基于动态权值的关联数据语义相似度算法研究

贾丽梅 郑志蕴 李 钝 王振飞

(郑州大学信息工程学院 郑州 450001)

摘 要 语义相似度计算对关联数据的信息检索有重要作用,直接影响数据的语义挖掘效果。实例的属性信息是关联数据语义相似度计算的一个重要因素。针对传统的关联数据语义相似度算法未考虑属性的重要性和取值类型导致计算精度较低的问题,提出基于动态权值的关联数据语义相似度计算方法,即根据待匹配的数据集中属性不同取值的数量、属性值的分布以及属性的有效性3个因素动态计算属性的权值,然后依据属性取值类型选用匹配相似度算法,最后结合属性的动态权值对概念进行实例的相似度计算。实验表明,基于动态权值的相似度计算方法与传统方法相比,实例相似度的计算精度得到了一定的提高。

关键词 关联数据,语义相似度,实例属性,动态权值

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.055

Research on Semantic Similarity Algorithm of Linked Data Based on Dynamic Weight

JIA Li-mei ZHENG Zhi-yun LI Dun WANG Zhen-fei

(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract Semantic similarity calculation has an important role in information retrieval of linked data, and the results of calculation directly affect the effect of data mining. The attribute information of instance is an essential factor for semantic similarity computation of linked data. To solve the problem of lower computation precision caused by lack of considering the importance of attribute and type of attribute value, this paper proposed a new semantic similarity calculation method based on dynamic weight. This method dynamically computes the attribute weight according to quantity of different attribute values, distribution of attribute values, and validity of attribute. Then, it chooses the matching similarity algorithm of attributes according to the types of attribute value. Finally, it combines the dynamic weight of attributes to calculate the semantic similarity of instances. The experiment confirms that the computation precision of the semantic similarity of instances obtained from the methods in this thesis is better than existing methods.

Keywords Linked data, Semantic similarity, Instance attributes, Dynamic weight

1 引言

目前,关联数据[1](linked data)作为发布和互联结构化数据的新媒介,受到了学术界、工业界和政府部门的广泛关注。越来越多的数据(特别是政府数据和企业数据[2])以关联数据的形式发布到网上,并通过RDF链接起来,形成了巨大的数据资源网。发布数据量的增多,使得整合分布式、异构或自治数据的复杂性减弱,同时也推动了关联数据走向应用[3]。在关联数据的应用中,RDF数据的爆炸式增多无疑给关联数据的分析和处理带来了挑战。如何从海量数据中进行有效的语义发现是目前亟待解决的难题。

在关联数据语义发现研究中,通常采用实例之间的相似度计算来进行有效的语义挖掘。目前国内外针对实例间相似度计算进行了很多探索性研究。Tversky[4]提出基于属性的计算方法,利用实例的属性集信息进行相似度计算,即两个实例拥有的共同属性越多,相似度越大。文献[5]基于属性进行实例差异度计算,差异度越大,实例相似度越低。文献[6]提出符合RDF数据模型规范的语义关联表示模型,即基于数据间的属性使用统计方法进行数据间语义相似性的推导。这些研究都是基于属性进行语义相似度计算,但是没有考虑属性的重要性和取值类型因素。在实际中,各维属性的重要性是不同的,要结合不同的属性取值类型选用相应的属性计算方法。

2 语义相似度计算相关技术

语义相似度是指实例对间的相似程度,可以用实例对某些共有属性的相似情况来表示。文献[7]对相似度计算方法系统地进行分类,并对各类方法进行了详细的比较。该文献虽然是基于本体进行分析的,所述的相似度计算方法对于RDF型数据并不是都适用,但也是很好的借鉴。例如,基于属性语义相似度计算方法就是关联数据实例相似度计算中常用的方法。

到稿日期:2013-09-16 返修日期:2013-12-07

贾丽梅(1988—),女,硕士,主要研究方向为智能信息处理,E-mail:klimg2008@126.com;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;李 钝(1975—),女,博士,讲师,主要研究方向为信息检索、数据挖掘;王振飞(1973—),男,副教授,主要研究方向为分布式计算、信息安全。

经典的基于属性语义相似度计算[4]的核心思想是：两个实例如果共同拥有的属性越多，则相似度就越高；反之，相似度就越低。计算方法如式(1)所示：

$$\mathrm{Sim}(X,Y)=\frac{|X\cap Y|}{|X\cap Y|+|X-Y|+|Y-X|}=\frac{|X\cap Y|}{|X\cup Y|-|X\cap Y|} \tag{1}$$

