本体语义相似度自适应综合加权算法研究

郑志蕴 阮春阳 李 伦 李 钝
(郑州大学信息工程学院 郑州 450001)

摘 要 本体语义相似度计算是解决语义网中语义异构的关键环节。通过对传统语义相似度计算方法的分析研究,引入本体层次结构,给出基于信息内容、距离、属性的语义相似度改进计算方法,并采用主成分分析法,提出一种自适应相似度综合加权计算方法(ACWA),以解决传统综合加权计算时人工赋权的不足。实验结果表明,提出的ACWA算法的计算结果与参照标准之间的皮尔森系数较传统算法平均高出了8.1%,有效提升了本体语义相似度计算的准确性。

关键词 本体,语义相似度,主成分分析法
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.046

Adaptive Ontology Semantic Similarity Comprehensive Weighted Algorithm

ZHENG Zhi-yun RUAN Chun-yang LI Lun LI Dun
(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract Ontology semantic similarity computation is the key to solve the semantic heterogeneity in semantic Web. Through analysis and study of the traditional ontology semantic similarity computation, this article introduced the ontology hierarchy, proposed an improved semantic similarity method, which is based on information content, distance, and attribute, and put forward a ACWA using the principal component analysis, to address the deficiencies of artificial weight in traditional comprehensive weighted calculation. The experimental results show that the pearson cofficient of the proposed ACWA algorithm results compared with the reference value is 8.1% higher than that of the traditional method, and the accuracy of ontology semantic similarity calculation is effectively increased.

Keywords Ontology, Semantic similarity, Principal component analysis

当前语义相似度计算已广泛应用于本体学习与合并、语义标注、知识管理中的信息抽取及自然语言理解等相关领域。随着本体日益广泛的应用,各领域都出现了大量的本体,然而本体的建立却没有统一的规范约束,因此产生了语义异构等许多问题。为了解决语义异构问题并把这些领域本体进行合并,本体映射(Ontology Mapping)的研究应用而生。本体映射是本体集成(Ontology Integration)、本体串联(Ontology Alignment)、本体合并(Ontology Merging)等的技术基础,是解决知识共享与重用的有效途径[1],而本体概念间的语义相似度计算是本体映射的关键技术之一,已成为当今信息技术研究的一个热点[2]。目前本体概念的语义相似度计算方法主要有基于信息内容的计算、基于距离的计算、基于属性的计算及混合3种方法的综合加权计算。本文在对传统概念语义相似度计算方法进行分析研究的基础上,提出一种以自适应综合加权为核心的语义相似度计算方法。该方法综合考虑概念在本体树中的节点密度信息、深度信息、公共父节点信息,采用改进的基于信息内容的概念语义相似度算法,结合基于距离和属性的改进方法计算,利用主成分分析法对3个因素计算出的相似度进行动态加权线性求和。实验表明,与传统算法相比,本文提出的ACWA算法的计算结果与参照标准值之间的皮尔森系数比传统算法平均高出8.1%。

1 相关工作

目前国内外已有大量研究者对本体概念语义相似度的计算进行了研究,并形成了许多成熟的相似度计算方法。主要包括3种计算方法。

1.1 基于信息内容的计算方法

基于信息内容的计算方法通过一个给定本体中两个概念之间的共享信息内容来确定概念之间的相似度[3]。本体结构中的子节点表示的概念通常都是对其祖先节点表示概念的细化,所以一个节点的内容能够代表其所有祖先节点的信息内容,最近邻公共祖先节点代表的是在本体树结构中距离两个概念最近的共同祖先节点,可以采用两个概念最近公共祖先节点的信息内容对其语义相似度进行量化,得到本体树中任

到稿日期:2015-08-07 返修日期:2015-11-07 本文受河南省国际科技合作项目(144300510007),郑州市科技攻关计划项目(141PPTGG368)资助。

郑志蕴(1962—),女,博士,教授,CCF会员,主要研究方向为分布式计算、智能信息处理;阮春阳(1991—),男,硕士,主要研究方向为语义网、本体;李 伦(1978—),男,博士,副教授,主要研究方向为云计算;李 钝(1975—),女,博士,讲师,主要研究方向为信息检索、数据挖掘,E-mail:ielidun@zzu.edu.cn(通信作者)。

