# 一种本体概念的语义相似度计算方法

李文清 [1] 孙 新 [1] 张常有 [1, 2] 冯 烨 [3]

**摘 要** 概念语义相似度已广泛应用于 Web 服务发现、本体映射等领域, 但现有的概念语义相似度计算方法对概念间语义相似程度的区分不够细致. 本文从本体结构出发, 首先提出了自底向上的本体概念出现概率计算方法, 并在此基础上改进了基于节点信息量的概念语义相似性度量方法; 然后又设计了基于边计算的本体概念语义相似度计算方法; 最后对上述两种方法线性加权, 提出了一种加权的本体概念语义相似度计算方法. 实验结果表明该方法能进一步正确区分本体中父子概念及兄弟概念间的相似程度.

**关键词** 本体, 语义相似度, 概念出现概率, 信息量

**DOI** 10.3724/SP.J.1004.2012.00229

## A Semantic Similarity Measure between Ontological Concepts

LI Wen-Qing[1] SUN Xin[1] ZHANG Chang-You[1, 2] FENG Ye[3]

**Abstract** Concept semantic similarity is wildly used in web service matchmaking, ontology mapping and so on. But the existing concepts semantic similarity measuring methods cannot distinguish the similarities further. So in this paper, we firstly propose a bottom-up concept probability computation method based on ontology structure, and based on this probability, we improve an information content based semantic similarity method. Then, we design an edge based concept semantic similarity method. Finally, we linearly combine the two previous semantic similarity methods to form a weighted one. Result shows that the weighted one can distinguish similarity between concept and its children, or between siblings.

**Key words** Ontology, semantic similarity, probability of concept occurrence, information content

语义相似度源于计算语言学等领域, 主要研究术语、词汇或概念之间的相似程度, 被看作概念在分类上的相似程度, 广泛应用于自然语言处理中词义消歧[1]、知识管理中信息抽取[2]、语义标注[3] 以及本体学习[4] 与合并[5]、Web 服务发现[6−7] 等相关领域. 目前已出现很多概念语义相似性度量方法的研究成果. Sánchez 等[8] 根据这些方法所使用的数据源以及这些数据源的使用方式对它们进行分类.

基于词汇共同分布的方法[9−10] 利用词汇在给定语料库中的信息分布来度量语义. Lemaire 等[10] 采用统计和浅层语言分析技术来估计词汇在语料库中共现的程度, 从而计算词汇间的语义相似度. 此类方法认为词汇共现频率是它们相关的有力证明. 为了度量社会范围内词汇的使用情况, 此类方法需要尽可能通用的、完善的语料. 另外, 由于缺乏对文本语义的分析, 会产生对词汇的误解, 从而影响词汇共现的统计结果. 受益于 Web 的大规模性和通用性, 研究人员利用搜索引擎的查询技术来检索 Web 资源以度量概念间的相似性[8, 11].

基于结构化知识表示的方法将层次结构、WordNet、领域本体等结构化知识表达方式作为相似性计算的基础. 其中, 基于特征的方法根据本体概念描述模型中相同和不同的概念属性来估计概念间相似度. Tversky[12] 是基于特征的相似性度量方法的代表.

基于边计数的方法将概念相似性度量建立在本体中分割两个概念的语义连接数目之上[13−15]. Rada 等[13] 在概念层次树中, 用概念间最短距离来度量它们之间的相似性. Wu 等[14] 用两个概念及它们的最近共同祖先 (Nearest common ancestor, NCA) 概念分别到根概念的最短距离来度量它们之间的相似度. Leacock 等[15] 利用 WordNet 的结构, 考虑概念所在分类树的深度, 将文献 [13] 的语义距离转化为语义相似度. 本体中概念间的连接边仅表明概念间具有某种特定语义关系, 并不能量化概念间的语义距离, 但此类方法却认为所有连接边的长度相等, 表示相同的语义距离[11], 所以说该方法忽略了边之间语义连接强度的差异. 此外, 基于路径长度的方法依赖于本体中语义连接的完整性和本体的覆盖能力, 所以此类方法适用于具有良好的域间覆盖能力的本体, 如 WordNet.

收稿日期 2010-12-14 录用日期 2011-07-16
Manuscript received December 14, 2010; accepted July 16, 2011
国家自然科学基金 (60873208), 河北省自然科学基金 (F2009000929) 资助
Supported by National Natural Science Foundation of China (60873208) and Natural Science Foundation of Hebei Province (F2009000929)
本文责任编委 徐波
Recommended by Associate Editor XU Bo
1. 北京理工大学计算机学院 北京 100081 2. 石家庄铁道大学信息科学技术学院 石家庄 050043 3. 北京控制工程研究所 北京 100190
1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081 2. School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043 3. Beijing Institute of Control Engineering, Beijing 100190

基于信息论的相似性度量方法将概念信息量与本体知识相结合[16−19]. Resnik[16] 用两个概念的 NCA 概念的信息量度量它们的相似性. Lin[17] 将两个概念的 NCA 概念信息量与它们各自信息量之和的比值定义为两者之间的相似度. 杨立等[18] 用父子概念的差异信息量表示连接边的信息量. 白东伟[6] 结合概念局部密度和 Lin 的定义[17], 利用基于信息量的相似度函数计算边权重, 并将概念的深度影响内置于边权重的计算过程中. Pirro[19] 利用基于特征相似性理论中的一些概念并将它们转变到信息理论中, 充分利用信息量度量相似度. 概念信息量通常由概念在语料库中的分布来计算[8, 16−17, 20].

