

# 基于语义树的概念语义相似度计算方法研究

安建成, 武俊丽

(太原理工大学 计算机与软件学院, 山西 太原 030024)

**摘要:** 在分析传统语义相似度计算方法的基础上, 综合考虑了边的深度、密度、强度及两个概念的语义重合度、层次差等主要影响因素, 提出了一种基于语义树的概念语义相似度计算方法, 并验证了该算法的合理性。

**关键词:** 概念语义相似度; 语义距离; 信息检索

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-7180(2011)01-0138-04

## Research on Method of Concept Semantic Similarity Computation Based on Semantic Tree

AN Jian-cheng<sup>1</sup>, WU Jun-li<sup>2</sup>

(Institute of Computer and Software, Taiyuan University of Technology, Taiyuan 030024, China)

**Abstract:** The main influential factors: depth, density, strength of the edge and concept overlap degree, concept level difference are considered comprehensively on the basis of analyzing the traditional method of calculating semantic similarity, proposes a semantic tree-based concept semantic similarity computation method, and proves the algorithm's rationality.

**Key words:** concept semantic similarity; semantic distance; information retrieval

### 1 引言

概念相似度在信息检索、信息抽取、机器翻译、词义排歧、文本聚类、文本映射等方面都有着广泛的应用<sup>[1-2]</sup>。语义检索在信息检索领域扮演着越来越重要的角色, 它突破了传统信息检索思想的束缚, 对信息进行语义匹配。语义检索首先对用户输入的检索信息进行语义分析、语义推理和语义扩展, 形成扩展后的语义查询, 最终获得相关信息资源。在进行语义推理和扩展的过程中需要运用到概念间的语义相似度计算, 来尽可能地提高信息的查全率和查准率。

文中在分析传统语义相似度计算方法的基础上, 提出了一种综合考虑多个因素的概念语义相似度计算方法, 并验证了该算法的有效性。

### 2 概念语义相似度

#### 2.1 概念语义相似度定义

概念语义相似度是一个主观性非常强的概念,

在不同的应用领域中可能有着不同的含义。在基于实例的机器翻译中, 语义相似度主要用于衡量文本中词语的可替换程度; 而在信息检索中, 相似度更多的要反映文本与用户查询在意义上的符合程度, 相似度越高, 表明该文本与用户的查询请求越接近。

当两个概念元素具有某些共同特征时, 则定义它们是相似的, 用  $\text{Sim}(C_1, C_2)$  来表示概念  $C_1, C_2$  之间的相似度。形式上概念语义相似度计算应满足:

(1) 值为  $[0, 1]$  区间中的一个实数, 即:  $\text{Sim}(C_1, C_2) \in [0, 1]$ 。

(2) 两个概念是完全相同的, 则相似度为 1, 即:  $\text{Sim}(C_1, C_2) = 1$ , 当且仅当  $C_1 = C_2$ 。

(3) 两个概念没有任何共同特征, 那么相似度为 0, 即:  $\text{Sim}(C_1, C_2) = 0$ 。

(4) 相似关系是对称的, 即:  $\text{Sim}(C_1, C_2) = \text{Sim}(C_2, C_1)$ 。

#### 2.2 概念语义相似度计算

概念相似度计算是计算两个概念内在含义之间

的相似度.其前提是两个概念具有一定的语义相关性,并且,当它们在概念间的结构层次网络图中存在一条通路.对于概念的语义相似度计算方法,国内外的学者已经对此做了大量的研究工作<sup>[2-4]</sup>.

为了使得本体内部概念间的相似度计算更加客观、准确,进行语义相似度计算时应考虑到以下的原则和因素:量化原则、简单性原则、充分利用本体的特性、可调节性等原则;考虑到概念语义重合度、语义距离、层次深度、调节因子等因素.

### 2.2.1 传统的语义相似度计算方法

#### (1) 基于距离的语义相似度计算

首先求出两个概念的语义距离,然后再将其转化为语义相似度,计算公式如下:

$$\text{Sim}(C_1, C_2) = \frac{\lambda}{\lambda + \text{Dist}(C_1, C_2)} \quad (1)$$

式中  $C_1, C_2$  表示语义树中的两个概念,  $\text{Dis}(C_1, C_2)$  表示概念  $C_1, C_2$  之间的语义距离,  $\lambda$  为一个可调节参数,一般其含义为语义相似度为 0.5 时的语义距离.当  $\text{Dis}(C_1, C_2) \in [0, \infty)$  时  $\text{Sim}(C_1, C_2) \in [0, 1]$ , 当  $\text{Dis}(C_1, C_2) = 0$  时  $\text{Sim}(C_1, C_2) = 1$ ,  $\text{Dis}(C_1, C_2) \rightarrow \infty$  时  $\text{Sim}(C_1, C_2) = 0$ .