其中，$|X\cap Y|$表示概念 X 和 Y 共有属性的个数，$|X-Y|$表示属于 X 但不属 Y 的属性个数，$|Y-X|$表示属于 Y 但不属于 X 的属性个数。

在计算两个实例相似度的过程中，涉及到单对属性相似度计算。目前关于单对属性的相似度计算方法主要有字符相似度计算、数值型相似度计算等。

①字符相似度计算[8]：通过字符的插入、删除和替换等操作将字符串 x_1 转化为字符串 y_1，用转化过程操作的次数计算两个字符串的相似度。计算方法如式(2)所示：

$$\mathrm{Sim}(x_1,y_1)=1-\frac{N(x_1,y_1)}{\max\{(|x_1|,|y_1|)\}} \tag{2}$$

其中，$N(x_1,y_1)$指将 x_1 转化为 y_1 的操作次数，$\max\{(|x_1|,|y_1|)\}$指两个字符串的最大长度。

②数值型相似度计算：通过计算两个数值型属性值的差异计算数值型属性的相似度，不要求数值完全相同，只需将两者的差别限制在一定的范围，两者的差的绝对值越大，相似度越小。计算方法如式(3)所示：

$$\mathrm{Sim}(x_1,y_1)=1-\frac{|x_1-y_1|}{\max(x_1,y_1)} \tag{3}$$

在实际应用中，应结合具体场景调整相应的属性相似度计算方法，即在计算单对属性之间相似度时，要根据属性的取值类型选取恰当的相似度计算方法。

在基于属性的语义相似度计算方法中，属性对于实例相似度计算的影响并不是完全相同的，如果在相似度计算中考虑属性的重要性，则相似度计算的准确性可以得到进一步的提高。目前针对属性的重要性方面已有一些探索性研究。文献[9]利用 OWA 算子对属性的权值进行计算，并通过实验证明利用该方法计算的权值得到的实例相似度更能准确揭示实例之间的语义相似关系。但是该方法需要人为地对属性的重要性进行判断，主观因素过大，不能自动计算属性的权值。文献[10]提出根据数据集中相关属性的统计信息确定属性的权值，即对于一个属性，假设在数据集中，该属性有 n 个取值，而以该属性为谓词的三元组有 N 个，则该文献把 n/N 作为该属性的权值，如果权值越大，则该属性越重要(该方法可简称为文献[10]方法)。该文献只考虑了属性的不同取值数量对该属性权值的影响，实际上，属性的权值与属性值的分布也有一定的关系。文献[11]基于属性的不同取值数量以及属性值分布情况对属性权值进行计算，并对提出的权值计算模型进行合理性证明，弥补了文献[9]的部分缺陷(该方法可简称为文献[11]方法)。这些研究所提出的权值确定方法是基于整个数据集的统计信息，虽然在一定程度上提高了语义相似度计算的精度，但是这些研究都是固定权值的计算方法，权值一旦确定就不再改变。在实际中，对实例相似度进行计算时，研究对象不一定拥有数据集中所有的属性，所以基于整个数据集的统计信息得出的属性权值并不能很好地反映属性在实例对相似度计算中的重要性。对此，本文提出一种动态权值确定方法(简称为动态权值方法)，即在计算不同实例对的语义相似度时，根据研究对象所具有的属性缺失情况动态确定属性的权值。

3 基于动态权值的关联数据语义相似度算法

合理的属性权值能够提高相似度计算的准确度，目前属性权值的计算方法主要有主观赋权法和客观赋权法。主观赋权法根据专家主观上对各属性的重视程度确定属性权值，难以避免人为因素带来的偏差，有时会使赋权结果很不合理，而且主观赋权法往往因人而异，并不能客观反映属性在相似度计算中的重要性。客观赋权法利用数据集的统计信息得出相应的属性权值。由于避免了人为因素的干扰，客观赋权法得到的权值较主观赋权法更为合理科学。基于客观赋权法的优势，本文利用数据集的统计信息，提出一种动态赋权的关联数据语义相似度计算方法。