意两个概念之间的语义相似度计算[4]，如式(1)所示：

$$Sim(a,b)_{ic}=\frac{2IC(Lcan(a,b))}{IC(a)+IC(b)} \tag{1}$$

其中，$Lcan(a,b)$表示概念 a 和 b 在本体树中的最近邻公共祖先节点，$IC(a)$和 $IC(b)$表示概念 a 和概念 b 所拥有的信息内容，也称信息量，这种信息的量化表示提供了一种测量语义相似性的方法，信息量的计算公式如式(2)所示：

$$IC(c)=-\log\frac{N(c)}{N} \tag{2}$$

其中，$N(c)$为概念 c 在训练样本中出现的次数，N 为训练样本的总数。文献[4]用式(2)计算概念节点的信息量，这是传统的基于语料库的方法，计算结果依赖于语料库训练样本的个数，对于不同的语料库则有可能得到不同的 $IC(c)$。这是因为语料库中的概念是有限的，不同的语料库概念的数量也不同，出现的频率也不一定相同，所以这种利用语料库的概念节点信息量计算方法存在不足之处。经过对本体树的分析，本文认为影响概念信息内容及概念间相似度的因素有如下几方面。

(1)被比较概念在本体树中的深度。根据 Resnik[5] 和 Seco[6] 理论，概念深度越小，说明出现频率越高，则该概念越抽象，所涵盖的信息内容也就越少。反之，底层概念越具体，所继承的信息内容也越多，概念间所共享的上层信息概率越大，因此底层概念间的语义相似度一般大于高层概念间的相似度。如 IC(风景名胜)<IC(少林寺)，Sim(风景名胜，休闲场所)<Sim(少林寺，龙门石窟)。

(2)被比较概念在本体树中的密度。局部区域的密度说明此区域对概念的划分越具体，信息内容也就越大，所共享信息的概率也就越大，区域内概念词间的语义相似度相对较大。如图 1 所示，Sim(休闲场所，景点)<Sim(风景名胜，景点)。

(3)被比较概念在本体树中相隔的路径长度。在密度及路径类型相同的情况下，概念间路径长度越长，相似度越小。如图 1 所示，Sim(少林寺，景点)<Sim(风景名胜，少林寺)。

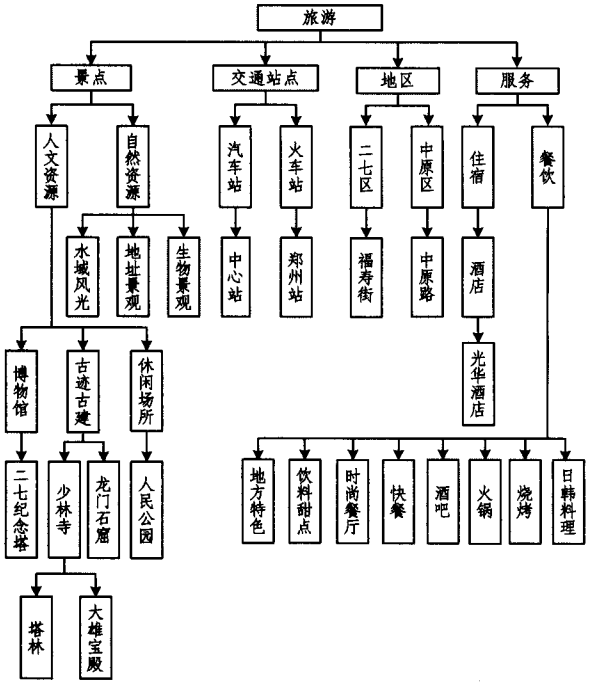

图 1　旅游本体树局部结构

1.2　基于距离的计算方法

基于距离的计算方法通过本体树中两个概念词的几何距离来计算它们之间的语义距离，计算模型如式(3)所示：

$$Sim(a,b)=\frac{1}{Dis(a,b)} \tag{3}$$

其中，$Dis(a,b)$表示概念之间的最短距离。WuAndPalmer 算法[7]与 LeacockAndChodorow 算法[8]是基于语义距离的两种经典算法。WuAndPalmer 算法通过与概念词最近的公共父结点概念词的位置关系来计算其相似度；LeacockAndChodorow 算法则是将两概念间的路径长度转化为信息量来进行相似度计算。文献[7]和文献[8]都忽略了边的关系类型对计算的影响，不同的关系类型对路径的影响也不相同。目前，基于距离的相似度计算研究中考虑边的类型的方法较多，但大部分没有将边的类型来作为计算的主要因素进行研究。

1.3　基于属性的计算方法

本体概念通过属性来表明概念特征，基于属性的计算方法通过统计概念所具有的公共属性的个数来获得概念的相似度。概念的相似度与概念拥有的公共属性个数成正比。Tversky[9] 提出基于属性最经典的语义相似度计算方法，计算模型如式(4)所示。

$$Sim(a,b)=\alpha\times Properties(a\cap b)-\beta\times Properties(a-b)-\gamma\times Properties(b-a) \tag{4}$$