该方法的基本思想是把概念间的语义距离用它们在语义树层次结构中的几何距离来量化,该算法是建立在具有一定语义相关性的两个概念在语义网络中存在一条通路的假设基础上的,这种算法直观、简单、易于理解.但没有考虑到语义树结构中有向边的差异,其计算结果比较依赖于所构建领域本体语义树的完备性,有时不能很好地反映客观结果.

#### (2) 基于内容的语义相似度计算

基于内容的语义相似度计算模型的基本原理是:两个概念共享的信息越多,它们之间的语义相似度也就越大,反之则越小.这种计算方法充分利用了概率统计理论的相关知识,因此更具有说服力,但这种方法只能粗略地量化概念之间的语义相似度,不能更加细致地区分各概念语义相似度的值.

#### (3) 基于属性的语义相似度计算

在现实世界中,人们在区别和联系不同事物的过程中一般都通过比较事物之间的属性.如果两个事物有很多属性相同,则说明这两个事物很相似,反之,则相反.

### 2.2.2 改进的语义相似度计算方法

首先给出有关层次树中概念之间关系分类的三个相关定义<sup>[4-6]</sup>:

**定义 1** 在本体概念的树状层次结构中,如果

概念  $A$  和概念  $B$  存在这样的关系: $A$  是  $B$  的祖先,则  $A$  和  $B$  为同支概念,概念  $A$  称为  $A$  和  $B$  的最近根节点,记为  $LN(A, B)$ ,  $A, B$  之间的距离为:  $\text{Dist}(A, B) = \text{Dep}(B) - \text{Dep}(A)$ , 其中  $\text{Dep}(C)$  为概念  $C$  在层次结构中的深度.如图 1(a) 所示.

**定义 2** 在本体概念的树状层次结构中,如果概念  $A$  和  $B$  之间存在这样的关系: $A$  不是  $B$  的祖先且  $B$  也不是  $A$  的祖先,则称  $A$  和  $B$  为异支概念,如果概念  $R$  是  $A$  和  $B$  共同的祖先,并且是符合此条件的所有节点中距离语义树的根节点最远的一个,则称  $R$  为  $A$  和  $B$  的最近根节点,也记为  $LN(A, B)$ ,  $A, B$  之间的距:  $\text{Dist}(A, B) = \text{Dist}(A, R) + \text{Dist}(B, R)$ . 如图 1(b) 所示.

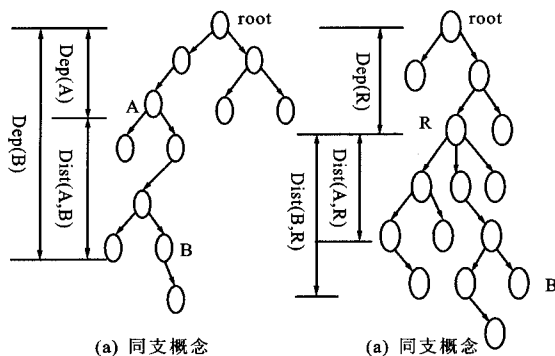


图 1 同支与异支概念

**定义 3** 概念  $C$  称为概念  $A$  和  $B$  的语义相关概念,当且仅当概念  $C$  满足如下的条件:当  $A, B$  为同支概念时,  $C$  在以  $A$  为根的子树中且不在以  $B$  为根的子树中;当  $A, B$  为异支概念时,  $C$  在以  $R$  为根的子树中且不在以  $A$  或  $B$  为根的子树中.

