一种改进的本体相似度计算方法

朱珍元,郑 诚

(安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:本体映射的关键技术是本体相似度计算。本文基于已有的V-Doc(虚拟文档)技术提出一种新的NV-Doc本体相似度计算方法,其中不仅用到了本体中实体自身以及其第一层相邻节点的信息,而且还充分利用了第二层相邻节点的信息。

关键词:语义网;本体映射;虚拟文档;本体相似度

中图分类号:TP301 **文献标识码**:A **文章编号**:1674-7720(2011)01-0007-03

An improved ontology similarity calculation

Zhu Zhenyuan, Zheng Cheng

(Department of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: How to calculate ontology similarity is a critical technique in ontology mapping, this paper presents a new NV-Doc algorithm that based on V-Doc (virtual document) to calculate ontology similarity. It makes full use of information not only about entity itself in ontology and one-step neighbors, but also two-step neighbors.

Key words: semantic Web; ontology mapping; virtual document; ontology similarity

本体是共享概念化的明确具体规范,随着语义网的发展,本体的应用越来越多。用RDF[1]或OWL[2]书写的Web本体在语义网的出现和应用方面起到了很大作用,本体的数量也与日俱增。

Web的分布式特点使得大量的本体由不同组织开发,并且在很大程度上覆盖相同或者相交的领域,因此Web本体之间存在一定的相似性,但相关领域的不同本体之间也存在很大的异构性。

解决本体异构问题的最好方法是本体映射。本体映射的目的是架起异构本体之间的桥梁,在使用不同本体的Web应用之间建立互操作,从而实现语义网环境下数据的集成与管理。而本体映射的关键技术是本体的相似度计算,即计算两个不同本体中实体之间的相似度,当相似度值大于某个给定的阈值时,可以认为这两个实体之间存在着一定的语义关系。

目前,关于本体相似度计算方法的自动化程度不高,而且不能充分利用本体的各种描述信息。已有的V-Doc技术能够较好地解决这两方面的问题,但也存在一些不足。

基于虚拟文档的本体相似度计算方法V-Doc[3]将本体看成一个有向图,图中的每个节点对应本体中的一个实体,为每个实体自动建立虚拟文档,充分利用了节点自身和邻接节点的描述信息。但该方法也存在不足:节点的特征不仅与邻接节点有关,而且还与邻接节点的邻接节点信息有关,即实体的描述信息还应该考虑节点的第二层邻接节点的信息。针对其不足,本文提出一种新的基于虚拟文档的本体相似度计算方法NV-Doc。

1 V-Doc简介

1.1 虚拟文档的构建

虚拟文档是为了描述概念特点而建立起来的文档,为每一个节点构建虚拟文档,充分利用节点自身和邻接节点的描述信息。

定义1 (URIrefs描述):假设e是一个URIref,对e的描述通过与其有关的名字、标签、注释和其他自然语言描述信息组成,其定义[3]为:

$$\begin{aligned}Des(e)=&\alpha_1\times\text{collection of words in the local name of e}+\\&\alpha_2\times\text{collection of words in the rdfs:lable of e}+\\&\alpha_3\times\text{collection of words in the rdfs:comment of e}+\\&\alpha_4\times\text{collection of words in other annotations of e}\end{aligned}\quad(1)$$

其中α_1、α_2、α_3、α_4是在区间[0,1]内固定的实数,分别表

示名字、标签、注释和其他自然语言描述信息在对 e 描述中的权重，其具体值可以通过实验确定。

定义 2（虚拟文档）：假设 e 是一个 URIref，e 的虚拟文档 VD(e)定义为：

$$VD(e)=Des(e)+\gamma_1\times\sum_{e'\in sn(e)}Des(e')+\gamma_2\times\sum_{e'\in on(e)}Des(e') \quad (2)$$

其中，$sn(e)$代表关于实体 e 的子概念集合，$on(e)$代表关于实体 e 的父概念集合，γ_1、γ_2 是在[0，1]内固定的实数，其具体值需要通过实验确定。