通过对相关研究分析发现, 目前基于本体的概念语义相似性度量方法普遍无法细致区分本体中兄弟概念以及父子概念间的相似程度. 另外, 从本体内部得到的概念信息量有助于提高概念相似度的准确性, 但概念信息量本身还需要进一步提升. 因此本文利用本体内部结构信息, 首先提出了自底向上的概念出现概率计算方法 (Bottom-up concept probability computation method, B-U) 以改善概念信息量质量, 并在此基础上改进了文献 [17] 中的方法; 然后设计了一种基于边计算的概念相似性度量方法; 最后, 对上述方法线性加权, 提出了加权概念相似度计算方法 (A weighted similarity measures between ontological concepts, WSim_OC).

## 1 本体概念信息量

根据信息论中信息量的定义, 现有概念信息量均采用 Resnik 提出的方法来计算, 如下所示.

$$IC_{\text{res}}(c) = -\log p(c) \tag{1}$$

其中, $IC_{\text{res}}(c)$ 为本体中概念 $c$ 所含信息量; $p(c)$ 为本体中概念 $c$ 的出现概率.

本体中概念 $c$ 的出现概率的定义如下:

**定义 1.** $p(c)$: 一个实例属于概念 $c$ 的概率, 简称概念 $c$ 的概率.

概念信息量由其概率决定, 因此获取准确的概念概率是信息量计算过程中的关键.

### 1.1 概念出现概率的计算

为了获得准确的概念相似度, 信息量的计算是非常关键的. 经典的信息论方法[16−17, 20] 通过计算语料中概念出现概率的倒数来获得概念的信息量, 以确保频繁出现的概念的信息量比较少出现的低. Resnik[16] 将名词在语料库中出现的次数累加到包含该名词的所有概念的出现次数. Richardson 等[20] 引入词汇类别因子, 改进了 Resnik 的方法. 然而基于语料库的方法扩展性弱, 存在数据稀疏等问题. Sánchez 等[8] 通过 Web 计算概念概率并利用特定语境内的联合查询来减弱词汇歧义性带来的影响. 此外, 本体有意义的组织方式为计算概念概率带来新思路, 研究人员提出基于本体内部信息的概念概率计算方法[6, 21]. Seco 等[21] 首先利用概念的下义概念计算信息量, 用概念所覆盖下义概念的总数与总概念数之比来度量概念概率. 白东伟[6] 根据层次概念树的结构, 将概念的概率平均分配给它所有子概念 (子概念指概念的直接孩子概念). 文中称为由上而下的概率平均分配法 (Top-down-aver, T-D-A), 公式如下:

$$p(c) = \begin{cases} 1, & c \text{ 是根概念} \\ \dfrac{1}{c(p)} \times p(p), & c \text{ 是其他概念} \end{cases} \tag{2}$$

其中, $p$ 为 $c$ 的父概念; $c(p)$ 为概念 $p$ 的子概念数; $p(c), p(p)$ 分别为概念 $c$, $p$ 的出现概率. T-D-A[6] 认为概念的实例与其各子概念实例的概率相等, 无法区分概念的各子概念对其概率贡献的差异.

### 1.2 B-U 概念概率计算法

概念的各个子概念对其概率的贡献不同, 为准确估计概念概率, 需在计算过程中体现出各子概念贡献的差异性. 本文根据本体结构, 提出了自底向上的概念概率计算方法 (B-U).

本体中的上层概念是其下层概念的泛化, 下层概念是其上层概念的具体化. 对于一个完整的领域本体而言, 叶子概念集合描述了该领域内所有具体概念, 也表达了全体具体概念的语义信息, 因此我们认为本体中任意概念的语义信息可由它所覆盖的子树中的所有叶子概念共同表达. 通常领域内的实例属于其本体中叶子概念的概率相等, 即本体中所有叶子概念的概率相等. 对于非叶子概念, 领域中属于非叶子概念 $c$ 的实例必定是其某个子概念的实例, 即非叶子概念概率等于其所有子概念概率之和, 也就是非叶子概念概率是其包含的全部叶子概念的概率之和. B-U 方法如下所示:

$$p(p) = \begin{cases} \dfrac{1}{l_{\text{count}}}, & p \text{ 是叶子节点} \\ \displaystyle\sum_{i=1}^{c(p)} p(c_i), & p \text{ 是其他节点} \end{cases} \tag{3}$$

其中, $p(p)$ 为概念 $p$ 的概率; $l_{\text{count}}$ 为本体中叶子概念的总数; $c_i$ 为概念 $p$ 的第 $i$ 个子概念; $p(c_i)$ 为 $c_i$ 的概率; $c(p)$ 为概念 $p$ 的子概念个数.