在关联数据集中，每个实例都有多个属性，这些属性都是实例的语义信息，但是不同属性所起的作用却是不同的，即属性的重要程度不同。在相似度计算时，主要属性对计算结果的影响大于次要属性。例如，计算两个电影是否相似时，标题和类型属性应该比时长和出版日期等属性对结果的影响更大。因此，计算实例相似度时，属性的重要性因素可以进一步提高计算的精度。本文首先基于属性的重要性因素，利用动态赋权方法确定属性的权值(反映属性的重要性)，然后利用属性相似度混合计算方法对属性相似度进行计算，最后结合属性权值和属性相似度计算结果对实例相似度进行计算。计算方法如式(4)所示：

$$S(X,Y)=\frac{f_{com}(X,Y)}{f_{total}(X,Y)}=\frac{\sum_{i=1}^{k}\mathrm{Sim}_{p_i}(X,Y)w_{p_i}}{\sum_{i=1}^{k}w_{p_i}} \tag{4}$$

其中，$S(X,Y)$表示实例 X 和 Y 的相似度，k 表示实例 X 和 Y 共有属性的个数，p_i 表示第 i 个共有属性，$\mathrm{Sim}_{p_i}(X,Y)$表示实例 X 和 Y 关于属性 p_i 的相似度，w_{p_i}表示属性 p_i 的权值。

3.1 动态赋权方法

文献[10,11]都是基于整个数据集的统计信息计算属性的权值，属性权值一旦确定就不会发生改变，这种固定属性权值计算方法会影响相似度结果的准确性，例如，两个比较对象的某些属性缺失，在进行相似度计算时，这些缺失的属性相似度值为 0，会导致整个相似度结果偏低，从而影响最后的判断。因此，属性的权值要根据实例属性的缺失情况做动态的调整。

首先，对数据集信息进行统计分析，得到属性集 $P=\{p_1,p_2,\cdots,p_m\}$(m 表示属性个数)以及每个属性的取值分布情况。比如，属性 p_i 在数据集中出现的次数为 N，其不同取值集为 $O=\{o_1,o_2\cdots o_n\}$，属性的不同取值个数为 $n(n\leqslant N)$，每个不同取值 o_l 出现的次数为 n_l，且 $\sum_{l=1}^{n}n_l=N$。

根据数据集的统计信息，利用属性的不同取值个数和属性值的分布情况对属性 $p_i(i=1,\cdots,m)$ 的权值进行初始化(初始化的属性权值是基于整个数据集的统计信息)。从属性 p_i 的 N 个取值中任选 2 个，其值相等的概率如式(5)所示：

$$P(V_{p_{i_j}}=V_{p_{i_k}})=\frac{\sum_{l=1}^{n}n_l(n_l-1)}{N(N-1)} \tag{5}$$

属性取值相等的概率越大，则属性的重要性越弱，进而属性的权值越小。例如，性别属性值相同的概率比姓名属性值

相同的概率大，但在计算两个人的相似度时，姓名比性别更重要。因此，属性的权值可以用 $P(V_{p_{i_j}}=V_{p_{i_k}})$ 的倒数计算。为了使计算的权值变化平缓，可以对权值计算公式进行平滑处理，处理后的权值计算公式如式(6)所示：

$$w_{p_i}=\log\left(\frac{N(N-1)}{\sum_{l=1}^{n}n_l(n_l-1)+1}+1\right) \tag{6}$$

通过式(6)可以计算数据集中所有属性的权值，为了将权值限定在[0,1]范围内，可以对式(6)进行改进，改进后的权值计算方法如式(7)所示：

$$w_{p_i}=\log\left(\frac{N(N-1)}{\sum_{l=1}^{n}n_l(n_l-1)+1}+1\right)/w_{p_{\max}} \tag{7}$$

其中，$w_{p_{\max}}$ 表示利用式(6)计算的最大属性权值。式(7)计算的属性权值是基于整个数据集的，每个属性权值都受到所有属性的影响，然而，在计算具体实例相似度时，并不是每个实例都具有数据集中的所有属性，有些属性是缺失的，其他属性的权值应该不受这些缺失属性的影响，所以本文根据实例属性的缺失情况动态确定属性的权值。属性是否缺失可以用属性是否有效表示，属性无效则表示属性缺失，用式(8)表示：

$$val_{p_i}=\begin{cases}0, & \text{属性 } p_i \text{ 无效}\\ 1, & \text{属性 } p_i \text{ 有效}\end{cases} \tag{8}$$