其中，$Properties(a\cap b)$表示概念 a 和 b 所具备的公共属性集合，$Properties(a-b)$表示概念 a 具备而概念 b 不具备的属性集合，$Properties(b-a)$则表示概念 b 具备而 a 不具备的属性集合。基于属性的计算方法可以对人类认识和辨别现实生活中各种事物的过程进行模拟，但需要给出事物各个属性的详细信息。

随着语义相似度计算研究的深入，许多研究人员提出了一些混合算法，如文献[1]和文献[10]所述方法都是对上述算法的改进和综合，在一定程度上提高了计算结果的准确性。但综合加权过程中的权值是由本领域的专家确定的，存在主观性、不准确性而且对不同的本体不具有适应性。

2　自适应综合加权算法

本体概念语义相似度综合计算分别考虑多个因素计算概念间的语义相似度，并线性加权求和计算出最终的语义相似度。综合计算过程中权值的确定是关键的一环，权值确定的准确与否直接影响相似度计算的准确性，为了使权值确定方法适用于不同的本体，本文基于经济学领域中经典的主成分分析法，提出一种自适应的综合加权算法(Adaptive Comprehensive Weighted Algorithm，ACWA)。首先考虑本体概念的信息内容、概念距离和概念属性，分别计算出概念的相似度，然后用主成分分析法对计算的 3 个结果进行加权求和，计算出最终的相似度。

2.1　多因素相似度计算

为了计算结果更加全面准确，从多个因素出发，分别计算概念间的相似度。本文从本体概念信息量、距离和属性 3 个因素分析，给出相应的语义相似度计算公式。

(1)概念信息量

提出一种利用本体概念自身的信息量进行求解的方法，如式(5)所示：

$$IC(c)=\left(\frac{mNode(c)}{mNode(T)}+\frac{\log(D(c))}{\log(mD(T))}\right)*\left(1-\frac{\log(h(c)+1)}{\log(mNode(T))}\right) \tag{5}$$

其中，$mNode(c)$是以概念 c 节点所在的以根节点为直接父结点的子树所拥有的概念节点的总个数，$mNode(T)$是概念 c 所在本体树 T 所有概念节点的个数，$h(c)$是概念 c 的所有子节点个数，$D(c)$是概念 c 的深度，$mD(T)$ 是本体树 T 的最大深度。求出概念的信息量后再使用式(1)计算两个概念的相似度。

(2)概念间距离

基于距离的传统相似度计算方法忽略了边的关系类型对计算的影响。实际上，不同的关系类型对路径的影响不相同，本文将边的关系类型作为权重加入到计算过程中，针对本体树中 3 种主要概念关系，定义对应边的权重如式(6)所示：

$$Wedge(a,b)=\begin{cases}0.9, & \text{is } A\\ 0.5, & \text{partof}\\ 0.1, & \text{other}\end{cases} \tag{6}$$

结合 WuAndPalmer 算法，提出一种改进计算模型如式(7)所示：

$$Sim(a,b)_{dis}=\frac{2*sPath(c,r)+sPath(a,b)}{sPath(a,b)+sPath(a,c)+sPath(b,c)+2*sPath(c,r)} \tag{7}$$

$$sPath(a_1,a_n)=\sum_{i=1}^{n}wedge_i*path(a_i,a_{i+1}) \tag{8}$$

其中，$wedge$ 为相邻概念间边的权重，$path$ 为相邻概念间直接距离，$sPath(a,b)$表示从概念 a 到概念 b 的加权最短路径，$sPath(a,c)$和 $sPath(b,c)$分别表示概念 a 和 b 到最近公共节点 c 的最短路径，$sPath(c,r)$表示概念 c 到根节点 r 的最短路径。

(3)概念属性

1.3 节提到的 Tversky[9] 计算方法仅采用概念对共有属性个数度量相似度，即共有的属性越多相似性就越高。实际上，属性的结构信息对概念相似度计算也有影响，例如父节点拥有的属性子节点一定拥有，子节点拥有的属性父节点不一定拥有。结合属性结构信息，提出一种改进的计算方法，如式(9)所示：

$$Sim(a,b)_{pro}=\frac{Properties(a\cap b)}{Properties(a\cap b)+\alpha*Properties(a-b)+\beta*Properties(b-a)} \tag{9}$$

$$\alpha=\begin{cases}\dfrac{d(a)}{d(a)+d(b)}, & a\leqslant b\\ 1-\dfrac{d(a)}{d(a)+d(b)}, & a>b\end{cases},\alpha+\beta=1 \tag{10}$$

