一种基于语义的本体概念相似度的计算方法

吴开贵　万红波　朱郑州
（重庆大学计算机学院　重庆 400044）

摘　要　本体映射是实现异构本体互操作的有效方法，其关键技术在于概念相似度的计算。本文提出了一种概念相似度的计算方法，主要从概念名称、概念属性、概念关系来计算概念相似度，通过引入候选概念集和信息增益，提高了相似度的准确率，简化了相似度的计算过程。

关键词　本体映射，概念相似度，信息增益

A Computation Method of Conceptual Similarity in Ontology Based on Semantic Web

WU Kai-gui　WAN Hong-bo　ZHU Zheng-zhou
(College of Computer, Chongqing University, Chongqing 400044, China)

Abstract　Ontology mapping is an effective method to realize the interoperation of heterogeneous ontologies, the key of ontology mapping is to calculate similarities between concepts. A new ontology mapping approach is put forward, the concept similarity is computed from the aspects of concept name, concept attribute and concept relationship. What's more, the candidate set and information gain are brought to improve the accuracy of similarity and to simplify the calculation process.

Keywords　Ontology-mapping, Concept similarity, Information gain

1　引言

语义网采用多层次的表示框架，而本体位于从文档描述到知识推理转折的层次，因此本体的构建是实现语义网的关键环节。本体就是用来描述某个领域（领域本体）甚至更广范围（通用本体）内的概念以及概念之间的联系，使得这些概念和联系在共享的范围内有着明确唯一的定义，这样人和机器之间就可以进行交流[1]。但是，由于在网络中可以获取的本体数量越来越多，并且本体的创建者不同，使用的建模方法不同，因而即使对同一个领域内的问题建模，不同的领域专家开发出来的本体必然存在着差别。为了使这些领域专家们所创建的本体，可以互相"理解"，本体映射应运而生。本体映射的目的就是找到这些本体之间的语义联系，实现本体合并[2]。

文献[3]认为本体可理解为概念、属性和关系的集合。属性即概念的属性，关系即概念间的关系，所以本体映射主要是集中在概念间的相似度计算，并求出本体中概念的相似矩阵。当其相似度大于某个阈值时就认为这两个概念间存在一定的映射关系。

现有的概念相似度计算方法可分为基于句法的方法和基于语义的方法。ACAOM[4]是一种基于语义的方法，采用了基于名称策略和基于实例策略，该方法在名称比较中没有考虑词与词之间的结构，认为每个词都有相同的重要性，这样增加了计算复杂度；向量比较法采用的算法仅对词出现的频率进行简单的计算，没有分析每个词在文章中的重要性，因此向量的比较正确率不高。针对 ACAOM 存在的问题，本文提出一种基于语义的计算方法。

2　本体映射

本体映射通过定义条件规则、函数、逻辑以及表达关系的集合来实现不同本体间的映射，是完成本体集成的重要一步工作（本体集成的概念包括本体的重用、本体合并、本体修正等）。或者说，本体映射是不同的本体在概念层语义相关联[5]，源本体的实例可根据语义关联的关系转换为目的本体。下面举例说明本体映射的概念。

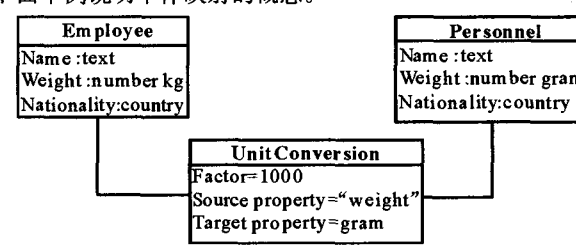

图 1　Employee 本体与 Personnel 本体的映射

如图 1 所示，Employee 和 Personnel 是同一个公司两个不同部门建立的雇员本体，因此存在一定差异，即 Weight 属性的度量单位不同，因而可以利用 Unit Conversion 的映射规则来建立这两个本体之间的映射关系。

3　概念相似度的计算

3.1　方法的设计思想

本文通过两个概念的基距离来过滤出候选概念集，在计算概念相似度时基于名称策略、概念属性、概念关系分别计算概念相似度，最后通过赋予一定的权值进行相似度合并，这样可使概念相似度的计算更加全面，计算结果更加准确。同时在计算概念属性相似度时引入信息增益，进一步缩小概念范围，减少概念相似度的计算量，其整个过程如图 2 所示。

3.2　基于名称策略计算概念相似度

利用 WordNet 计算语义关系的方法很多，最为简单的方法就是利用路径长度的方法，它将 WordNet 看作一个图，通过识别两个词义之间的最短路径从而发现语义关系上的相似

