度定义为句子间词性及其依存关系的最佳匹配程度，取值在[0, 1]之间。取值为 0 时，表明 2 个句子在词性及其依存关系上完全不相同。取值为 1 时，表明 2 个句子在词性及其依存关系上完全相同。

数据预处理后，设长句由($L_1, L_2,\cdots,L_m$)表示，短句由($S_1, S_2,\cdots,S_n$)表示，其中，m 为长句中词的总数；n 为短句中词的总数，且 $m\geqslant n$。两句词性相似度矩阵如式(1)所示：

$$\boldsymbol{P}_{\mathrm{Sim}(m\times n)}=\begin{vmatrix} S_{11} & S_{12} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & S_{23} & \cdots & S_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ S_{m1} & S_{m2} & S_{m3} & \cdots & S_{mn} \end{vmatrix} \tag{1}$$

其中，S_{ij} 表示长句中第 i 个词性和短句中第 j 个词性的相似度，若两词性相等，则 S_{ij}=1；否则 S_{ij}=0。

下面给出定义和句子结构相似度计算方法。记短句中第 1 个在长句中找到相同词性的词性为 sF，其在短句中的位置为 sL，在长句中的位置为 lL。用 $sCurPOS$ 表示短句当前词性，初值为 sF，用 $lCurPOS$ 表示长句当前词性，初值为 sF。

定义 1(对应词) 对应词有 2 种：正对应词和偏对应词。称长句 lL 位置上的词性为短句 sL 位置上词性的正对应词，$lCurPOS$ 左边第一个词性为 $sCurPOS$ 左边第一个词性的偏对应词，以此类推。

定义 2(相邻对应词间距 d) 初值 d=0。若短句 sL+1 位置上的词性和长句 i 位置上的词性相同，且 lL+1≤i<m，则长句 i 位置上的词性为短句 sL+1 位置上词性的正对应词，此时 d=$i-lL-1$。否则，长句 lL+1 位置上词性为短句 sL+1 位置上词性的偏对应词，d++。长句当前词性 $lCurPOS$ 更新为当前对应词，短句当前词性 $sCurPOS$ 更新为 sL+1 位置上的词性，以此类推。相邻正对应词间距体现出了距离越大、相似度越低的特点。

定义 3(短句前余词) 记 $sPreC$ 为短句前余词数，若 sL>lL，有 $sPreC$=$sL-lL$。称从短句 0 位置到 $sPreC-1$ 位置的所有词为短句的前余词。

定义 4(短句后余词) 记 $sSufC$ 为短句后余词数，若短句最后一个词性在长句中无对应词，且短句中最后一个在长句中有对应词的词性位于短句中的 i 位置，则 $sSufC$=$n-i-1$。称从短句 i+1 位置到 $n-1$ 位置的所有词为短句的后余词。用前余词和后余词来刻画句中词的顺序。

定义 5(配对路径) 短句中每个词性在长句中对应词的位置及其依存关系称为配对路径。

例如：

(1)句 1：北京铁路运输检察院以京铁检刑诉[2009]0014 号起诉书指控被告人崔世亮犯掩饰、隐瞒犯罪所得罪。

(2)句 2：检察院指控被告人崔世亮犯掩饰、隐瞒犯罪所得罪的事实清楚。

数据预处理后：

(1)句 1 词性序列：

ns n vn n p b n ng n vg n n vn n nr v v v vn usuo vg

(2)句 2 词性序列：

n vn n nr v v v vn usuo vg n a

有 sF=n, sL=0, lL=1, $sPreC$=0, $sSufC$=2，配对路径示意图如图 1 所示，其中，#表示该词性在长句中没有对应词。

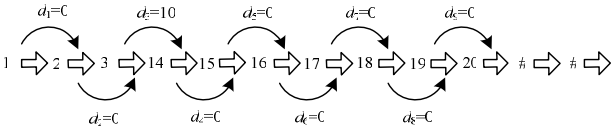

图 1 配对路径示意图

记两句结构相似度为 $stru_{\mathrm{Sim}}(S_1, S_2)$，则：

$$struSim(S_1,S_2)=\frac{\sum_{i=1}^{c}\frac{W_i}{1+d_i^2}}{\sum_{j=1}^{r}\frac{WS_j+WL_j}{2}+\sum_{k=1}^{sPreC+sSufC}WS_k+\sum_{t=1}^{m-r}WL_t} \tag{2}$$

其中，c 为 S_1 和 S_2 中相同词性的个数；r 为对应词(包括正、偏对应词)的总数；$\sum_{j=1}^{r}\frac{WS_j+WL_j}{2}$ 表示对应词对中两词性权值均值线性和；$\sum_{k=1}^{sPreC+sSufC}WS_k$ 表示短句中所有前余词、后余词词性权值线性和；$\sum_{t=1}^{m-r}WL_t$ 表示长句中所有非对应词词性权值线性和。