其中，$Properties(a\cap b)$表示概念 a 和 b 所具备的公共属性集合，$Properties(a-b)$表示概念 a 具备而概念 b 不具备的属性集合，$Properties(b-a)$则表示概念 b 具备而 a 不具备的属性集合。$d(a)$和 $d(b)$分别为概念 a 和 b 在本体层次结构中的深度。

2.2 基于主成分分析法的动态权值计算方法

PCA 的思想是在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法。通常转化生成的综合指标称为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关，这就使得主成分比原始变量具有某些更优越的性能。PCA 根据主成分的贡献率来分配各主成分的权重，而不是人为确定的，从而克服了多因素分析中人为确定权值的缺陷，保证结果客观、合理、准确。

PCA 计算模型如下：

(1)为了消除不同变量的量纲的影响，首先需要对变量进行标准化。设检测数据样本共有 n 个，指标共有 p 个，分别设变量 x_i,x_j,x_p，令 $X_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,p)$为第 i 个样本第 j 个指标的值。作变换：

$$Y_j=\frac{X_j-E(X_j)}{\sqrt{Var(X_j)}}\ (i,j=1,2,\cdots,p) \tag{11}$$

得到标准化数据矩阵 $y_{ij}=\dfrac{X_{ij}-\overline{X_j}}{s_j}$，其中$\bar{x}_j=\dfrac{1}{n}\sum_{i=1}^{i}x_{ij}$，$s_j^2=\dfrac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$

(2)在标准化数据矩阵 $Y=(y_{ij})n*p$ 的基础上计算 p 个原始指标相关系数矩阵：

$$R=(r_{ij})_{p\times p}=\begin{bmatrix}r_{11} & r_{12} & \cdots & r_{1p}\\ r_{21} & r_{22} & \cdots & r_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ r_{p1} & r_{p2} & \cdots & r_{pp}\end{bmatrix}$$

其中 $r_{ij}=\dfrac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}\ (i,j=1,2,\cdots,p)$

(3)求相关系数矩阵 R 的特征值并排序 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$，再求出 R 的特征值相应的正则化特征向量 $e_i=(e_{i1},e_{i2},\cdots,e_{ip})$，则第 i 个主成分表示为各指标 X_k 的组合 $Z_i=\sum_{i=1}^{p}e_{ik}\cdot X_k$。

(4)计算累积贡献率确定主成分的数目。主成分 Z_i 的贡献率为 $w_i=\dfrac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}(i=1,2,\cdots,p)$，累计贡献率为 $\dfrac{\sum_{k=1}^{i}\lambda^k}{\sum_{k=1}^{p}\lambda_k}(i,j=1,2,\cdots,p)$，一般取累计贡献率达 85%～95%的特征值 λ_1，$\lambda_2,\cdots,\lambda_m$ 对应第 1,第 2,…,第 $m(m\leqslant p)$个主成分。

(5)计算主成分载荷，确定综合得分。当主成分之间不相关时，主成分载荷是主成分和各指标的相关系数，相关系数越大，说明主成分对该指标变量的代表性越好，计算公式为：

$$l_{ij}=p(z_i,x_j)=\sqrt{\lambda_i}e_{ij}\ (i,j=1,2,\cdots,p) \tag{12}$$

(6)确定各主成分得分，确定综合评分函数。得到各主成分的载荷以后，可以计算各主成分的得分：

$$\begin{cases}z_1=l_{11}x_1+l_{12}x_2+\cdots+l_{1p}x_p\\ z_2=l_{21}x_1+l_{22}x_2+\cdots+l_{2p}x_p\\ \cdots\\ z_m=l_{m1}x_1+l_{m2}x_2+\cdots+l_{mp}x_p\end{cases} \tag{13}$$

$$Z=(z_{ij})_{n\times m}=\begin{bmatrix}z_{11} & z_{12} & \cdots & z_{1m}\\ z_{21} & z_{22} & \cdots & z_{2m}\\ \vdots & \vdots & \vdots & \vdots\\ z_{n1} & z_{n2} & \cdots & z_{nm}\end{bmatrix}$$，其中 z_{ij} 表示第 i 个样本第 j 个主成分得分，则第 i 个样本的综合得分 $f_i=\sum_{k=1}^{m}w_k\cdot z_{ik}(i=1,2,\cdots,n)$。