如果某一个属性无效，则该属性对实例相似度计算的影响可以忽略，$w_{p_{\max}}$ 也会做相应的变化，一旦 $w_{p_{\max}}$ 发生改变，所有属性的权值也要随着发生改变，这样可以减少由于属性的缺失对相似度计算的影响。其中属性的有效性判断是在实例相似度计算时进行的，因此，属性权值的动态调整在此过程中自动进行。动态属性权值的计算公式如式(9)所示：

$$w'_{p_i}=val_{p_i}*\log\left(\frac{N(N-1)}{\sum_{j=1}^{n}n_j(n_j-1)+1}+1\right)/w'_{p_{\max}} \tag{9}$$

其中，$w'_{p_{\max}}$ 表示的是利用式(6)计算的有效属性集中最大的属性权值，大小随着属性的有效性动态发生改变。

最后，运用式(4)进行实例对的相似度计算，将权值计算式(9)代入，式(4)可转化为：

$$S(X,Y)=\sum_{i=1}^{k}\mathrm{Sim}_{p_i}(X,Y)\frac{\mathrm{val}_{p_i}*w'_{p_i}}{\sum_{i=1}^{k}\mathrm{val}_{p_j}*w'_{p_j}} \tag{10}$$

其中，k 表示实例 X 和 Y 共有属性的个数，$\mathrm{Sim}_{p_i}(X,Y)$ 表示实例 X 和 Y 关于共有属性 p_i 的相似度计算值，本文采用属性相似度混合计算方法对其进行计算。

3.2 相似度混合计算方法

关联数据语义相似度计算除了受属性权值影响外，还受属性的取值类型影响。为了提高相似度计算的精度，本文还对属性的取值类型分别进行研究。根据属性的取值类型将属性值分为两大类：URI 型和文本型，不同类型的属性值选用不同的属性相似度计算方法。

3.2.1 URI 型属性值

如果属性的取值类型为 URI 型，则称该类属性值为 URI 型属性值。本文采用深度搜索方法，首先判断两个 URI 是否相同，若相同，则关于该属性的相似度为 1，否则，挖掘这两个 URI 所具有的属性，计算这两个 URI 的属性相似度，依次递归，最后 URI 型属性值相似度计算会转化为文本型属性值相似度计算。

3.2.2 文本型属性值

如果属性的取值类型非 URI，则称该类属性值为文本型属性值。本文按照文本的字符串特点将其分为数值型、日期型和普通文本型，再按照文本的长度将普通文本型分为长文本型和短文本型。

(1)数值型属性值

如果属性的取值全部为数字或带有小数点的数字，则称该类属性值为数值型。采用第 2 节中的式(3)计算。

(2)日期型属性值

如果属性的取值类型为日期型，则称该类属性值为日期型。对于日期型属性值的相似度计算公式，本文同样不要求日期完全相同，而是将日期字符串按一定的比例进行截取，通过相应截取字符串的相似度与截取字符串的比例计算日期型属性值的相似度。计算方法如式(11)所示：

$$\mathrm{Sim}(x_1,y_1)=\frac{len(h_1)}{len(x_1)}\left(1-\frac{N(h_1,h_2)}{\max\{(|h_1|,|h_2|)\}}\right) \tag{11}$$

其中，$h_i(i=1,2)$ 表示日期串 x_i 截取的字符串，$len(h_1)$ 表示截取的长度，$len(x_1)$ 表示日期的长度。

(3)长文本型属性值

如果文本属性值(非数值和日期型)长度大于一定的阈值，则称该类属性值为长文本型属性值，例如标题或描述等属性，这些属性对应的属性值一般都是长文本型。一般的字符相似度计算方法并不能很好地适用于长文本型属性值相似度计算。针对该类型的属性值，本文通过术语抽取等预处理，将长文本型属性值的相似度计算转化为词组集合的相似度计算。对于词组集合的相似度文本，采用向量空间模型进行计算，向量中每一维都对应一个词组，其值就是该词组出现的频率。两个长文本型属性值的相似度就是相应向量的接近度，可以用向量之间夹角的余弦表示，计算方法如式(12)所示：

$$\mathrm{Sim}(x_1,y_1)=\cos\theta=\frac{\sum_{k=1}^{n}w_{1_k}\times w_{2_k}}{\sqrt{\left(\sum_{k=1}^{n}w_{1_k}^2\right)\left(\sum_{k=1}^{n}w_{2_k}^2\right)}} \tag{12}$$