性，即从一个节点到另外一个节点的路径越短，则表明它们相似的程度越高。

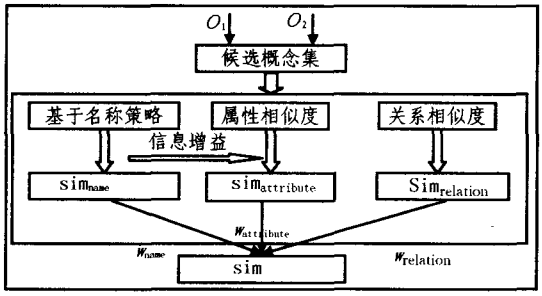

图 2　概念相似度的计算过程

Leacock-Chodorow 方法是在 1998 年提出的，利用了路径的方法计算两个概念的相似程度，它只限于 is a kind of 关系，因此受限程度比较大。

Lin 利用概率的方法计算两个概念的相似度。

$$\mathrm{Sim}_{name}(A,B)=\frac{2\cdot\log(p(A,B))}{\log(p(A))+\log(p(B))} \tag{1}$$

由于 WordNet 所包含的名词和动词组织成了 is-a 关系的词汇数据库，所以 WordNet 非常适合进行概念相似度计算。本文采用 WordNet 作为辅助信息来计算两个本体中的概念的相似度，通过结合路径的方法[6]，对 Lin 计算概念相似度的方法进行了改进，得到下面公式。

$$\mathrm{Sim}_{name}(A,B)=\frac{2\cdot\log(p(A,B))}{\log(p(A))+\log(p(B))}\cdot\frac{1}{2}a^{l} \tag{2}$$

在寻找词义 A 和 B 的共同上位词时，通过路径的方法，设定了一个系数 $\frac{1}{2}a^{l}$。其中 a 是一个介于 0 和 1 之间的常数，用来调整随层次加深，相似度随之递减的程度，l 表示在某个同义次层次结构中，寻找词义 A 和词义 B 的共同上位词的最大路径。由于 A 和 B 两个词义，故乘上 1/2。

3.3　基于概念属性计算相似度

基于属性计算概念相似度的理论依据是：如果两个概念的属性都相同，那么这两个概念是相同的；如果两个概念具有相似的属性，那么这两个概念也是相似的。属性有属性名称、属性数据类型、属性实例数据等要素，因此判断两个属性是否相似主要从这三个要素的相似度进行考虑。

设概念 A 的属性为 a，概念 B 的属性为 b，两个属性间的相似度记为 $\mathrm{Sim}(a,b)$。属性相似度计算公式如下：

$$\mathrm{Sim}(a,b)=w_1\mathrm{Sim}_{name}(a_{name},b_{name})+w_2\mathrm{Sim}_{data}(a_{data},b_{data})+w_3\mathrm{Sim}_{ins}(a_{ins},b_{ins})$$

其中 w_1,w_2,w_3 是权重，代表属性名称、类型、数据对属性相似度计算的重要程度 $w_1+w_2+w_3=1$。

设概念 A 和概念 B 之间共计算出 m 个 $\mathrm{Sim}(a,b)$，并设置相应的权值 $w^k_{attribute}$。概念 A 和概念 B 基于属性的相似度计算公式为：

$$\mathrm{Sim}_{attribute}(A,B)=\frac{\sum_{k=1}^{m}w^k_{attribute}\mathrm{Sim}(a,b)}{\sum_{k=1}^{m}w^k_{attribute}} \tag{3}$$

另外，由于一个概念可能有多个属性，每个属性对概念的描述程度和作用也各不相同。如果每个属性都考虑，则计算量会大大增加。所以在计算属性相似度时，可以先依据机器学习方法计算出属性的信息增益[7]，并以此为依据来确定各个属性的优先级。最后，只选取几个信息增益大的属性进行相似度的计算，这样可以减少计算量。

3.4　基于概念关系计算相似度

本体中的概念之间都存在一定的关系。和概念属性的理论依据一样，关系有关系名称、关系类型、关系实例数据等要素，因此判断两个关系是否相似主要从这三个要素的相似度进行考虑。

设概念 A 的关系为 r，概念 B 的关系为 s，两个关系间的相似度记为 $\mathrm{Sim}(r,s)$。关系相似度计算公式如下：

$$\mathrm{Sim}(r,s)=w_1\mathrm{Sim}_{name}(r_{name},s_{name})+w_2\mathrm{Sim}_{data}(r_{data},s_{data})+w_3\mathrm{Sim}_{ins}(r_{ins},s_{ins})$$

设概念 A 和概念 B 之间共计算出 n 个 $\mathrm{Sim}(r,s)$，并设置相应的权值。概念 A 和概念 B 基于关系的相似度计算公式为：

$$\mathrm{Sim}_{relation}(A,B)=\frac{\sum_{k=1}^{n}w^k_{relation}\mathrm{Sim}(r,s)}{\sum_{k=1}^{n}w^k_{relation}} \tag{4}$$

