第34卷 第19期 计 算 机 工 程 2008年10月
Vol.34 No.19 Computer Engineering October 2008
·软件技术与数据库· 文章编号：1000—3428(2008)19—0056—02 文献标识码：A 中图分类号：TP301.6

基于相似度计算的本体映射优化方法

谷志锋[1]，刘 勇[1]，郭跟成[2]
(1. 河南科技大学电子信息工程学院软件工程实验室，洛阳 471003；2. 阿尔伯塔大学电气与计算机工程系，加拿大)

摘 要：在基于相似度计算的本体映射中，相似度计算量大的主要原因是待映射概念和待计算属性过多。该文采用过滤策略，利用候选映射策略和信息增益策略减少待映射概念和待计算属性数量。该过滤策略充分利用本体特点和数据挖掘思想，有效滤除没有计算意义的概念和属性，减少了相似度计算量。实验结果证明，滤除的概念和属性对映射效果的影响很小。

关键词：本体映射；候选映射；信息增益

Optimizing Method for Ontology Mapping Based on Similarity Computation

GU Zhi-feng[1], LIU Yong[1], GUO Gen-cheng[2]
(1. Laboratory of Software Engineering, School of Electron and Information, Henan University of Science and Technology, Luoyang 471003;
2. Department of Electrical & Computer Engineering, University of Alberta, Canada)

【Abstract】The account quantum of ontology mapping based on similarity computation is vast, because the amount of concepts and attribute waiting for computing is large. This paper adopts a filtering tactic which uses candidate mapping tactic and information gain tactic to reduce the amount of concepts and attribute waiting for computing. This filtering tactic makes full use of the characteristic of ontology and the idea of data digging, and gets rid of those worthless concepts and attribute to reduce the account quantum of similarity computation. Experimental results indicate that the effect of those concepts and attribute is little.

【Key words】ontology mapping; candidate mapping; information gain

1 概述

随着本体应用的增多，异构本体间的互操作成为一个技术难题，目前最有效的方法是在本体间进行映射，即本体映射[1]。本体映射通常以 2 个本体作为输入，并为其中的各个概念建立相应语义关系。

基于相似度计算的本体映射是最常用的映射方法，但存在相似度计算量过大的缺点，成为制约本体映射效率的瓶颈。相似度计算量大主要存在于 2 个方面：(1)传统方法计算 2 个本体的概念相似度时，由于本体中每对概念都被考虑，即需要计算每对概念之间的相似度，因此计算量很大。而一些概念对完全不相似，无须计算其相似度。(2)传统方法在计算概念的属性相似度时通常考虑概念的每个属性，而概念的属性对概念的影响程度不同，有些属性对概念的影响很小，没有必要全部考虑。

针对上述问题，本文采用相应的过滤策略，利用候选映射策略解决待映射概念过多的问题，利用信息增益策略解决待计算属性过多的问题，从而滤除没有计算意义的概念和属性，在不影响映射效果的前提下最大限度地减少相似度计算的工作量。

2 候选映射策略的应用

假设有本体O1和本体O2，对于本体O1中的一个概念A，本体O2中一般只有部分概念与它相似。为了提高映射效率，可以对本体O2 中的概念进行筛选，选出最有可能与本体O1中概念A相似的概念组成候选映射集[2]，仅对A与候选映射集中的概念计算相似度，以减小计算量。传统研究已涉及如何进行筛选，例如，文献[3]使用COSA算法提取有效概念，文献[4]把概念A与本体O2中概念的名称相似度作为筛选条件，对本体O2中的概念进行筛选。因为上述方法没有充分考虑本体自身的特点，所以一开始就漏掉了很多概念，影响了本体映射的准确性。本文候选映射策略充分利用本体特点及数据挖掘思想，以综合的概念相似度计算结果作为筛选条件，提出一种改进的过滤方法。初步实验表明，此策略在保证映射准确性的前提下，有效减少了本体映射的计算量。