其中，w_{1_k} 表示属性值 x_1 第 k 个特征项的权值，n 表示属性值含有的特征项个数。

(4)短文本型属性值

如果文本属性值(非数值和日期型)长度小于一定的阈值，则称该类属性值为短文本型属性值。在相似度计算时采用第 2 节中的式(2)计算。

在两个实例相似度计算时，共有属性的取值类型并非只有一种，本文选用的属性相似度混合计算方法可以进一步提高相似度计算的准确性。

4 实验结果与分析

为了验证提出的动态权值关联数据语义相似度计算方法的优越性，基于 Java 编辑工具 My Eclipse、语义网开发工具开源包 JENA 以及 MySQL 数据库对 RDF 数据集进行存储和分析，实现权值和相似度的计算，并利用 RKB Explorer[12] 管理的部分数据集和 FOAF 部分数据集进行测试，比较本文提出的动态权值确定方法与文献[10,11]提出的固定权值方法在不同阈值下得到的实例对相似度计算的精度。

4.1 实验环境和实验步骤

Jena 对 MySQL、HSQLDB、Oracle 和 Microsoft SQL Server 等提供了支持的程序接口，用以将 RDF 数据集存储到关系数据库中。本文使用 Jena 2.5 API 和 MySQL 5.0.4-beta-nt 数据库，MySQL 驱动包使用的是 mysql-connector-java-

3.1.10-bin.jar。

测试数据集选用 RKB Explorer[12] 管理的部分数据集和 FOAF 部分数据集。这些数据集都是 RDF 三元组格式，针对的主要是语义网的应用，而且 RKB 还提供了 Consistent Reference Service(CRS)在线服务，可以对实例对的相似度进行在线验证。实验核心流程图如图 1 所示。

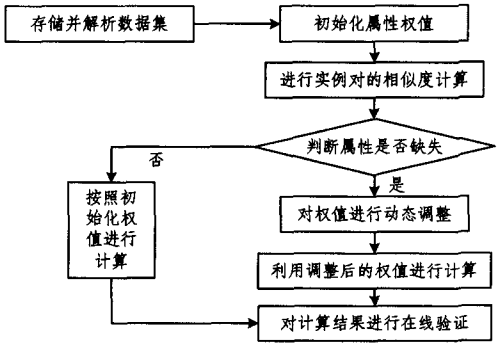

图 1　实验核心流程图

实验步骤：

1)下载 RDF 数据集，通过 Jena 工具包中的 model 将 RDF 数据集存储到 MySQL 数据库中。

2)对存储到数据库中的数据进行统计分析处理，并对本文提出的权值确定方法进行编程实现。

3)计算数据集中实例对的相似度，在计算过程中，根据实例对属性缺失情况实现权值的动态调整。

4)对计算结果进行在线验证，计算实验结果的精度。

4.2　实验分析

RKB Explorer 管理的 ACM 关联数据集格式和内容比较统一，容易进行相似数据的发现，而且 CRS 针对数据集的相似数据提供在线验证，本文针对 ACM 测试数据集进行了测试。本文使用的 ACM 测试数据集中包含 100 多万个 Publication 实例数据，其属性的取值类型包含 URI 型、数值型、日期型、长文本型以及短文本型。首先，通过 TDB 工具对测试数据集进行解析，从属性的统计信息可以得到该数据集中的属性个数以及包含每个属性的三元组个数，该数据集的属性缺失情况并不是很频繁。根据属性 http://www.aktors.org/ontology/portal#has-title 对应的属性值采用式(1)计算实例对的相似度，过滤最不相似的实例对。然后，基于本文提出的动态权值相似度计算方法对过滤后的实例对进行深层次的相似度计算。本文设定 4 个阈值，其对应的相似度分别为 60%，70%，75%，80%。如果相似度的值大于设定的阈值，则把对应的实例对输出到相应的文档中。最后，分别从 4 个不同阈值得到的文档中随机选取 50 个实例对，利用 CRS 在线服务进行验证。