在计算概念相似度时,  $A$  和  $B$  各自的子节点数以及它们的语义相关概念数对相似度计算结果也有影响,当  $A, B$  为同支概念时,  $A$  的子概念由  $B$  的子概念和  $A, B$  的语义相关概念组成,前者所占的比重越大,概念  $A, B$  的语义相似度越大;当  $A, B$  为异支概念时,  $R$  的子概念由  $A$  的子概念、 $B$  的子概念以及  $A, B$  的语义相关概念三部分组成,前两部分的比重越大,  $A, B$  的语义相似度越大. 因此有:

当  $C_1, C_2$  为同支概念,并且  $C_2$  为  $C_1$  的下位概念时,  $\text{Son}(C)$  表示概念  $C$  的子节点个数.

$$\text{Sim}(C_1, C_2) = \gamma * \frac{\text{Son}(C_2)}{\text{Son}(C_1)} \quad (2)$$

当  $C_1, C_2$  为异支概念时:

$$\text{Sim}(C_1, C_2) = \gamma * \frac{\text{Son}(C_2) + \text{Son}(C_1)}{\text{Son}(LN(C_1, C_2))} \quad (3)$$

层次树中的概念有且只有这两种关联关系,因

此在文中最后定义概念相似度计算公式时,把概念分为这两种情况来考虑。

下面给出几个影响语义树有向边权值的因素定义式:

**定义 4** 有向边的深度:

将有向边的深度定义为

$$\text{Dep}(e_{ij}) = \min(\text{Dep}(C_i), \text{Dep}(C_j)) \quad (4)$$

其中  $e_{ij}$  表示连接概念  $C_i$  和  $C_j$  的有向边,  $\text{Dep}(C)$  表示概念  $C$  在层次树中的深度. 用  $R$  表示层次树的根节点, 并且设根节点的输入为 1:  $\text{Dep}(R) = 1$ ,  $\text{Dep}(C) = \text{Dep}(\text{Parent}(C)) + 1$ . 对与根节点相连的边来说其深度为 1. 随着有向边深度的增加, 其权值减小, 概念间的语义距离减小, 语义相似度增加. 即有向边深度对边的权值的影响公式为:

$$\text{Weight}_{\text{dep}}(e_{ij}) = \frac{1}{m^{\text{Dep}(e_{ij})}} \quad m \geq 2 \quad (5)$$

**定义 5** 有向边所处局部密度:

将有向边的密度定义为:

$$\text{Des}(e_{ij}) = \frac{\bar{G}}{\text{outd}(\text{LN}(C_i, C_j))} \quad (6)$$

其中  $\bar{G} = \frac{\sum_{i=1}^n \text{outd}(C_i)}{n}$  为整个层次树的平均密度:

即所有节点出度之和和节点个数的比值. 随着有向边所处密度的增加, 边的权值逐渐减小, 概念的语义距离减小, 语义相似度增大, 所以有向边密度对权值的影响公式为:

$$\text{Weight}_{\text{des}}(e_{ij}) = \frac{\bar{G}}{\text{outd}(\text{LN}(C_i, C_j))} \quad (7)$$

**定义 6** 有向边的强度:

两个概念共享的信息越多, 它们之间的相似度就越大, 反之则越小. 每一个概念都可以看成是对其祖先节点的细化, 即每一子节点包含其所有祖先节点的信息. 为了充分利用信息理论和概率统计理论的相关知识, 文中将本体库的统计特征引入到相似度的计算上来, 这里记  $P(C)$  为

$$P(C) = \frac{\text{概念 } C \text{ 在训练资料中出现的次数}}{\text{训练资料的总数}} \quad (8)$$

则概念  $C$  的信息量  $\text{Info}(C)$  为:

$$\text{Info}(C) = -\lg(P(C)) \quad (9)$$

接着定义有向边  $e_{ij}$  的强度为:

$$\text{Int}(e_{ij}) = |\text{Info}(C_i) - \text{Info}(C_j)| \quad (10)$$

经过分析, 有向边的权值随着其强度的增大而减小, 所以有向边强度对其权值的影响公式为:

$$\text{Weight}_{\text{int}}(e_{ij}) = \frac{1}{\text{Int}(e_{ij})} \quad (11)$$

**定义 7** 两概念间的语义距离:

$$\text{Dist}(C_1, C_2) = \sum_{i=1}^{\text{Dist}_0(C_1, C_2)} \text{Weight}(e_{ij}) \quad (12)$$

其中  $\text{Dist}_0(C_1, C_2)$ , 为概念  $C_1, C_2$  之间的最短路径长度,  $e_{ij}$  为连接两概念的  $C_i, C_j$  边, 所以层次树中有向边的深度、密度、强度对其权值的影响公式为:

$$\text{Weight}(e_{ij}) = \text{Weight}_{\text{dep}}(e_{ij}) * \text{Weight}_{\text{des}}(e_{ij}) * \text{Weight}_{\text{int}}(e_{ij}) \quad (13)$$