B-U 方法所得概念概率具有以下属性:

1) $p(p) \geq p(c_i)$: 当本体中的概念向下移动时, 对应概念的概率单调递减, 概念信息量则单调递增. 也就是说, 如果概念 $b$ 是概念 $a$ 的上义概念, 则 $p(b) \geq p(a)$.

2)　$p(r)=1$, 其中 $r$ 是本体的根概念节点. 根节点概率为 1, 这是由于领域内任何实例属于该领域本体顶层概念的概率为 1.

算法 1 列出了计算概念信息量的具体步骤.

**算法 1.** B-U 方法

**输入.** 本体结构;

**输出.** 概念概率, 概念 $IC_{\text{res}}$;

**步骤 1.** 遍历本体, 初始化概念 $IC_{\text{res}}$ 的计算标识 $ICF=0$, 置概念概率为 0, 统计叶子数 $l_{\text{count}}$;

**步骤 2.** 从根概念开始遍历本体结构中的概念;

**步骤 3.** 如果当前概念是根概念, 并且概念的 $ICF=1$, 则算法结束;

**步骤 4.** 如果当前概念是叶子概念, 计算其概率及该叶子的信息量, 并置其 $ICF=1$;

**步骤 5.** 如果当前概念的 $ICF=1$, 回溯到其父概念并返回概念的概率, 父概念成为当前概念;

**步骤 6.** 若当前概念的所有子概念都已返回, 则按式 (3) 计算当前概念的概率, 用式 (1) 计算其 $IC_{\text{res}}$, 置 $ICF=1$, 转步骤 3; 否则遍历当前概念的下一子概念, 转步骤 5.

## 2　加权概念相似度计算法

影响概念相似度的因素主要有概念深度、概念间距离、概念局部密度等. 本文将密度的影响融入信息量的计算过程中, 在基于边计算的概念语义相似性度量方法中体现深度对相似度的影响.

### 2.1　基于信息量的概念相似度

基于节点信息量的概念相似性度量方法认为共享信息量越高的两个概念的语义相似程度越高. 此类方法中, 概念语义相似性由它们的共享信息量和差异信息量共同决定. 该相似度随着概念间共享信息量增大而增大, 随它们差异信息量的增大而减小. 概念共享信息量用其 NCA 的信息量表示, 差异信息量由它们各自信息量与 NCA 信息量差之和表示.

Lin 在文献 [17] 中提出的方法是该类方法的代表, 然而 B-U 所得根概念信息量为 0, 直接应用于原方法时无法区分 NCA 是根节点的概念间相似度, 因此本文对此方法做如下改进:

$$sim_{IC}(a,b)=\frac{2IC_{\text{res}}(\text{NCA}(a,b))+\delta}{IC_{\text{res}}(a)+IC_{\text{res}}(b)+\delta} \tag{4}$$

其中, $a,b$ 为本体中的概念; $\text{NCA}(a,b)$ 是 $a$ 和 $b$ 的最近共同祖先概念. $IC_{\text{res}}(a)$, $IC_{\text{res}}(b)$ 是由式 (3) 和式 (1) 得到的概念 $a,b$ 的信息量; $IC_{\text{res}}(\text{NCA}(a,b))$ 是 $a,b$ 的共享信息量.

$\delta$ 是一个大于 0 的实数. 加入 $\delta$ 旨在避免分母等于 0 的情况, 从而有效区分最近共同祖先是根节点的概念相似度. 当 $sim_{IC}(a,b)$ 中的其他参数不变时, 其值随 $\delta$ 增大而增大. 如果 $\delta$ 过大, 会极大地削弱信息量本身对相似度的影响, 从而降低区分概念相似程度的能力. 因此 $\delta$ 的取值最好在 $[0,2\max(IC_{\text{res}}(a))]$ 之间.

### 2.2　基于边计算的概念相似度

基于边计算的概念语义相似性度量方法用概念间最短距离来衡量它们的语义相似性. 该类方法认为距离越近的概念间语义相似程度越高, 却忽略了概念以及它们 NCA 概念的深度所产生的影响, 从而导致概念间相似程度区分不细致. 我们认为概念间相似度与它们相同抽象程度和差异抽象程度相关, 通常可用概念的深度表示其抽象程度.

文中用两概念 NCA 的深度表示它们相同抽象程度, 用概念间最短距离来度量它们差异抽象程度. NCA 的深度越大, 它们相同抽象越具体, 相似程度越高; 概念间最短距离越大, 它们差异抽象程度越大, 相似度越小. 基于边的相似度如下所示:

$$sim_{sd}(a,b)=\frac{d(\text{NCA}(a,b))}{d(a)+d(b)-d(\text{NCA}(a,b))} \tag{5}$$

其中, $d(a)$, $d(b)$ 分别是 $a,b$ 的深度; $d(\text{NCA}(a,b))$ 为最近共同祖先概念的深度.