首先，利用 ACM 测试数据集，对仅仅考虑属性重要性和仅仅考虑取值类型这两种情况进行分析，计算相似度精度，结果如表 1 所列。

表 1　ACM 测试数据集不同情况下的相似度计算精度

精度/阈值 \ 方法	仅仅考虑属性重要性	仅仅考虑取值类型	动态权值方法
0.6	0.6900	0.4250	0.7650
0.7	0.8350	0.5600	0.8900
0.75	0.9100	0.6500	0.9400
0.8	0.9400	0.8100	0.9700

分析表 1，本文提出的动态权值方法较前两种情况所得的相似度计算精度高。从表 1 中也可以得出充分利用数据的语义信息，其可以有效地提高相似度计算精度。

然后，比较本文提出的方法与文献[10]方法和文献[11]方法的计算精度。ACM 测试数据集在不同阈值 t 下的相似度计算精度验证结果如表 2 所列。

表 2　ACM 测试数据集下不同阈值的相似度计算精确度

精度/阈值 \ 方法	文献[10]方法	文献[11]方法	动态权值方法
0.6	0.6600	0.7100	0.7650
0.7	0.7850	0.8300	0.8900
0.75	0.8600	0.8950	0.9400
0.8	0.9100	0.9300	0.9700

该数据集下不同方法计算结果的统计检验度量值如表 3 所列。

表 3　ACM 测试数据集下不同方法的统计检验度量

度量值/度量 \ 方法	文献[10]方法	文献[11]方法	动态权值方法
方差	0.0089	0.0071	0.0062
最大值	0.9100	0.9300	0.9700
最小值	0.6600	0.7100	0.7650

另外，本文还使用了 FOAF 测试数据集进一步进行测试，因为 FOAF 测试数据集与 ACM 测试数据集相比，格式和内容统一性较差，属性的缺失情况频繁出现，所以对该数据集进行测试可以有效地验证属性缺失影响因素对实例相似度计算的影响。同样采用上述方式，得到不同阈值 t 下的相似度计算精度验证结果，如表 4 所列。

表 4　FAOF 测试数据集下不同阈值的相似度计算精确度

精度/阈值 \ 方法	文献[10]方法	文献[11]方法	动态权值方法
0.6	0.5200	0.6150	0.7100
0.7	0.6950	0.7400	0.8200
0.75	0.7800	0.8200	0.8900
0.8	0.8400	0.8700	0.9350

该数据集下不同方法计算结果的统计检验度量值如表 5 所列。

表 5　FAOF 测试数据集下不同方法的统计检验度量

度量值/度量 \ 方法	文献[10]方法	文献[11]方法	动态权值方法
方差	0.0147	0.0097	0.0070
最大值	0.8400	0.8700	0.9350
最小值	0.5200	0.6150	0.7100

通过对表 2－表 5 的分析，可以得到下面结论：

(1)实际应用中，一般要达到的都是模糊挖掘效果，即要挖掘的并不仅限于语义完全相同的实例对，对具有一定相似程度的实例对也要进行挖掘。从表 2 可以看出，3 种方法都能对具有一定相似度的实例对进行有效的挖掘，计算精度都呈现递增趋势，阈值越大，挖掘的实例对越相似，阈值达到 1 时，实例对是精确匹配的，完全相同。在阈值比较小的情况

(下转第 273 页)

er-Verlag,2012:274-285

[11] Steichen B,Carenini G,Conati C. User-adaptive information visualization:using eye gaze data to infer visualization tasks and user cognitive abilities[C]// Proceedings of the 2013 international conference on intelligent user interfaces. New York: ACM,2013:317-328

[12] Fekete J D,Van Wijk J J,Stasko J T,et al. The value of information visualization [C]// Information Visualization. Berlin: Springer-Verlag,2008:1-18

[13] Peterson D J,Berryhill M E. The Gestalt principle of similarity benefits visual working memory [J]. Psychonomic bulletin & review,2013:1-8

[14] Rusu A,Fabian A J,Jianu R. Using the gestalt principle of closure to alleviate the edge crossing problem in graph drawings [C]// Proceedings of the 2011 15th International Conference on Information Visualisation (IV). London:IEEE,2011:488-493

[15] Healey C G, Enns J T. Attention and visual memory in visualization and computer graphics [J]. IEEE Transactions on Visualization and Computer Graphics,2012,18(7):1170-1188

[16] Michalski R, Grobelny J. The role of colour preattentive processing in human-computer interaction task efficiency:a preliminary study [J]. International Journal of Industrial Ergonomics, 2008,38(3):321-332