**定义 8** 概念节点的深度差:

$$\text{DepthDis}(C_1, C_2) = |\text{Dep}(C_1) - \text{Dep}(C_2)| \quad (14)$$

其中,  $\text{Dep}(C_1), \text{Dep}(C_2)$  分别表示两概念节点在层次树中的深度. 随着两概念层次差的增大, 概念的相似度减小, 这个因素和边的权值没多少联系, 而是直接影响了概念的语义相似度, 所以有:

$$\text{Sim}(C_1, C_2) = \alpha * \frac{1}{\text{DepthDis}(C_1, C_2)} \quad (15)$$

其中  $\alpha$  为调节因子

**定义 9** 两概念的语义重合度:

语义重合度是指本体内部概念之间包含相同上位概念的个数. 语义重合度表明了两个概念间的相似程度. 在实际的计算中, 可以转化为两节点公共节点的个数. 若用  $UP(X)$  代表从  $X$  节点上溯到根节点的集合 (包括自身节点), 则可以用  $UP(X) \cap UP(Y)$  来表示  $X, Y$  两个概念节点上溯到根节点的公共节点的集合, 用  $|UP(X)|$  表示  $UP(X)$  的值.

定义概念  $C_1, C_2$  的语义重合度计算公式为:

$$OL(C_1, C_2) = \frac{|UP(C_1) \cap UP(C_2)|}{\text{Dep}(T)} \quad (16)$$

其中  $\text{Dep}(T)$  为层次树的最大深度. 随着两概念语义重合度的增大, 它们的相似度也随之增大, 即有:

$$\text{Sim}(C_1, C_2) = \beta * OL(C_1, C_2) \quad (17)$$

其中  $\beta$  为调节因子.

最后结合式(1)和上面的同支、异支概念给出概念语义相似度计算公式:

$C_1, C_2$  为同支概念, 且  $C_1$  为  $C_2$  的上位节点:

$$\text{Sim}(C_1, C_2) =$$

$$\omega * \frac{\lambda}{\lambda + \eta \text{Dist}(C_1, C_2)} + \beta * OL(C_1, C_2) + \delta * \left( \alpha * \frac{1}{\text{DepthDis}(C_1, C_2) + 1} + \gamma * \frac{\text{Son}(C_2)}{\text{Son}(C_1)} \right)$$

$C_1, C_2$  为异支概念:

$$\text{Sim}(C_1, C_2) =$$

$$\omega * \frac{\lambda}{\lambda + \eta \text{Dist}(C_1, C_2)} + \beta * OL(C_1, C_2) + \delta *$$

$$\left(\alpha * \frac{1}{DepthDis(C_1, C_2) + 1} + \gamma * \frac{Son(C_2) + Son(C_1)}{Son(LN(C_1, C_2))}\right)$$

(18)

其中  $\omega, \beta, \delta$  分别为概念间的语义距离、两概念的语义重合度、概念层次差影响概念语义相似度的调节因子,  $\omega + \beta + \delta = 1$ .

3 计算结果及分析

文中通过计算同一个语义树结构下其中一个概念与其他若干个概念的语义相似度来检验算法的可行性. 基于上述的相似度算法, 选取了 WordNet2. 0 中关于计算机的局部语义树来进行实验. 语义树结构图如图 2 所示, 节点旁边的数字为概念节点的信息量.

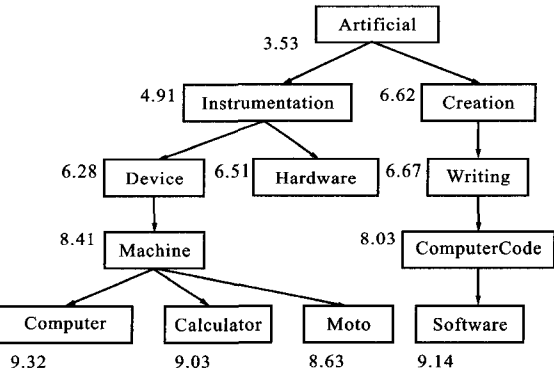


图 2 WordNet2. 0 关于 Computer 的语义树局部图

首先利用式(19)和(20)计算, 其中  $\alpha$  取 2, 然后利用文中算法即式(18)来计算 Computer 和其他几个概念间的语义相似度.