3.5　信息增益计算

在数据挖掘的判定树归纳方法中，信息增益（information gain）可以度量选择测试属性。这种信息理论方法使得对一个对象分类所需的期望测试数目达到最小，也就是对概念实例的分类能最有效。因此可以利用属性的信息增益来评价属性对概念所起的作用。文献[9]中信息增益的计算公式为：

$$E(k)=\sum_{i=1}^{m}p_i*\log_m(1/p_i)=-\sum_{i=1}^{m}p_i*\log_m(p_i) \tag{5}$$

其中 K 表示实例集，$E(K)$ 表示 K 的信息量。p_i 为类 k_i 在实例集 K 中的比率，m 为实例集的类别数。

记属性 X_j 的属性值个数为 $|X_j|$，因此可以将实例集划分为 $|X_j|$ 个子集 s_i，$i=1,2\cdots|X_j|$，则属性 X_j 的熵 $E(X_j)$ 为：

$$E(X_j)=\sum_{i=1}^{|X_j|}w_i*E(S_i) \tag{6}$$

w_i 为权重系数，其值为：$w_i=\frac{S_i\text{ 中的实例数}}{K\text{ 中的实例数}}$

属性 X_j 的信息增益定义为：$gain(X_j)=E(K)-E(X_j)$

3.6　寻找概念的候选概念集

计算两个本体 A 和 B 中概念的相似度时，本体中的每一对概念都被考虑在内。如果本体 A 中含有 m 个概念，本体 B 中含有 n 个概念，那么就要计算 $m\times n$ 次相似度，也就是每对概念之间的相似度都要计算出来，并形成 $m\times n$ 维的相似矩阵，因此计算量很大。有的两个概念根本就不相似，也就是它们的相似度为 0，所以计算它们的相似度是不必要的，大大增加了计算的空间复杂度和时间复杂度。因此计算时可对概念对的数量进行约减，以减少计算量。本文主要根据两个概念名称的基距离（basic distance）进行判断。

文献[7]中提出的两个节点间的基距离定义为：

$$dist(N_1,N_2)=1-\frac{2m}{n_1+n_2} \tag{7}$$

其中，n_1,n_2 分别表示节点 N_1 在本体 O_1、节点 N_2 在本体 O_2 中的词的个数，m 为其中重叠的词的个数。

这样可以计算出本体 O_2 中与本体 O_1 中概念 A 最相似的 N 个概念，即基距离小于某个阈值的 N 个概念，记为 $B[1,\cdots,n]$，这样可以得到与 A 进行相似性比较的后选概念集为：

$Candidate(A)=B[1,\cdots,n]$

3.7　概念相似度的综合计算

把基于名称策略、基于属性、基于关系计算得到的概念相似度进行合并，得到最后的概念相似度 $\mathrm{Sim}(A,B)$，由公式

（下转第 146 页）

可以通过设定参数 α 和 β 来控制所提取决策规则的强弱性，当 $\alpha=1$ 且 $\beta\neq0$ 时，所提取的规则就是完全确定的强决策规则，这时规则的容错性能就比较差。

最后，通过实例分析对算法的有效性进行了验证。从销售量信息决策表中提取的规则可以看出，该方法所提取出的规则是比较合理的，所提取出的规则可以很好地对不一致信息进行分类。

参考文献

[1] Pawlak Z. Rough classification[J]. International Journal of Man-machine Studies, 1984, 20:469-483

[2] Pawlak Z. Rough sets and Decision Analysis[C]. Fifth IIASA Workshop on Decision Analysis and Support, 1998:123-127

[3] 印勇，曹长修，张邦礼. 基于粗糙集理论的分类规则发现[J]. 重庆大学学报，2000，23(1)：63-65

[4] Pawlak Z. Rough sets. International Journal of Computer and information Science, 1982, 11: 341-356

[5] Kryszkiewicz M. Rules in incomplete information system[J]. Information Sciences, 1999, 113:271-292

[6] Kryszkiewicz M. Rough set approach to incomplete information system. Information Sciences[J], 1998, 112: 39-49

[7] 黄兵，魏大宽，周献中. 不完备信息系统最大分布约简及规则提取算法[J]. 计算机科学，2004，32(8A)：53-56

[8] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法. 北京：科学出版社，2001

[9] 王基一，许黎明. 概率粗糙集模型[J]. 计算机科学，2002，28(8A)：76-78

[10] 谢宏，程浩忠，牛东小. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报，2005，28(9)：1570-1574

(上接第 124 页)