### 2.3　加权概念语义相似度

由于基于节点信息量的方法 $sim_{IC}$ 无法区分信息量相同、深度不同的概念间相似度, 而基于边计算的方法 $sim_{sd}$ 又无法区分深度相同、密度不同的概念间相似度. 因此, 为了弥补单个方法存在的不足, 将上述两种方法线性加权, 提出加权的本体概念语义相似度计算方法 (WSim_OC), 如下所示:

$$\begin{aligned} sim(a,b)=&\alpha sim_{IC}(a,b)+\\ &(1-\alpha)sim_{sd}(a,b) \end{aligned} \tag{6}$$

其中, $\alpha$ 为 $sim_{IC}$ 的影响因子, 调整 $\alpha$ 可改变概念信息量和深度对相似度的影响, 有利于更合理地度量相似性 (详见本文实验分析). 当 $sim_{IC}>sim_{sd}$ 时, 相似度随 $\alpha$ 增大而增大; 反之, 相似度则随 $\alpha$ 增大而减小. 通过对大量本体分析发现, 通常概念相似度之间的差异随 $\alpha$ 增大而减小, 随 $\alpha$ 的减小而增大.

WSim_OC 方法满足概念相似度的特性:

1)　$sim(a,b)\in[0,1]$: 概念相似度在 [0, 1] 之间;

2)　$\forall c, sim(c,c)=1$: 概念自身相似度为 1;

3)　$\forall a,b, sim(a,b)=sim(b,a)$: 概念相似度的对称性.

通常对其父概念概率贡献越大的概念与其父概念所共享的信息量越高, 相似度越高. 而概念对其父概念概率的贡献与它所覆盖叶子的数量成正比. 因

此, 概念与其子概念的相似程度随着该子概念所覆盖叶子数目的增大而增大. 此外, 由于兄弟是通过它们父概念关联的, 所以父子概念相似度越高的两个兄弟概念的相似度越高.

假设概念 $a, b, c$ 在相同本体中, 且 $a, b$ 为 $c$ 的子概念. 其中, $c$ 的深度为 $d$. $c$ 所覆盖的子树含有 $n$ 个叶子, 以 $a, b$ 为根的子树分别含有 $m, z$ 个叶子, 并且叶子的概率为 $p$.

由于 $a, b$ 的 NCA 为 $c$, 则 $d(a) = d(b) = d + 1$. 则由式 (5) 可得 $a, c$ 相似度中的 $sim_{sd}$ 分量值为

$$sim_{sd}(a, c) = d(d + (d + 1) - d)^{-1} = (1 + d^{-1})^{-1}$$

同理, $sim_{sd}(a, b) = (1 + 2d^{-1})^{-1}$.

由此可知, 父子相似度中的 $sim_{sd}$ 分量只与父概念深度相关, 所以具有相同父概念的所有父子相似度中的 $sim_{sd}$ 分量相等, 即 $sim_{sd}(a, c) = sim_{sd}(b, c)$; 兄弟相似度中的 $sim_{sd}$ 也只与它们父概念的深度相关, 因此互为兄弟关系的任意两个概念相似度中的该分量相等. 因此, $sim_{sd}$ 无法细致区分父子、兄弟间的相似度.

由式 (4) 得 $a, c$ 相似度中的 $sim_{IC}$ 分量为

$$sim_{IC}(a, c) = \frac{2IC_{\text{res}}(c) + \delta}{IC_{\text{res}}(a) + IC_{\text{res}}(c) + \delta}$$

由式 (3) 得: $p(a) = mp$, $p(b) = zp$, $p(c) = np$. 根据式 (1) 得:

$$sim_{IC}(a, c) = \frac{-2\log p(c) + \delta}{-\log p(c) - \log p(a) + \delta} = \frac{-2\log n + \delta - 2\log p}{-\log nm + \delta - 2\log p}$$

当 $n$ 一定时, $sim_{sd}$ 随 $m$ 的增大而增大 (分子、分母必定大于 0, 且分子不变, 分母减小); 当 $m$ 一定时, 函数随 $n$ 的增大而减小 (证明略). 所以, $n$ 不变时, 若 $m > z$ 则 $sim_{IC}(a, c) > sim_{IC}(b, c)$. 即具有相同父概念的父子概念相似度的 $sim_{IC}$ 分量, 随着以该子概念所覆盖的叶子数目的增大而增大. 同理, $a, b$ 相似度的 $sim_{IC}$ 分量为

$$sim_{IC}(a, b) = \frac{-2\log n + \delta - 2\log p}{-\log zm + \delta - 2\log p}$$

当 $n$ 一定时, $sim_{IC}$ 随 $zm$ 的增大而增大, 所以父子概念间相似度越高的两个兄弟概念相似度越高.

综上分析, 可知 WSim_OC 方法不仅满足概念相似度的公认特性, 还可证明该方法能够按上述常识合理区分父子概念、兄弟概念间的相似度.

算法 2 列出了 WSim_OC 方法的具体步骤.