$$Sim(C_1, C_2) = -\lg \frac{Dist(C_1, C_2)}{2Dep(C_1, C_2)}$$

(19)

$$Sim(C_1, C_2) = 1 - \sqrt{\frac{\alpha - 1}{\alpha} \times \frac{Dep(C_2)}{Dep(C_2) + Dep(C_1)} \times Dist(C_1, C_2)}$$

(20)

取参数  $\lambda = 2, \eta = \frac{1}{2}, \omega = \frac{8}{10}, \beta = \frac{1}{10}, \delta = \frac{1}{10}, \alpha = \frac{1}{2}, \gamma = \frac{1}{2}, m = 2$ .

在计算语义相似度时应根据具体的情况设置调节因子, 实验中的语义树深度为 5, 即  $Dep(T) = 5$ ,  $Dep(e_{ij})$  的取值为  $[1, 5]$ ,  $Dist_0(C_1, C_2)$  的取值为  $[1, 8]$ , 式中的调节因子  $\alpha, \gamma$  实质上都是影响两概念间层次差的, 所以取  $\alpha = \frac{1}{2}, \gamma = \frac{1}{2}$ , 式(18)中  $\omega, \beta, \delta$  的取值直接影响到计算结果, 考虑到边的深度和

密度对语义相似度的影响较大, 所以取  $\omega = \frac{9}{10}, \beta = \frac{1}{20}, \delta = \frac{1}{20}$ . 首先确定以上几个调节因子的值, 之后根据专家对 Computer—Calculator 这一对概念的相似度的经验值来判断, 发现取  $\eta = \frac{1}{2}$  时更接近专家判断值, 所以取  $\eta = \frac{1}{2}$  来计算以下几对概念的语义相似度. 计算结果如表 1 所示.

表 1 方法得到的概念相似度的比较

| 概念对                 | 算法<br>(19) | 算法<br>(20) | 文中算法<br>(18) | 专家<br>经验值 |
|---------------------|------------|------------|--------------|-----------|
| Computer—Computer   | 1          | 1          | 1            | 1         |
| Computer—Calculator | 0. 500     | 0. 9412    | 0. 953       | 0. 98     |
| Computer—Motor      | 0. 500     | 0. 9412    | 0. 941       | 0. 93     |
| Computer—Machine    | 0. 667     | 0. 9697    | 0. 948       | 0. 95     |
| Computer—Device     | 0. 500     | 0. 9143    | 0. 921       | 0. 94     |
| Computer—Hardware   | 0. 333     | 0. 7442    | 0. 894       | 0. 91     |
| Computer—Software   | 0. 200     | 0. 4848    | 0. 870       | 0. 92     |

目前, 对于语义相似度算法的优劣评判还没有一个较为通用的标准. 文中采用人工判别方法来看是否符合专家的经验值.

从表 1 中的数据可以看出: 文中算法在量化概念语义相似度方面更接近专家的经验值, 为了更进一步地证明该算法的合理性, 现在引入兼容度 (Compat) 作为评价相似度的一个指标来衡量公式, 计算结果和专家判断结果的吻合程度. Compat 越大表示计算结果和专家经验值越吻合. Compat 取值在  $[0, 1]$  之间.

定义 10 设  $Sim_i$  和  $Sub_i$  分别表示用公式计算和专家判断的第  $i$  对概念的相似度, 则记第  $i$  对概念的算法计算结果与专家判断结果的百分差为

$$E_i = \frac{|Sim_i - Sub_i|}{Sub_i} \times 100\%$$

(21)

定义 11 兼容度 基于上述定义, 定义算法结果与专家判断结果的兼容度 Compat 为

$$Compat = 1 - \frac{1}{N} \sum_{i=1}^N E_i$$

(22)

其中  $N$  为实验所选的概念对数目. 这样, 对一组概念对而言, Compat 表示了算法计算结果与专家判断它们相似度的吻合程度. 根据表 1 的计算结果, 得到三种算法所得与专家经验判断值之间的兼容度, 如表 2 所示. 可以看出, 文中方法计算结果与专家的判断较为吻合.

(下转第 146 页)

线程运行到稳定状态点后,由式(2)与式(3) $S_j = T_{hexx} + A_i$ 可分别计算出各 CPU 的带权运行时间,数据如表 2 所示.