考虑 2 种特殊情况：

(1)S_1 和 S_2 结构完全相同。如果 S_1 和 S_2 结构完全相同，则式(2)中 n=m=c=r, $sPreC$=$sSufC$=0，所有的 d_i=0，分母等于分子，句子结构相似度等于 1。

(2) S_1 和 S_2 结构完全不相同。如果 S_1 和 S_2 结构完全不相同，有 c=0，句子结构相似度等于 0。

3.3 算法描述

输入 待计算句子结构相似度的 2 个句子 S_1 和 S_2

输出 S_1 和 S_2 的结构相似度

Step1 数据预处理

去掉 S_1 和 S_2 中的停用词，分别得到 S'_1 和 S'_2，对 S'_1 和 S'_2 进行分词处理后提取两句的词性序列。

Step2 计算词性相似度矩阵

计算短句中每个词性和长句中所有词性的相似度，得出两句子词性相似度矩阵 $\boldsymbol{P}_{\mathrm{Sim}(m\times n)}$，保存 $\boldsymbol{P}_{\mathrm{Sim}(m\times n)}$中首次出现 1 的列号 j，并保存该列中所有 1 的位置，列 j 有多少个 1，就有多少条配对路径。

Step3 搜索最佳配对路径

分别从正向和逆向进行搜索，计算每条配对路径的结构相似度，最后选择最大值作为 S_1 和 S_2 的结构相似度。

为了获得最佳的配对路径，本文从正向和逆向进行搜索，下面给出正向搜索过程：

(1)若词性相似度矩阵 $\boldsymbol{P}_{\mathrm{Sim}(m\times n)}$中没有 1 值，则 $stru_{\mathrm{Sim}}$=0.0，结束。

(2)若列 j 只有一个 1 值，说明只有一条配对路径，按照式(2)计算该条配对路径的结构相似度，将其作为 S_1 和 S_2 的结构相似度值，结束。

(3)若列 j 有多个 1 值，按照式(2)计算每条配对路径的结构相似度，将相似度最大的路径作为正向的最佳配对路径，其相似度值作为正向最大结构相似度。

将句子 S_1 和 S_2 词性序列逆序，重复步骤(2)和步骤(3)计算逆向最大结构相似度，最后从正向最大结构相似度和逆向最大结构相似度中选择最大值作为 S_1 和 S_2 的结构相似度。

4 实验结果与分析

从北京法院网刑事文书的 100 个案例中抽取 1 000 个句子作为句子库，从句子库中手工获取 20 个不同结构的句子作

为源句集，剩下的 980 个句子作为测试集。为了方便计算出句子结构相似度后基于人名进行聚类，本文分配权值 9 给 nr(人名)，分配权值 7 给 n(名词)，分配权值 4 给 v、vi、vn(动词类)，分配权重 2 给 p(介词)，分配权值 1 给其他词性。为体现本算法的有效性，将与文献[3]提出的基于词类串的句子结构相似度计算方法进行比较，部分实验比较结果如表 1 所示。

表 1　部分句子结构相似度实验结果

源句与相似句	词性序列	句子结构相似度	
		本文方法	文献[3]方法
源句 1：指控被告人崔世亮犯故意伤害罪	vn n nr v b n	—	—
相似句 A_1：指控被告人崔世亮犯故意伤害罪	vn n nr v b n	1.0	1.0
相似句 A_2：判处被告人林勇犯故意伤害罪	v n nr v b n	0.875	0.875
相似句 A_3：被告人崔世亮犯故意伤害罪罪名成立	n nr v b n vi	0.778	0.875
相似句 A_4：代理人赵鸿江	n nr	0.500	0.667
相似句 A_5：延庆县人民检察院指派检察员聂智慧出庭支持公诉	ns nt v n nr1 n vi v vn	0.177	0.138
源句 2：职员张岩因琐事与韩冬发生口角	n nr p n p nr v n	—	—
相似句 B_1：邻居钱新华用脚将朱孝伟踹伤	n nr p n nr v v	0.726	0.701
相似句 B_2：被告人郭利荣因同事田玉兰指责自己划破其三轮车车胎	n nr p n nr v rr v n n	0.581	0.625
相似句 B_3：因犯人郭志强在取保候审期间逃跑	p n nr p vi vi f v	0.413	0.456
相似句 B_4：附带民事诉讼原告人韩冬要求被告人张岩赔偿医药费 3 707.49 元	v b vn n n nr v n nr v n q	0.350	0.553
相似句 B_5：且能赔偿被害人的经济损失	v vn n n n	0.160	0.204
源句 3：被告人崔世亮在开庭审理过程中亦无异议	n nr p vi v n f d v n	—	—
相似句 C_1：当事人王东辉在开庭审理过程中均无异议	n nr p vi v n f d v n	1.0	1.0
相似句 C_2：被告人吴广顺、吴广德在开庭审理过程中均无异议	n nr nr1 ns p vi v n f d v n	0.910	0.938
相似句 C_3：证人蒋鹏飞的证言及辨认笔录	n nr vi n v n	0.623	0.629
相似句 C_4：根据被告人崔世亮的犯罪情节和悔罪表现	n nr vn n vi v	0.357	0.415
相似句 C_5：待原告人霍志斌实际做出后	p n n nr ad v f	0.213	0.241