**算法 2.** WSim_OC **方法**

**输入.** 两个概念;
**输出.** 概念间的语义相似度;
**步骤 1.** 遍历本体树, 找到这两个概念;
**步骤 2.** 获取概念的深度、信息量;
**步骤 3.** 查找它们的最近共同祖先概念 NCA;
**步骤 4.** 获取其 NCA 的深度及信息量;
**步骤 5.** 依式 (6) 计算它们的语义相似度.

## 3 实验

本文以图 1 所示汽车本体片段为例, 计算该本体中概念间的相似度. 图中节点由概念和标识 $c_i$ 表示. 针对文中所提方法, 分别设计两组实验. 第一组实验将 B-U 概念概率计算方法与 T-D-A[6] 方法进行比较. 第二组实验将 WSim_OC 方法分别与文献 [6] (本文简称 WESD) 和文献 [7] (简称 EXSim) 中的方法进行比较, 验证 WSim_OC 方法的合理性. 下述图中, 曲线标识分别对应所采用方法的名称; Sim 表示相似度; Dis 则表示语义距离.

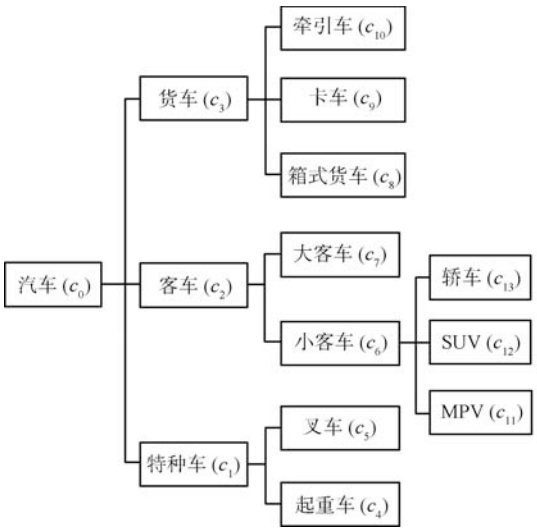



图 1 汽车本体片段示例图

Fig. 1 Example of automobile ontology fragment

### 3.1 实验一

表 1 为 T-D-A 和 B-U 所得概念概率. 结果表明 T-D-A 使得概念的各子概念概率相等, 如客车的概率与特种车的概率相等. 然而该本体中, 客车对汽车的贡献大于特种车的贡献, 即客车概率应大于特种车概率; 同理, 小客车概率应大于大客车概率. 因此 B-U 方法能更合理地计算概念概率. 该实验表明 B-U 方法为概念的各子概念分配不同的概率, 为细致区分父子、兄弟概念相似度奠定了基础.

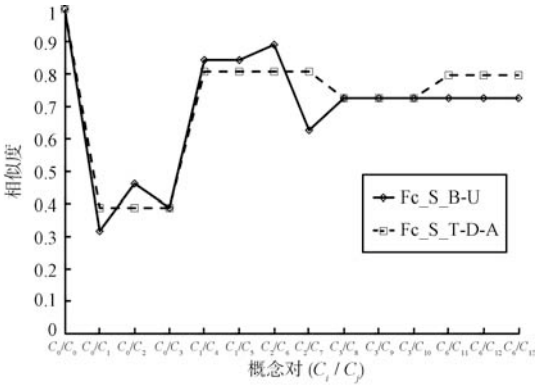
进一步将上述两类概念概率所得概念信息量分别用于 $sim_{IC}$ 方法中, 通过所得概念相似度来比较这两种概率计算方法的性能. 式 (4) 中 $\delta$ 取 1. 图

2 (a)、图 2 (b) 所示分别为父子、兄弟概念相似度. 由 P_T-D-A 所得相似度中, 概念与其所有子概念的相似度相等, 任意两兄弟的相似度相等. 如汽车与客车的相似度等于其与特种车的相似度, 货车与客车的相似度等于其与特种车的相似度. 然而, 我们认为汽车与客车的相似性比它与特种车的相似度高, 货车和客车的相似性比特种车的高. 显然, P_B-U 所得相似度更加合理.

表 1　概念概率表

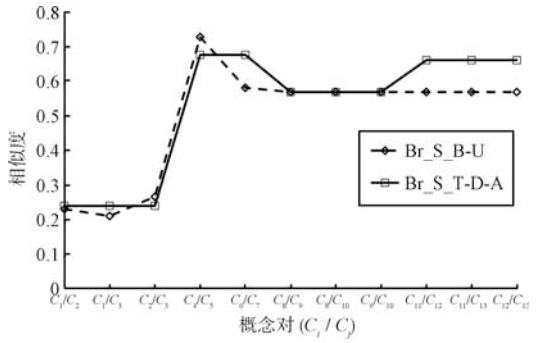
Table 1　Concept probabilities

| 同深度概念 | P_B-U | P_T-D-A |
|---|---|---|
| $c_0$ | 1 | 1 |
| $c_1, c_2, c_3$ | $p(c_1)=0.222,\ p(c_2)=0.444$ $p(c_3)=0.333$ | 0.3333 |
| $c_4, c_5$ | 0.1111 | 0.1667 |
| $c_6, c_7$ | $p(c_6)=0.333,\ p(c_7)=0.1111$ | 0.1667 |
| $c_8, c_9, c_{10}$ | 0.1111 | 0.1111 |
| $c_{11}, c_{12}, c_{13}$ | 0.1111 | 0.0556 |