PCA 是一种降维的思想，本文采用主成分分析法动态计

算权值并不需要减少多因素的维度，而是利用主成分分析法计算出各因素的贡献率作为权值。原始主成分分析法是按累计贡献率大于设定阈值来确定主成分。本文提出的3个因素(信息量、距离、属性)都要作为主成分，可以忽略这个环节，提高算法的效率。ACWA中基于PCA的动态权值计算方法的主要思想如下：

(1)将3个因素计算出的3个相似度作为3个维度，通过多个样本的计算得到相似度矩阵作为输入样本矩阵；

(2)将样本矩阵矩阵标准化变换为标准矩阵 Z，并求出相关系数矩阵 R；

(3)解样本相关矩阵 R 的特征方程得3个特征根，确定主成分 λ_1，λ_2，λ_3；

(4)解方程组 $R*b=\lambda_i*b(i=1,2,3)$ 单位特征向量 b_i^0；

(5)将标准化后的指标变量转换为主成分 $U_i=Z_i^T*b_i^0$ $(i=1,2,3)$；

(6)对3个主成分进行加权求和，即得最终相似度值，权值为每个主成分的贡献率。

2.3 综合加权计算

综合加权的计算方法中涉及权值设定，以往在权值确定的过程中都是专家进行人工赋权，使结果具有主观不确定性，而且不同领域本体的权值确定还需要不同领域专家来确定，使得权值的计算不具有自适应性。本文考虑概念信息量、距离和属性3个因素分别计算出相似度后，对3个相似度用改进的主成分分析法进行动态加权线性求和。假设被比较概念对集合中有 m 对概念词，设 $X_i=(Sim_{ic(i)},Sim_{dis(i)},Sim_{pro(i)})$ 为主成分输入样本集合中的一个向量，其中每一维变量分别代表综合相似度计算模块中各部分语义相似度计算的结果，则概念相似度矩阵表示为 $X_{sim}=(x_{i1},x_{i2},x_{i3})^T(i=1,2,\cdots,m)$。对构建出来的相似度矩阵 X_{sim} 进行主成分分析，提取出的主成分为 $Y=(y_{sim1},y_{sim2},y_{sim3})$，各主成分的贡献率为 (r_1,r_2,r_3)，则最终的概念语义相似度计算公式为：

$$Sim_{total}=r_1*y_{sim1}+r_2*y_{sim2}+r_3*y_{sim3} \tag{14}$$

ACWA算法：

输入：节点概念 a,b,公共节点概念 c,根节点概念 r

输出：综合加权语义相似度 Sim_{total}

Begin

1. 计算节点概念信息量 IC(a)、IC(b)、IC(c)

2. 基于信息量的语义相似度计算：$Sim(a,b)_{ic}=\dfrac{2IC(Lcan(a,b))}{IC(a)+IC(b)}$

3. 确定节点概念之间边上权重：wedge

4. 计算节点概念之间最短加权距离：$sPath(a_1,a_n)=\sum_{i=1}^{n}wedge_i*path(a_i,a_{i+1})$

5. 基于距离的语义相似度计算：$Sim(a,b)_{dis}=\dfrac{2*sPath(c,r)+sPath(a,b)}{sPath(a,b)+sPath(a,c)+sPath(b,c)+2*sPath(c,r)}$

6. 确定影响因子：α，β

7. 基于属性的语义相似度计算：$Sim(a,b)_{pro}=\dfrac{Properties(a\cap b)}{Properties(a\cap b)+\alpha*Properties(a-b)+\beta*Properties(b-a)}$

8. PCA计算3个因素各自的动态权值：$PCA(Sim(a,b)_{dis},Sim(a,b)_{dis},Sim(a,b)_{pro})$

9. 综合加权计算：Sim_{total}

End

3 实验与结果分析

为了验证提出的自适应综合加权算法的优越性，本文先用改进的语义相似度算法从信息内容、距离和属性3个方面计算出本体概念的相似度，然后用改进的PCA计算出各个因素的权值并对其进行线性相加得到最终的相似度，并与经典算法进行比较。