表 1 给出 3 个源句作为输入句，分别从句子库中找出结构相似度范围不同的 5 个句子来进行比较。从词性序列可以很容易看出，A_2 与源句 1 结构相似度应大于 A_3 与源句 1 结构相似度，文献[3]提出的方法没有考虑到词的顺序，得出两者的结构相似度相等。A_4 词性序列是源句 1 词性序列的子序列，且所有词权值和是源句 1 所有词权值和的 1/2，因此，两句结构相似度为 0.5，符合判断。而文献[3]由于加大了所有相同词权值和的比重，得出的结果大于 0.5。在考虑了对应词对权值均衡和距离越大相似度越低 2 个因素后，本算法计算 B_3 与源句 2 的结构相似度大于 B_4 与源句 2 的结构相似度。由于文献[3]提出的算法并不能真正体现出距离越大、结构相似度越低的特点，因此得出来的结果正好相反。从词顺序和词总数可以明显看出，B_2 与源句 2 结构相似度远大于 B_4 与源句 2 结构相似度，而文献[3]得出的相似度值相差不大。可见，本算法在反映句子结构相似度时其结果比文献[3]提出的方法更趋于合理。

5　结束语

本文通过词性及其依存来计算句子结构相似度，在一定程度上提高了句子结构相似度的准确率。但本文算法没有对语义进行详细的分析，下一步的研究内容是在本算法的基础上加上语义分析，结合语义相似度和结构相似度计算句子相似度，进一步提高句子相似度准确率。

参考文献

[1] 裴　婧, 包　宏. 汉语句子相似度计算在 FAQ 中的应用[J]. 计算机工程, 2009, 35(17): 46-48.

[2] Liu Yi, Liu Qiang. Sentence Similarity Computation Based on Feature Set[C]//Proc. of the 13th International Conference on Computer Supported Cooperative Work in Design. Santiago, Chile: [s. n.], 2009: 751-756.

[3] Wang Rongbo, Wang Xiaohua, Chi Zheru, et al. Chinese Sentence Similarity Measure Based on Words and Structure Information[C]//Proc. of the 7th International Conference on Advanced Language Processing and Web Information Technology. Dalian, China: [s. n.], 2008: 27-31.

[4] Li Yuhua, McLean D, Bandar Z A, et al. Sentence Similarity Based on Semantic Nets and Corpus Statistics[J]. IEEE Trans. on Knowledge and Data Engineering, 2006, 18(8): 1138-1150.

[5] 李　彬, 刘　挺, 秦　兵, 等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研究, 2003, 20(12): 15-17.

编辑　顾姣健

(上接第 46 页)

于网站拓扑图的路径补充算法。在用户聚类阶段，提出一种综合访问次数，访问次序和访问时间等多种评价向量的用户相似度计算方法，并在此基础上采用直接聚类的方法进行聚类操作。通过 Davies-Bouldin 指标，清楚地反映了本文算法的聚类效果较文献[6-7]更有效。在今后的研究中，可以对如下方面进行改进：预处理阶段用户识别中动态设置断开阈值，同时直接聚类算法的效率有待进一步提高。

参考文献

[1] 李　燕, 冯博琴, 鲁晓峰. Web 日志挖掘中的数据预处理技术[J]. 计算机工程, 2009, 35(22): 44-46.

[2] 付志涛. 基于 Web 日志的网络用户聚类研究与实现[D]. 南京: 南京理工大学, 2007.

[3] 刘茂福, 何炎祥, 彭　敏. Web 模糊聚类方法及其应用[J]. 计算机科学, 2005, 32(1): 155-158.

[4] 方元康. 基于模糊聚类的 Web 日志挖掘研究[D]. 合肥: 合肥工业大学, 2008.

[5] Davies D L, Bouldin D W. A Cluster Separation Measure[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1979, 1(4): 224-227.

[6] 汪永成. 模糊聚类算法研究及在 Web 日志挖掘中的应用[D]. 阜新: 辽宁工程技术大学, 2008.

[7] 马晓艳, 唐　雁. 一种基于用户浏览路径的 Web 用户聚类方法[J]. 西南师范大学学报: 自然科学版, 2009, 34(3): 93-97.

编辑　陈　文