本体中的概念是分层的，可以把本体看成一棵概念树，树中每个节点代表一个概念。因此，概念之间的关系也具有树的一些性质。因为树中有子节点、父节点和兄弟节点等一些特有概念，所以概念树中也存在子概念、父概念和兄弟概念。根据启发规则“在同一个本体中，如果 2 个概念属于同一个父概念，那么这 2 个概念是相似的，即兄弟概念是相似的”，本体中具有同一父概念的兄弟概念具有较相近的相似关系。因此，先把本体 O2 中的概念以是否是“兄弟概念”为标准进行分类，如果以图 1 所示本体为例，经过分类后可以得到集合 $B1=\{b1\}$,$B2=\{b2,b3,b4\}$,$B3=\{b5,b6,b7,b8\}$,$B4=\{b9,b10,b11,b12\}$；然后从集合 $B1$, $B2$, $B3$, $B4$ 中分别抽取一个概念，与本体 O1 中的概念 A 进行相似度计算(例如，从 $B1$ 中

基金项目：国家自然科学基金资助项目(60475021)；河南省教育厅自然科学计划基金资助项目(200510464021)
作者简介：谷志锋(1978－)，男，硕士研究生，主研方向：GIS 互操作，本体映射；刘 勇、郭跟成，副教授
收稿日期：2008-01-03 **E-mail：**gzf62296918@163.com

取 $b1$，从 $B2$ 中取 $b3$，从 $B3$ 中取 $b7$，从 $B4$ 中取 $b11$，如果某个集合中的元素较多，可以根据情况抽取多个，分别和 A 进行相似度计算，取平均值)，比较计算出来的相似度值，把它们从大到小进行排列，并设定一个阈值，认为大于该阈值的概念和 A 相似。根据相似的传递性及上述启发规则，假设与 A 相似的概念存在于 $B1, B2$ 中，则可以把 $B1, B2$ 组成的集合{$b1,b2,b3,b4$}作为候选映射集。下文相似度计算工作只要在求得的候选映射集中进行。

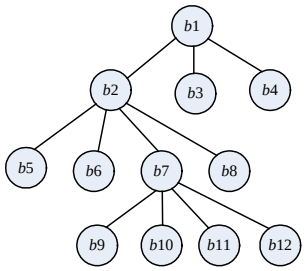

图 1　本体 O2

3　信息增益策略的应用

概念相似度的计算趋向于综合多种信息，例如综合概念名称、概念实例、概念属性等，其中，概念属性相似度计算在概念相似度计算中占有重要地位。但传统属性相似度研究普遍存在计算量过大的缺点，这是因为每个概念一般都有多个属性，在传统属性相似度计算中，通常要考虑概念的每个属性。概念的每个属性对概念的描述程度和作用各不相同，如果都考虑会加大计算量，且没有必要。需要利用合适的过滤策略筛选待计算的属性，选出对概念影响大的属性进行计算。本文根据数据挖掘中的判定树归纳思想，利用一组概念实例计算各个属性的信息增益。根据每个属性信息增益的大小确定各个属性的优先级。对属性按信息增益递减的顺序进行排列，只选取几个信息增益大的属性进行相似度计算，从而在一定程度上减少计算量。信息增益的计算过程[5]如下：

设 S 是 s 个数据样本的集合，s_i 是类 l_i 中的样本数。假定类标号属性具有 m 个不同值，定义 m 个不同类 l_i (i=1,2,⋯, m)。一个给定样本分类需要的期望信息由式(1)给出，即

$$I(s_1,s_2,\cdots,s_n)=-\sum_{i=1}^{n}(s_i/s)\mathrm{lb}(s_i/s) \tag{1}$$

设属性 X 具有 v 个不同值 $\{a_1,a_2,\cdots,a_v\}$，可以用属性 X 将 S 划分为 v 个子集 $\{s_1,s_2,\cdots,s_v\}$，其中，s_j 包含 S 中在 X 上具有值 a_j 的样本。设 s_{ij} 是子集 s_j 中类 l_i 的样本数。由 X 划分成子集的熵或期望信息由式(2)给出，即

$$E(X)=\sum_{j=1}^{v}[(s_{1j}+s_{2j}+\cdots+s_{mj})/s]I(s_{1j}+s_{2j}+\cdots+s_{mj}) \tag{2}$$