3.1 实验数据

本文采用部分旅游本体作为实验数据，以验证算法的准确性和有效性。旅游本体是根据河南省旅游局数据通过protege生成的旅游本体，数据的统计信息如表1所列。

表1 数据统计信息表

本体	概念个数	属性个数	边类型数
旅游本体	41	33	2

3.2 实验设计

3.2.1 3组对比实验

为了验证提出算法的性能，设计了3组实验：

(1)改进3因素计算方法与传统算法的比较实验；

(2)改进PCA计算3个因素的权重实验；

(3)加权综合计算相似度与传统算法的比较实验。

3.2.2 实验步骤

(1)选取数据集中的概念对；

(2)从信息内容、距离和属性计算本体概念对的语义相似度并构造相似度矩阵；

(3)用相似度矩阵作为样本矩阵，实现改进的PCA算法，输入样本矩阵计算出各个因素的贡献率，并将其作为权值；

(4)对应相似度值加权并线性相加计算出最终的概念语义相似度；

(5)计算结果与经典的算法计算结果比较。

实验的核心流程如图2所示。

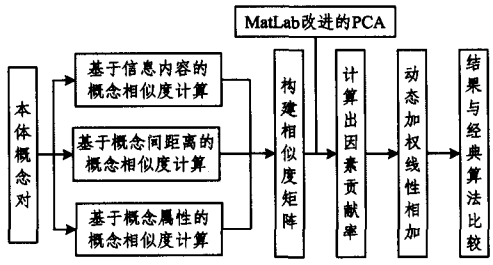

图2 实验核心流程示意图

3.3 结果分析

针对所提算法计算结果和传统算法计算结果的准确度进行比较分析。

1. 改进算法与传统算法对比

3种改进算法与传统算法的对比结果如表2所列。

表2中以人工判定值作为标准，它是由领域专家确定，具有权威性。计算值与人工判定标准比较是领域本体相似度计算中一种常用的比较分析方法。通过图3可以看出，基于本体层次结构改进的单因素的相似度计算结果更加接近人工判定标准，准确度总体上高于传统方法，其原因主要有：

(1)改进的信息量方法不需从样本文本中取样，消除了计算的随机性；

(2)改进的距离方法计算中考虑了边的关联类型,更具有区分度;

(3)改进的属性方法计算中考虑属性的同时还考虑了概念的上下位关系。

表 2　多因素相似度计算结果

本体概念对的相似性	3 因素本体概念语义相似度计算						
	信息内容		距离		属性		人工判定
	本文	Resnik	本文	Wu & Palmer	本文	Tversky	
Sim(人文资源,自然资源)	0.379	0.500	0.437	0.540	0.592	0.625	0.430
Sim(人文资源,中心站)	0.203	0.300	0.270	0.360	0.118	0.103	0.192
Sim(景点,交通站点)	0.350	0.486	0.370	0.400	0.308	0.240	0.380
Sim(人文资源,少林寺)	0.450	0.668	0.558	0.600	0.853	0.874	0.550
Sim(人民公园,中原区)	0.180	0.293	0.210	0.300	0.158	0.136	0.190
Sim(人民公园,二七区)	0.183	0.293	0.200	0.300	0.158	0.139	0.190
Sim(少林寺,龙门石窟)	0.573	0.568	0.719	0.400	0.646	0.698	0.645
Sim(少林寺,塔林)	0.496	0.601	0.589	0.600	0.575	0.592	0.600
Sim(火车站,郑州站)	0.696	0.650	0.613	0.600	0.727	0.775	0.620
Sim(住宿,二七区)	0.172	0.340	0.220	0.300	0.137	0.129	0.200