1.2 相似度计算

本体中每一个实体（节点）的描述信息（语言学特征）通过该节点的虚拟文档表示。因此，两个本体中实体的相似度可通过计算与之对应的两虚拟文档之间的相似度而得到，即虚拟文档之间的相似度就是实体之间的相似度。虚拟文档之间的相似度通过在信息检索领域应用广泛的向量空间模型 VSM(Vector Space Model)[4]方法计算。将两个待匹配的虚拟文档用向量空间中的一个向量表示，当然在相似度计算之前还要对文档进行预处理，如分词、去除停用词、提取词干等。向量空间模型中，关键词的权重使用 TF/IDF 技术[5]表示。由此可以得到一个 $N\times W$ 的矩阵 X，其中 N 是虚拟文档的个数，W 表示所有虚拟文档中 token 的总数。可以通过矩阵与其倒置矩阵的积得到虚拟文档之间的相似矩阵，最后规范化相似矩阵，使相似度值在[0，1]区间内。规范化后所得矩阵即为虚拟文档之间的相似度矩阵，每个值也代表了两个虚拟文档之间的相似度，从而得到与之对应的两实体之间的相似度。

2 NV-Doc

2.1 改进的虚拟文档

为 RDF 图中每一个节点构建虚拟文档，不仅用到节点自身以及相邻第一层的邻居节点信息，还用到节点第二层的邻接节点信息。

定义 3（改进的虚拟文档）：假设 e 是一个 URIref，e 的虚拟文档 NVD(e)的表示方程为：

$$NVD(e)=Des(e)+\gamma_1\times\sum_{e'\in sn(e)}Des(e')+\gamma_2\times\sum_{e'\in on(e)}Des(e')+$$

$$\gamma_3\times\sum_{\substack{e'\in sn(e'')\\ e''\in sn(e)\cup on(e)}}Des(e')+\gamma_4\times\sum_{\substack{e'\in on(e'')\\ e''\in sn(e)\cup on(e)}}Des(e') \quad (3)$$

其中，γ_1、γ_2、γ_3、γ_4 是在[0，1]内固定的实数，其具体值需要通过实验确定。

2.2 简单示例

假设一个简单的本体片段模型如图 1 所示。

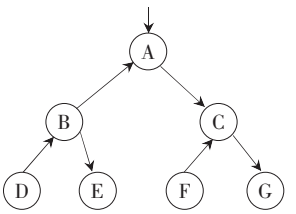

图 1 本体片段模型

按照式(3)得到节点 A 的虚拟文档为：

$$NVD(A)=Des(A)+\gamma_1\times\sum_{e'\in sn(A)}Des(e')+\gamma_2\times\sum_{e'\in on(A)}Des(e')+$$

$$\gamma_3\times\sum_{\substack{e'\in sn(e'')\\ e''\in sn(A)\cup on(A)}}Des(e')+\gamma_4\times\sum_{\substack{e'\in on(e'')\\ e''\in sn(A)\cup on(A)}}Des(e')=$$

$$Des(A)+\gamma_1\times Des(C)+\gamma_2\times Des(B)+$$

$$\gamma_3\times(Des(E)+Des(G))+\gamma_4\times(Des(D)+Des(F)) \quad (4)$$

3 实验结果及分析

3.1 实验数据

实验数据选用基于 KAON2 的开源资源 Framework for Ontology Alignment and Mapping 中所提供的 Test Ontologies and Alignments。从中选用本体规模较小的 russia1.owl 和 russia2.owl 作为数据源，其中 russia1 中共有 49 个节点，russia2 中共有 51 个节点。进一步的实验选用数据集 OAEI 2005 benchmark tests 中的五组规模稍大的本体作为数据源。本文两次实验中各参数的取值不变：α_1、α_2、α_3、α_4 的值分别为 1.0、0.5、0.25、0.25，参数 γ_1、γ_2、γ_3、γ_4 的值分别取 0.1、0.1、0.05、0.05。各参数的取值借鉴 Falcon-OA[6]系统在程序中所给的参数值。对于实体的描述，第一层邻接节点一般比第二层邻接节点更有影响力，所以 γ_3、γ_4 分别取 0.05、0.05，比 γ_1、γ_2 的值 0.1、0.1 都小是有道理的。