项 $(s_{1j}+s_{2j}+\cdots+s_{mj})/s$ 是第 j 个子集的权，且等于子集 s_j (即 X 的值为 a_j)中的样本个数除以 S 中的样本总数。熵值越小，子集划分的纯度越高。对于给定的子集 s_j 所需要的期望信息由式(3)给出，即

$$I(s_{1j},s_{2j},\cdots,s_{mj})=-\sum_{i=1}^{m}p_{ij}\,\mathrm{lb}(p_{ij}) \tag{3}$$

其中，$p_{ij}=s_{ij}/|s_j|$ 是 s_j 中的样本属于类 l_i 的概率。

属性 X 的信息增益由式(4)给出，即

$$Gain(X)=I(s_1,s_2,\cdots,s_m)-E(X) \tag{4}$$

重复以上步骤可获得概念每个属性的信息增益，对属性按信息增益递减的顺序进行排列，只选取信息增益大的属性进行相似度计算。

4　风险评估与分析

由于本体研究具有复杂性，因此本文在解决相似度计算量大的问题时，分别采用 2 种不同的过滤策略，下文分别对其进行分析和评估。

(1)候选映射策略分析。为了方便评估候选映射策略的优劣，称利用此策略得到的集合数为概念深度，把每个集合中概念的个数称为概念广度。只把相似度值最大的概念所在集合作为侯选映射集，设本体 O3 如图 2 所示。用 K 表示本体的概念深度，N 表示集合概念广度的平均值，M 表示选中集合的概念广度。当 N 足够小(最小为 1)，K 足够大时，本文方法的计算量与没有设置侯选映射集时基本相同，没有达到减少计算量的目的，如图 3 所示。当 K 足够小(最小为 2)，M 足够小时，本文方法效果显著。分析表明，除了如图 3 所示的特殊情况，本文方法在一定程度上减少了计算量，当 K 足够小(最小为 2)，M 足够小时达效果最好。由于考虑本体的结构特点及相应启发规则，因此利用本文方法得到的侯选映射集，很少遗漏概念，且漏掉的概念对映射查全率的影响极小。

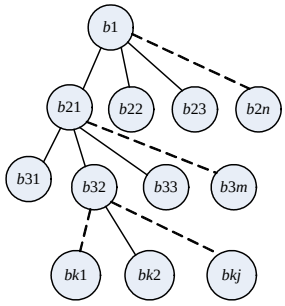

图 2　本体 O3

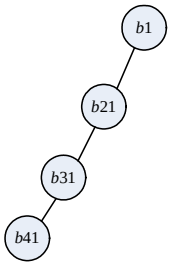

图 3　本体 O4

(2)对信息增益策略进行分析和评估。由于信息增益的计算依赖于具体实例，因此数据实例个数的多少直接影响信息增益的计算量。为了方便描述，用 m 表示数据实例个数，用 n 表示属性个数。如果 m 很大，n 很小，此时该策略对属性相似度计算量的影响不大。如果 m 很小，n 很大，本文策略将得到充分发挥。因此，信息增益策略不适合于数据实例特别多的计算。通常情况下，如果数据实例很多，可以选取一些具有代表性的实例来计算，从而减少计算复杂度，使信息增益策略得到充分发挥。

5　实验与测试

本文实验主要在文献[6]设计的典型数据集 Course Catalog ontology I 上进行。该数据集中的本体分别描述了康奈尔大学和华盛顿大学的课程信息。本文算法用 Jena 语言实现，开发平台为 Eclipse，实验测试在一台 2.93 GHz CPU、512 MB 内存的计算机上进行。

本文实验有 2 个目的：(1)验证本文算法是否减少了计算量，可以用 TempLoadRunner 工具进行测试；(2)验证本文算法是否影响了映射的准确性，可以使用查全率和查准率进行评估。查全率和查准率的计算式分别为 $R=A/C$, $P=A/B$，其中，C 是可能存在的概念映射对；B 是通过计算发现的概念映射对；A 是专家认为正确的概念映射对。

对同一组实验数据进行 2 次实验，一次使用本文算法，另一次不使用本文算法，测试结果如图 4 所示。由图 4 可以看出，本文策略极大提高了映射效率，且对查全率和查准率的影响很小。

（下转第 60 页）