单因素相似度计算值与人工判定标准的比较如图 3 所示。

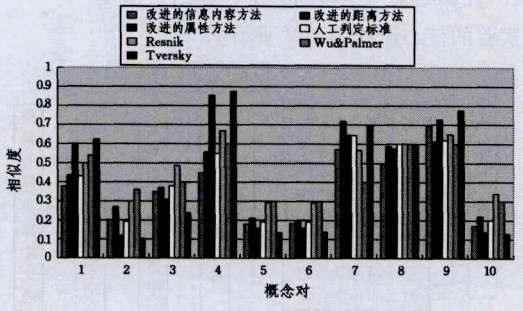

图 3　各算法计算结果与专家人工判断比较

2. 主成分分析法计算多因素贡献率

表 3 为采用改进的主成分分析法计算得到的信息内容、距离、属性 3 个因素的贡献率。

表 3　3 个因素的贡献率

因素	贡献率
信息内容	0.4869
距离	0.3280
属性	0.1851

分析表中数据可知,信息内容因素权值最大,所以对旅游本体的概念语义相似度影响最大,距离因素影响次之,属性因素影响最小。

3. ACWA 算法与传统算法的对比

将 3 个因素的贡献率作为权值进行加权计算得到最终相似度结果,与传统算法计算的结果比较如表 4 所列。

表 4　相似度计算结果比较

本体概念对的相似性	相似度算法			人工判定
	所提算法(ACWA)	基于信息内容(Resnik)	基于距离(Wu & Palmer)	
Sim(人文资源,自然资源)	0.456	0.500	0.540	0.430
Sim(人文资源,中心站)	0.215	0.300	0.360	0.192
Sim(景点,交通站点)	0.313	0.486	0.400	0.380
Sim(人文资源,少林寺)	0.495	0.668	0.600	0.550
Sim(人民公园,中原区)	0.193	0.293	0.300	0.190
Sim(人民公园,二七区)	0.193	0.293	0.300	0.190
Sim(少林寺,龙门石窟)	0.646	0.568	0.400	0.645
Sim(少林寺,塔林)	0.597	0.601	0.600	0.600
Sim(火车站,郑州站)	0.609	0.650	0.600	0.620
Sim(住宿,二七区)	0.187	0.340	0.300	0.200

3 种算法的计算结果与领域专家人工判定的参考值对比如图 4 所示,3 条曲线分别代表基于信息内容的算法、基于距离的算法和提出的 ACWA 算法的插值曲线。

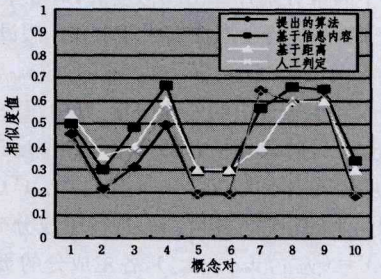

图 4　各算法计算结果与专家人工判断比较

从表 4 的实验结果和图 4 的对比结果可以看出:

(1)在信息内容值相差较小的两个弱概念计算中,ACWA 算法较基于信息内容的 Resnik 方法更加准确。例如:"人文资源"和不相关的"中心站"。

(2)ACWA 计算结果与人工判断匹配更多,从图 3 中可以清晰地看出。

(3)ACWA 算法使权值的确定可以根据计算出的样本动态确定,具有自适应性,更加客观。例如:"人民公园"和"二七区"的相似度结果更加接近专家人工判定。

4. 利用皮尔森相关系数计算不同语义相似度算法的准确性

皮尔森相关系数(Pearson Correlation Coefficient)是一种线性相关系数,是用来反映两个变量线性相关程度的统计量,计算模型如式(15)所示:

$$\rho=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{N})*(\sum Y^2-\frac{(\sum Y)^2}{N})}} \tag{15}$$

其中,X 表示算法计算出的相似度样本,Y 表示专家人工判定的样本,N 表示变量取值的个数。结果 P 表示 X 和 Y 的相关系数,P 的绝对值越大表明两个变量相关性越强。

本文利用皮尔森相关系数计算模型，分别计算不同的语义相似度算法的计算结果与专家人工判定的真实结果间的相关度，其中，X 分别表示 ACWA，Resnik 和 Wu & Palmer 算法计算出的相似度样本，Y 表示专家人工判定的样本，N 表示样本量，采用 Matlab 实现皮尔森系数计算的结果如表 5 所列。

表 5　相似度算法与人工判断相关系数

算法	参照标准	皮尔森相关系数
ACWA	人工判断	0.9431
Resnik	人工判断	0.9367
Wu & Palmer	人工判断	0.8168

从表 5 中的数据可以看出，本文 ACWA 算法计算出的结果与专家人工判定的真实结果（参照标准）的相关度最大（即皮尔森相关系数最大），比 Resnik 和 Wu&Palmer 算法的皮尔森相关系数平均高出 8.1%，从另一个侧面说明了基于本体层次结构和主成分分析法提出的 ACWA 算法的计算结果更接近人类专家的判断结果，相似度计算准确性较高。

结束语　语义网的快速发展以及本体的广泛应用产生了语义异构问题，本体语义相似度计算是解决语义网中语义异构的关键环节。针对传统本体概念语义相似度计算的不足，充分考虑相似度计算的影响因素和目前综合加权算法中专家人工赋权的严重不足，提出了改进的相似度计算算法 ACWA。ACWA 考虑影响本体概念的 3 个主要因素并对相应的已有算法进行改进，结合改进的主成分分析法对各个因素计算出的样本结果动态加权并线性相加计算出最终相似度。实验结果证明，该算法与人工判断的相似度值的相关度优于传统算法，尤其是主成分分析法自适应动态赋权，降低了人工赋权的误差，有效确保了计算的全面性、准确性。

参 考 文 献

[1] Liu H Z, Xu D. Ontology Based Semantic Similarity and Relatedness Measures Review[J]. Computer Science, 2012, 39(2): 8-13 (in Chinese)
刘宏哲，须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学，2012，39(2)：8-13