(a) 父子概念相似度

(a) Similarities between concepts and their children




(b) 兄弟概念相似度

(b) Similarities between siblings

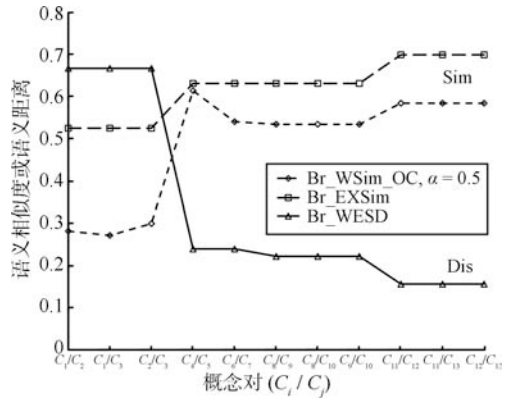
图 2　$sim_{IC}$ 所得父子、兄弟概念间相似度

Fig. 2　Similarities between concepts and their children, or between siblings by using $sim_{IC}$

## 3.2　实验二

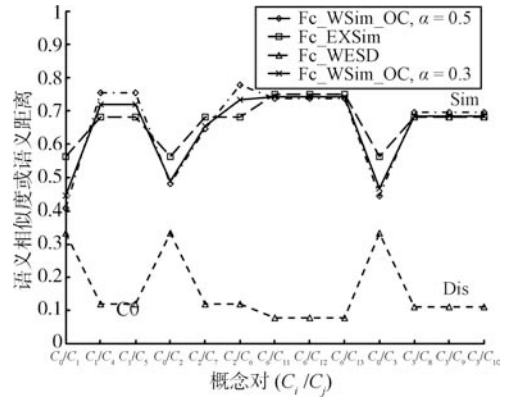
WESD[6] 与 EXSim[7] 是近年来用于 Web 服务匹配的两种语义相似度方法. 其中, WESD 首先计算边的连接强度, 然后利用最短加权距离来度量它们的相似程度. EXSim 则通过一个指数函数来度量概念间相似性. 本实验将 WSim_OC 与上述两种方法进行比较, 参数 $\delta$ 取 1; 权值 $\alpha = 0.5$. 根据概念结构关系, 可分为父子、兄弟以及位于不同分支的三种概念对, 分别讨论三种方法的性能. 图 3 (a)、图 3 (b) 分别为兄弟、父子概念间的相似度或语义距离.

图 3 (a) 表明 WSim_OC 方法使得父子相似度高的兄弟相似度高, 如 $sim(c_2, c_3) > sim(c_2, c_1)$. 表 2 为图 3 (b) 中父子相似度之间的关系. WESD 和 EXSim 使得位于同一分支上的父子间相似度 (语义距离) 随着深度的加深而增大 (减小), 如 $sim(c_6, c_{13}) > sim(c_2, c_6)$, $sim(c_2, c_6) > sim(c_0, c_2)$ 等; 虽然 WSim_OC 存在随深度的加深父子相似度减小的情况, 如 $sim(c_6, c_{13}) < sim(c_2, c_6)$. 这是由于 $c_6$ 对 $c_2$ 的贡献非常大, 而 $c_{13}$ 对 $c_6$ 的贡献相对一般, 也是合理的. 如果出现这种情况, 可通过调整 $\alpha$ 得到改善. 如图 3 (b) 所示, 当 $\alpha = 0.3$ 时, $sim(c_6, c_{13}) > sim(c_2, c_6)$.




(a) 兄弟概念

(a) Siblings




(b) 父子概念

(b) Concepts and their children

图 3　兄弟、父子概念间相似度或语义距离

Fig. 3　Similarities or distance between concepts and their children or between siblings