本文采用查准率和查全率的综合评估函数以及运行时间作为评价标准对实验结果进行评估。

$$查准率=\frac{已找出且正确的匹配对数}{已找出的匹配对总数}\times 100\%$$

$$查全率=\frac{已找出的匹配对总数}{实际存在的匹配对总数}$$

$$综合评估函数\ F=\frac{2\times查准率\times查全率}{查准率+查全率}$$

3.2 实验结果及分析

本文主要的改进之处是提出新的算法来构建本体中实体的虚拟文档，虚拟文档间的相似度计算也是通过描述的方法实现，初步实验结果如表 1 所示。

表 1 初步实验结果

	查准率/%	查全率/%	运行时间/s
V-Doc	0.83	0.75	7
NV-Doc	0.94	0.80	9

初步实验结果：表明改进的算法虽然在运行时间上有所延长，但查准率和查全率都有所提高，而且这种时间消耗不是很大。

其次，为了再一次验证 NV-Doc 较 V-Doc 的可行性，对数据集 OAEI 2005 benchmark tests 中的五组本体

进行实验，最后得到的实验结果如图 2、图 3 所示。

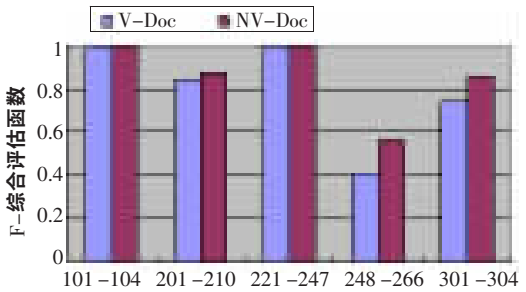

图 2 综合评估对比

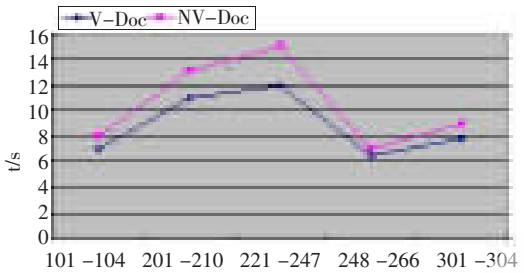

图 3 运行时间对比

再次实验结果表明，NV-Doc 能够取得比 V-Doc 更好的查全率和查准率，虽然在效率方面不及 V-Doc，但从整体上来看，效率上的部分损失换来更好的查准率和查全率也是值得的。

本文针对计算本体中实体相似度存在的问题提出改进方法，充分利用实体自身和实体的第一层及第二层邻接节点的描述信息(即实体的语言学上的特征)。实验结果分析表明，改进后的算法在查准率和查全率方面优于原先的算法。下一步的研究工作是：一方面将此方法和其他计算本体相似度的方法有效结合，从而更有效地实现本体映射；另一方面是减少运行时间，提高效率。最后还要充分利用本体其他的描述信息，如本体的属性、关系、实例等。

参考文献

[1] KLYNE G, CARROLL J J. Resource description framework (RDF): concepts and abstract syntax.//W3C Recommendation 10 February 2004. Latest version is available at http://www.w3.org/TR/rdf-concepts/.

[2] Patel-Schneider P F, HAYES P, HORROCKS I. OWL web ontology language semantics and abstract syntax. W3C Recommendation 10 February 2004. Latest version is available at http://www.w3.org/ TR/owl-semantics/.

[3] QU Yuzhong, HU Wei, CHENG Gong. Constructing virtual documents for ontology matching [C]//Proceedings of the 15th International Conference on W orld W ide W eb.Edinburgh, Scotland: [S.n.], 2006.

[4] VIJAY V, RAGHAVAN S K, WONG M. A critical analysis of vector space model for information retrieval. JASIS, 1986: 37(5), 279-287.

[5] SALTON G, MCGILL M. Introduction to modern information retrieval[M]. McGraw-Hill Book Company, 1984.

[6] Hu Wei, Qu Yuzhong. Falcon-AO: a practical ontology matching system [C]. Web Semantics: Science, Services and Agents on theWorldWideWeb, 2008: 237-239.

(收稿日期：2010-08-12)

作者简介：

朱珍元，女，1985 年生，硕士研究生，主要研究方向：数据库技术与语义 Web。

郑诚，男，1964 年生，副教授，主要研究方向：数据挖掘、数据库技术与语义 Web。