(2)-(4)得到公式(5)。

$$\mathrm{Sim}(A,B)=w_{name}\,\mathrm{Sim}_{name}(A,B)+w_{attribute}\,\mathrm{Sim}_{attribute}(A,B)+w_{relation}\,\mathrm{Sim}_{relation}(A,B) \quad (8)$$

其中，w_{name}，$w_{attribute}$，$w_{relation}$ 分别为概念名称相似性、概念属性集合相似性、相关概念相似性的权重，$w_{name}+w_{attribute}+w_{relation}=1$，$w_{name}\neq0$。

4 实验及分析

为了对以上方法进行评估，作者在两个本体数据上作了实验(表 1)。它们都描述了 Cornell 大学和 Washington 大学的课程体系。

表 1

	Ontologies	Concepts	Number of instances
Course Catalog Ⅰ	Cornell	34	1526
	Washington	39	1912
Course Catalog Ⅱ	Cornell	176	4360
	Washington	166	6975

根据公式(2)-(4)可以得出表 2。

表 2

Ontologies	sim_{name}	$\mathrm{sim}_{attribute}$	$\mathrm{sim}_{relation}$
Course Catalog Ⅰ Course Catalog Ⅱ	0.46	0.92	0.87

概念相似度的综合计算：

本文把概念相似度看成基于名称、属性和关系相似度计算的综合，分别取权值 $w_{name}=0.4$，$w_{attribute}=w_{relation}=0.3$。由综合公式(5)得：

$$\mathrm{Sim}(C,W)=0.4\times0.46+0.3\times0.92+0.3\times0.87=0.721$$

对表(1)中的数据我们用 ACAOM[2] 方法进行计算可得到。

$$\mathrm{Sim}(C,W)=0.5\times0.46+0.5\times0.96=0.710$$

时间复杂度分析：

影响复杂度的关键因素有生成候选映射空间的复杂度、候选映射对的个数、相似度度量的复杂程度。如果本体中的实体个数为 n，在不做任何改善的情况下，可能的候选映射对的个数为 $n*n$。为了提高映射效率，对候选映射对的选择采用以下策略：对基距离排序，选择基距离较小的作为候选映射对。所以生成后选概念集在最坏情况下所需时间：$T_{candidate}=O(n\log(n))$，在候选概念集中，对每个概念来说，原本有 n 个概念要与之比较，现在只固定 k 个，所以比较的概念 $N_{compare}$ 最多为 $n*k$ 个，所有相似度计算的最大复杂度为 $T_{compare}=m*O(\log^2 n)$，而对相似度的综合时复杂度为 $T_{integrate}=O(n)$。在使用候选概念集和信息增益的情况下为：$O(n\log(n))+O(n)*m*O(n*n)+O(n)=O(n\log(n))$，在直接使用所有相关概念的情况下为：$O(n\log(n))+O(n)*m*O(\log^2(n))+O(n)=O(n\log^2(n))$。

结束语 针对本体映射中概念相似度计算量大、计算精度不高等问题，本文给出了一种基于语义 Web 的本体概念相似度的计算方法，通过给定一定权值提高概念相似度的精确度，并引入信息增益和候选概念集。从相似度计算结果可得出该方法的精度比 ACAOM 方法高，从时间复杂度分析可以得出其复杂度要比 ACAOM 方法的复杂度小。

参考文献

[1] Doan A, Madhavan J Y, Domingos P, et al. Learning to map between ontologies on the semantic web// Proceedings of the World Wide Web Conference. 2002: 662-673

[2] 黄烟波，张红宇，李建华，等. 本体映射方法研究. 计算机工程与应用，2005：27-33

[3] 邓志鸿，唐世渭，张铭，等. Ontology 研究综述. 北京大学学报，2002(5)：730-738

[4] Zhou Chunguang, Wang Ying. A Composite Approach for Ontology Mapping//Proceeding of the 23rd British Database Conference. 2006

[5] Sekine S, Sudo S, Ogino S. Statistical Matching of two ontologies//Proceedings of the SIGLEX99: Standerdizing Lexical Resources. Maryland, USA, 1999: 69-73

[6] Pedersen T, Patwardhan S, Michelizzi S. WordNet: Similarity - Measuring the Relatedness of Concepts//Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics[C]. [s. l.], 2004

[7] Qian Peng-fei, Zhang Shensheng. Ontology Mapping Approach Based on Concept Partial. IEEE, 2006(7): 4107-4112

[8] Ehrig M, Sure Y. Ontology Mapping-An Integrated Approach. ESWS, 2004: 76-91

[9] 钟宁，尹旭日，陈世福. 基于信息增益的最佳属集性发现方法. 小型微型计算机系统，2002(4)：444-446