表 2 带权时间表

| 线程数 | 巷道 | 1 巷道 | 2 巷道 | 3 巷道 | 4 巷道 | 5 巷道 |
|-----|----|------|------|------|------|------|
|     |    | 2    | 2    | 1    | 1    | 1    |

由此当环形穿梭车载货到达线程调度起点 P 时,系统将通知其将货物直接运送到 5 巷道对应的入库输送机,并且堆垛机会在相对最短的时间内将其执行完成,这将是一条最优执行路径.

7 结束语

文中研究了大型 AS/RS 中调度策略的综合性能要求,提出了一种优化利用所有资源,提高整体入库效率的多处理器多线程的综合调度策略,并给出调度策略的测试用例.主要成果包括:(1)将 Windows 内核调度理论引入自动化立体仓库调度领域;(2)采用了 Windows 内核多处理器多线程调度中线程数量控制函数;(2)给出了在大型 AS/RS 中为环形穿梭车选择适当入库输送机的函数;(3)提出了综合考虑多种因素的多处理器多线程调度策略,并给出对其详细的分析论证.该研究成果可应用于对大型自动化立体仓库调度系统效率的优化.

参考文献:

[1] 江勇,吴建平.多性能目标分组调度策略[J].软件学报,2002,13(8):1514-1520.

[2] Linn R, Wysk R At. An ExPert system based controller for AS/RS[J]. Intemational Journal of Production Research,1990,32(2):176-191.

[3] Kim Byung-In, Heragu Sunderesh S, Graves Robert, et al. A hybrid scheduling and control system aichitecture for warehouse management[J]. IEEE Transactionson Robotics& Automation, 2003,12(6): 6-11.

[4] Krishnaiah Chetty M OV, Sarveswar Reddy. Genetic algorithms for studies on AS/RS integrated with machines[J]. Int J Adv Manuf Technol, 2003(22): 932-940.

[5] Love R. Linux 内核设计与实现[M].陈莉君,译.北京:机械工业出版社,2004.

[6] 赵树林,张璟,李军怀,等.基于工作流的办公自动化系统的研究与实现[J].微电子学与计算机,2009,26(5):135-141.

[7] 赵佳华,伦立军.软件体系结构层的耦合性度量研究[J].微电子学与计算机,2009,26(5):103-109.

[8] Russinovich M E, Solomon D A. 深入解析 Windows 操作系统, Microsoft windows server 2003/windows XP/windows 2000 技术内幕[M].4 版.潘爱民,译.北京:电子工业出版社,2007.

作者简介:

王海宾 男,(1982-),硕士研究生,教师,CCF 会员.研究方向为信息系统与决策支持、数据库与软件工程技术.

(上接第 141 页)

表 2 三种算法的兼容度

| 方法       | 兼容度    |
|----------|--------|
| 算法(19)   | 0.6006 |
| 算法(20)   | 0.8504 |
| 文中算法(18) | 0.9585 |

4 结束语

文中在算法中引入了语义树中有向边的深度、密度和强度对边的权值的影响以及概念语义重合度和概念层次差对语义相似度的直接影响来计算概念间的语义相似度.并抽取 WordNet2.0 中关于计算机的局部语义树来进行实验,根据具体情况设定了调节因子.实验结果比较理想,但有些地方仍然不够完善.进一步的研究方向是考虑影响语义相似度的其它更多因素,并将概念语义相似度的计算应用到有关信息检索应用系统中.

参考文献:

[1] 陈沈焰,吴军华.基于本体的概念语义相似度计算及其应用[J].微电子学计算机,2008,25(12): 96-99.

[2] Agirre, Rigau. A proposal for word sense disambiguation using conceptual distance[C]// International Conference of Recent Advances in Natural Language Processing. Borovets, Bulgaria, 2003.

[3] 王进,陈恩红,施德明,等.一种基于语义相似度的信息检索方法[J].模式识别与人工智能,2006(6):2-6.

[4] 江敏,肖诗斌,王弘蔚,等.一种改进的基于《知网》的词语语义相似度计算[J].中文信息学报,2008,22(5):84-89.

[5] 张选平,蒋宇.一种基于概念的信息检索查询扩展[J].微电子学与计算机,2006,23(4): 110-114.

[6] 夏天.汉语词语语义相似度计算研究[J].计算机工程,2007,33(6):191-194.

作者简介:

安建成 男,(1964-),副教授.研究方向为语义网、数据库.

武俊丽 女,(1982-),硕士.研究方向为数据库.