[2] Bae M, Kang S, Oh S. Semantic similarity method for keyword query system on RDF[J]. Neurocomputing, 2014, 146(C): 264-275

[3] Zhang C, Yang Y, Guo X, et al. The Improved Algorithm of Semantic Similarity Based on the Multi-dictionary[J]. Journal of Software, 2014, 9(2): 324-328

[4] Wang T, Wang L, Wu J Y, et al. Semantic Similarity Calucation Method of Comprehensive Concept in WordNet[J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36(2): 98-106 (in Chinese)
王桐，王磊，吴吉义，等. Word Net 中的综合概念语义相似度计算方法[J]. 北京邮电大学学报，2013，36(2)：98-106

[5] Resnik P. Using information content to evaluate semantic similarity in a taxonomy[C]// Proceedings of 14th International Joint Conference on Artificial Intelligence. 1995: 448-453

[6] Seco N, Veale T, Hayes J. An intrinsic information content metric for semantic similarity in Word Net[C]// Proc. of 16th European Conference on Artificial Intelligence. 2004: 1089-1090

[7] Wu Zhi-biao, Palmer M. Verbs semantics and lexical selection [C]// Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Morgan Kaufmann, 1994: 133-138

[8] Wei J Y, Zhong P S, Guo C F. Improved Semantic Similarity Algorithm Based on Ontology[C]// Applied Mechanics and Materials. 2011: 709-714

[9] Tversky A. Features of Similarity [J]. Psycological Review, 1977, 84(4): 327-352

[10] Sun H X, Qian J, Cheng Y. Review of Ontology-based Semantic Similarity Measuring[J]. New Technology of Library and Information Service, 2010(1): 51-56 (in Chinese)
孙海霞，钱庆，成颖. 基于本体的语义相似度计算方法研究综述[J]. 现代图书情报技术，2010(1)：51-56

[11] 雷钦礼. 经济管理多元统计分析[M]. 北京：中国统计出版社，2002：152-185

[12] Jia L M, Zheng Z Y, Li D, et al. Research on Similarity Algorithm of linked Data Based on Dynamic Weight[J]. Computer Science, 2014, 41(8): 263-266 (in Chinese)
贾丽梅，郑志蕴，李钝，等. 基于动态权值的关联数据语义相似度算法研究[J]. 计算机科学，2014，41(8)：263-266

[13] Wang X M, Yan J J. Approach of NON-ISA Concept Relation in Ontology Concept Similatity Computation[J]. Computer Science, 2011, 38(7): 250-254 (in Chinese)
王孝满，闫晶晶. 非 ISA 关系在本体概念相似度计算中的度量方法研究[J]. 计算机科学，2011，38(7)：250-254

[14] Wang Z X, Zhang D L. Optimization Algorithm for Edge-Based Semantic Similarity Calculation[J]. Pattern Recognition and Artifical Intelligence, 2010, 32(2): 273-277 (in Chinese)
王志晓，张大陆. 针对边计算法的语义相似度计算优化算法[J]. 模式识别与人工智能，2010，32(2)：273-277

[15] Huang Hong-bin, Dong Fa-hua, Deng Shu, et al. Approach of Determining Semantic Similarity among Concept between Different Ontologies[J]. Computer Science, 2008, 35(7): 153-156 (in Chinese)
黄宏斌，董发花，邓苏，等. 一种跨本体的语义相似度计算方法[J]. 计算机科学，2008，35(7)：153-156

[16] Zhang Z P, Tian S X, Liu H Q. Compositive Approach for Ontology Similarity Cpmputation [J]. Computer Science, 2008, 35(12): 142-145 (in Chinese)
张忠平，田淑霞，刘洪强. 一种综合的本体相似度计算方法[J]. 计算机科学，2008，35(12)：142-145

[17] Cao Z W, Qian J, Zhang W M, et al. A Compositive Approach for Concept Similarity Computation[J]. Computer Science, 2007, 34(3): 174-175 (in Chinese)
曹泽文，钱杰，张维明，等. 一种综合的概念相似度计算方法[J]. 计算机科学，2007，34(3)：174-175

[18] Lv H H, Song D W, Yang R. Weighted semantic similarity algorithm based on domain ontology[J]. Computer Engineering and Design, 2013, 34(12): 4209-4213 (in Chinese)
吕欢欢，宋伟东，杨睿. 基于领域本体的综合加权语义相似度算法研究[J]. 计算机工程与设计，2013，34(12)：4209-4213

[19] Wei J Y, Zhong P S, Guo C F. Improved Semantic Similarity Algorithm Based on Ontology[C]// Applied Mechanics & Materials. 2011: 709-714