表 2 父子概念相似度之间的关系表

Table 2 Relationships of similarities between concepts and their children

| 父子概念相似度关系 | Fc_WESD_Dis | Fc_EXSim_Sim | Fc_WSim_OC_Sim, $\alpha = 0.5$ |
|---|---|---|---|
| 同一分支上的父子概念 | $(c_0, c_2) > (c_2, c_6) > (c_6, c_{11})$ | $(c_6, c_{11}) > (c_2, c_6) > (c_0, c_2)$ | $(c_2, c_6) > (c_6, c_{11}) > (c_0, c_2)$ |
| | $(c_0, c_1) > (c_1, c_4)$ | $(c_1, c_4) > (c_0, c_1)$ | $(c_1, c_4) > (c_0, c_1)$ |
| | $(c_0, c_3) > (c_3, c_8)$ | $(c_3, c_8) > (c_0, c_3)$ | $(c_3, c_8) > (c_0, c_3)$ |
| 同父不同子的父子概念 | $(c_0, c_1) = (c_0, c_2) = (c_0, c_3)$ | $(c_0, c_1) = (c_0, c_2) = (c_0, c_3)$ | $(c_0, c_2) > (c_0, c_3) > (c_0, c_1)$ |
| | $(c_1, c_4) = (c_1, c_5)$ | $(c_1, c_4) = (c_1, c_5) = (c_2, c_6)$ | $(c_1, c_4) = (c_1, c_5)$ |
| | $(c_2, c_6) = (c_2, c_7)$ | $(c_2, c_6) = (c_2, c_7)$ | $(c_2, c_6) > (c_2, c_7)$ |
| | $(c_3, c_8) = (c_3, c_9) = (c_3, c_{10})$ | $(c_3, c_8) = (c_3, c_9) = (c_3, c_{10})$ | $(c_3, c_8) = (c_3, c_9) = (c_3, c_{10})$ |
| | $(c_6, c_{11}) = (c_6, c_{12}) = (c_6, c_{13})$ | $(c_6, c_{11}) = (c_6, c_{12}) = (c_6, c_{13})$ | $(c_6, c_{11}) = (c_6, c_{12}) = (c_6, c_{13})$ |
| 同密度不同深度 | $(c_3, c_8) > (c_6, c_{11})$ | $(c_6, c_{11}) > (c_3, c_8)$ | $(c_6, c_{11}) > (c_3, c_8)$ |
| 同深度父子概念 | $(c_2, c_6) = (c_2, c_7) =$ $(c_1, c_4) > (c_3, c_8)$ | $(c_2, c_6) = (c_1, c_4) =$ $(c_3, c_8) = (c_2, c_7)$ | $(c_2, c_6) > (c_1, c_4) >$ $(c_3, c_8) > (c_2, c_7)$ |

对于深度相同的父子概念, EXSim 方法所得父子概念间相似度相等, 如 $sim(c_2, c_6) = sim(c_1, c_4) = sim(c_3, c_8)$. 而 WESD 使得 $sim(c_3, c_8) > sim(c_1, c_4)$. 然而在概念深度相同时, 父子概念间相似度应随概念密度增大而减小, 即 $sim(c_3, c_8) < sim(c_1, c_4)$. 显然 WSim_OC 能够合理区分父概念深度相同而密度不同的父子间相似度.

同父不同子的父子相似度间的关系表明 WESD 和 EXSim 使得概念与其所有子概念的相似度相等, 如 $sim(c_0, c_1) = sim(c_0, c_2)$; 然而 $c_2$ 对 $c_0$ 的贡献远高于 $c_1$, 所以 $sim(c_0, c_2)$ 应大于 $sim(c_0, c_1)$. 然而 WSim_OC 方法可依据子概念对其父概念的贡献来区分父子间相似度, 使得 $sim(c_0, c_2) > sim(c_0, c_1)$.

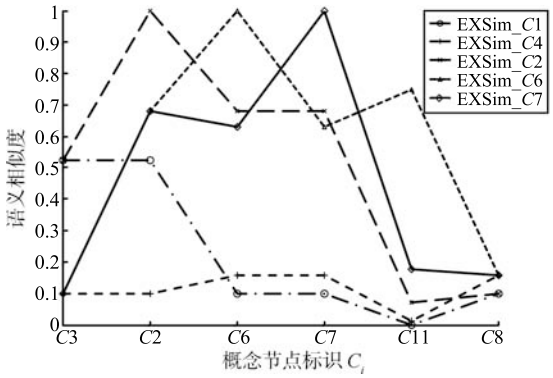



图 4 EXSim[7] 所得不再同一分支的概念间相似度

Fig. 4 Similarities between concepts at different branches by EXSim

对于两个不在同一个分支上的概念, 如 $(c_4, c_8), (c_1, c_8)$ 等, EXSim 存在不合理的度量: 当两个概念的最近共同祖先及其中一个概念不变时, 两概念的相似度随另一概念深度的加深而增大, 如图 4 所示: $sim(c_4, c_8) > sim(c_1, c_8)$. 图中曲线标识 EXSim_$C_i$ 表示概念 $c_i$ 与各概念间的相似度.

## 4 结论

本文以本体结构为依据, 提出了一种合理的概念概率计算方法. 并进一步利用概念的深度、信息量提出了加权概念相似性度量方法. 结果表明, 该方法能够合理计算概念概率, 细致区分概念间的相似度. 在后续的研究工作中, 我们将探索此类方法在分布式信息检索和 Web 服务发现中的性能.