[17] Cui W,Qu H. A survey on graph visualization [D]. Hong Kong University of Science and Technology,2007:16-29

[18] Basu A,Blanning R. Metagraphs and Their Applications [M]. Berlin:Springer-Verlag,2006:1-11,77-115

[19] 谭政华,胡光锐,任晓林. 模糊元图及其特性分析[J]. 计算机研究与发展,2000(3):272-277

[20] Dashore P, Jain S, Dashore S R. Fuzzy Metagraph and Rule Based System for Decision Making in Share Market [J]. International Journal of Computer Applications,2010,6(2):10-13

(上接第 266 页)

下,本文方法的相似度计算精度明显优于其他两种方法,在阈值较大时相似度计算精度的差异变小。分析表 2 和表 4,本文提出的方法较文献[11]的相似度计算精度平均提高了 5%和 7%,较文献[10]平均提高了 8%和 13%。

(2)在属性缺少情况频繁出现的数据集中,例如 FOAF 数据集,在实例相似度计算时,foaf-name 以及 foaf-topic 等重要属性都存在缺失情况,在计算时会导致权值的改变。本文方法计算的相似度精度明显优于其他两种方法的计算精度,究其原因,表 2 中,本文方法的优越性主要在于考虑了属性的取值类型影响因素,表 4 中,本文方法的优越性不仅与属性的取值类型有关,还与属性的缺失情况有关。从这两个表中,也可以得出属性的取值类型以及属性的缺失情况对实例相似度计算都有影响。

(3)分析表 3 和表 5,本文提出的方法相对于其他两种方法较稳定。

结束语 针对当前传统的关联数据语义相似度计算方法考虑因素单一,没有充分利用实例属性的语义信息从而造成相似度计算结果精确性不高的问题,本文提出一种基于动态权值的关联数据语义相似度计算方法。该方法首先根据实例属性不同取值的数量、属性值的分布以及属性的有效性 3 个因素对实例属性的权值进行动态确定,然后根据属性类型的不同自动选择不同的相似度计算方法。实验表明,本文提出的相似度计算方法充分利用了实例属性的语义信息,能更准确表示出实例之间的语义相似关系,提高了计算的精度。但是本文提出的算法针对的只是单个数据集,没有考虑跨数据集实例对的相似度计算以及计算精度与阈值的关系,而且计算性能也待进一步提高以应对大规模数据集的情况。下一步拟在本文的基础上,进一步研究跨数据集的实例对的语义相似度计算,并将新计算方法应用于关联数据应用领域,以提高检索效率。

参 考 文 献

[1] Berners-Lee T. Linked data-the story so far [J]. International Journal on Semantic Web and Information Systems,2009,5(3): 1-22

[2] Servant F P. Linking enterprise data [C]// Linked Data on the Web. 2008

[3] Hartig O, Sequeda J, Taylor J, et al. How to consume Linked Data on the Web: tutorial description[C]// Proceedings of the 19th international conference on World Wide Web. ACM,2010: 1347-1348

[4] Tversky A. Features of similarity [J]. Psychological review, 1977,84(4):327-352

[5] 高学东,吴玲玉,武森,等. 基于属性与对象关系信息的综合差异度计算[J]. 计算机工程,2011,37(22):35-38

[6] Sheth A, Aleman-Meza B, Arpinar I B, et al. Semantic association identification and knowledge discovery for national security applications[J]. Journal of Database Management,2005,16(1): 33-53

[7] 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

[8] Bhattacharya I, Getoor L. Iterative record linkage for cleaning and integration [C]// Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in data Mining and Knowledge Discovery. ACM,2004:11-18

[9] Zadeh P D H,Reformat M Z. Fuzzy semantic similarity in linked data using the OWA operator[C]// Fuzzy Information Processing Society (NAFIPS). 2012 Annual Meeting of the North American. IEEE,2012:1-6

[10] Song D, Heflin J. Domain-independent entity reference in RDF graphs[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM,2010: 1821-1824

[11] 张晓辉,蒋海华,邸瑞华. 基于属性权重的链接数据共指关系构建[J]. 计算机科学,2013,40(2):40-43

[12] Glaser H,Millard I C,Jaffri A. RKBExplorer. com:a knowledge driven infrastructure for linked data providers[M]// The Semantic Web:Research and Applications. Springer Berlin Heidelberg,2008:797-801