## References

1 Liu Yu-Peng, Li Sheng, Zhao Tie-Jun. System combination based on WSD using WordNet. *Acta Automatica Sinica*, 2010, **36**(11): 1575−1580
(刘宇鹏, 李生, 赵铁军. 基于 WordNet 词义消歧的系统融合. 自动化学报, 2010, **36**(11): 1575−1580)

2 Atkinson J, Ferreira A, Aravena E. Discovering implicit intention-level knowledge from natural-language texts. *Knowledge-Based Systems*, 2009, **22**(7): 502−508

3 Sánchez D D, Isern D, Millan M. Content annotation for the semantic web: an automatic web-based approach. *Knowledge and Information Systems*, 2011, **27**(3): 393−418

4 Sánchez D D. A methodology to learn ontological attributes from the web. *Data and Knowledge Engineering*, 2010, **69**(6): 573−597

5 Gaeta M, Orciuoli F, Ritrovato P. Advanced ontology management system for personalised e-learning. *Knowledge-Based Systems*, 2009, **22**(4): 292−301

6 Bai Dong-Wei. Research on Web Services Semantic Matchmaking and Discovery [Ph. D. dissertation], Beijing University of Posts and Telecommunication, China, 2007
(白东伟. 基于语义的 Web 服务匹配与发现技术研究 [博士学位论文], 北京邮电大学, 中国, 2007)

7 Qiu Tian, Li Peng-Fei, Lin Pin. A web service matching algorithm based on semantic similarity of concepts. *Acta Electronica Sinica*, 2009, **37**(2): 429−432
(邱田, 李鹏飞, 林品. 一个基于概念语义近似度的 Web 服务匹配算法. 电子学报, 2009, **37**(2): 429−432)

8 Sánchez D, Batet M, Valls A, Gibert K. Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems*, 2010, **35**(3): 383−413

9 Etzioni O, Cafarella M, Downey D, Popescu A M, Shaked T, Soderland S, Weld D S, Yates A. Unsupervised named-entity extraction form the web: an experimental study. *Artificial Intelligence*, 2005, **165**(1): 91−134

10 Lemaire B, Denhiere G. Effects of high-order co-occurrences on word semantic similarities [Online], available: http://cpl.revues.org/document471.html, December 9, 2011

11 Bollegala D, Matsuo Y, Ishizuka M. Measuring semantic similarity between words using web search engines. In: Proceedings of the 16th International Conference on World Wide Web. Banff, Canada: ACM, 2007. 757−766

12 Tversky A. Features of similarity. *Psychological Review*, 1977, **84**(2): 327−352

13 Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 1989, **19**(1): 17−30

14 Wu Z B, Palmer M. Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Las Cruces, USA: ACL, 1994. 133−138

15 Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press, 1998. 265−283

16 Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal, Canada: Morgan Kaufmann, 1995. 448−453

17 Lin D. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Madison, USA: Morgan Kaufmann, 1998. 296−304

18 Yang Li, Zuo Chun, Wang Yu-Guo. K-nearest neighbor classification based on semantic distance. *Journal of Software*, 2005, **16**(12): 2054−2062
(杨立, 左春, 王裕国. 基于语义距离的 K 最近邻分类方法. 软件学报, 2005, **16**(12): 2054−2062)

19 Pirro G. A semantic similarity metric combining features and intrinsic information content. *Data and Knowledge Engineering*, 2009, **68**(11): 1289−1308

20 Richardson R, Smeaton A F. Using WordNet in a Knowledge-Based Approach to Information Retrieval, Technical Report Working Paper CA-0395, School of Computer Applications, Dublin City University, Ireland, 1995

21 Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in WordNet. In: Proceedings of 16th European Conference on Artificial Intelligence, including Prestigious Applicants of Intelligent Systems. Valencia, Spain: IOS Press, 2004. 1089−1090



**李文清** 北京理工大学计算机学院博士研究生. 2000 年获北京理工大学计算机系学士学位. 主要研究方向为分布式 Web 服务发现, 信息检索.
E-mail: yeaphon_li@126.com
(**LI Wen-Qing** Ph. D. candidate at the School of Computer Science and Technology, Beijing Institute of Technology. She received her bachelor degree from Beijing Institute of Technology in 2000. Her research interest covers distributed web service discovery and information retrieval.)



**孙 新** 北京理工大学计算机学院讲师. 主要研究方向为对等网络, 网格计算和分布式系统. 本文通信作者.
E-mail: sunxin@bit.edu.cn
(**SUN Xin** Lecturer at the School of Computer Science and Technology, Beijing Institute of Technology. Her research interest covers P2P networks, grid computing, and distributed systems. Corresponding author of this paper.)



**张常有** 石家庄铁道大学信息科学技术学院教授. 主要研究方向为并行与分布式计算, 计算机支持协同工作.
E-mail: zhangchangyou@tsinghua.org.cn
(**ZHANG Chang-You** Professor at the School of Information Science and Technology, Shijiazhuang Tiedao University. His research interest covers parellel and distributed computing, and network and information security.)



**冯 烨** 北京控制工程研究所工程师. 2005 年获北京控制工程研究所工学硕士. 主要研究方向为制导、导航与控制.
E-mail: yeaphon@sohu.com
(**FENG Ye** Engineer at Beijing Institute of Control Engineering. He received his master degree from Beijing Institute of Control Engineering in 2005. His research interest covers guidance, navigation, and control.)