

文章编号:1001-9081(2011)04-1067-03

doi:10.3724/SP.J.1087.2011.01067

改进的本体匹配算法

张玉芳,李川,熊忠阳

(重庆大学计算机学院,重庆 400044)

(lanfker@gmail.com)

摘要:传统的利用本体结构信息对本体做匹配的方法,并未充分利用本体的树形结构特点,致使整个本体匹配的匹配过程具有大量的冗余计算。因此,提出一种改进的基于本体树形结构的本体匹配算法 TARA。该方法首先严格地以本体的树形结构为依据进行本体匹配,然后通过二次匹配来解决由于严格按照树形结构进行匹配而产生的不可避免的不足。实验结果表明,TARA 方法的查全率和准确率都有较好的表现。

关键词:语义网;本体匹配;编辑距离;本体匹配竞赛;Jena

中图分类号: TP18 **文献标志码:**A

Improved ontology matching method

ZHANG Yu-fang, LI Chuan, XIONG Zhong-yang

(College of Computer Science, Chongqing University, Chongqing 400044 China)

Abstract: The traditional ontology matching methods that use the ontology's structure to find the matches do not really make good use of the ontology's structural feature, which leads to considerable computation redundancies during the entire matching process. Therefore, a modified method named TARA was proposed to improve the matching process in this paper. The method firstly casted matching process by strictly using the ontology's structural information, and then a re-match process was applied to overcome the inevitable defect that caused by the matching process before. The experimental results show that the method has good performances in both recall and precision.

Key words: semantic Web; ontology matching; edit distance; Ontology Alignment Evaluation Initiative (OAEI); Jena

0 引言

自1998年9月,万维网之父Tim Berners-Lee提出语义网的概念以来,语义网因其具有现有网络无法比拟的优越性而引起了科研工作者的注意。本体(Ontology)作为语义网中的知识表现(Knowledge Representation)形式,成为了研究的热点。根据语义网的构想,语义网的智能特性必然需要整合语义网络中的现有知识,也就是利用各个相关本体,通过推理提供给用户需要的答案。而这个知识的整合过程作为关键的一步,需要有本体的匹配工作做基础。本体匹配(Ontology Matching)是指将本体中的概念相互比较,建立它们之间的映射关系。通过本体匹配,本体之间就能依靠概念之间的映射关系达到相互“理解”,从而使得基于多个本体之间的推理得以实现。

1 传统的本体匹配方法的问题

目前利用树形结构对本体做匹配的方法存在的问题大致如下。

1)自下而上的本体匹配顺序。自下而上的匹配顺序顾名思义就是在本体匹配的过程中,优先匹配本体树中的叶子节点,再匹配它们的上一级节点,依此类推。而在决定叶子节点的匹配过程中,需要对两个本体文件的所有叶子节点做笛卡尔乘集式的相似度运算。这样做有两个缺点,一是在无任何筛选规则的情况下计算所有的叶子节点的相似度,导致本

体匹配过程产生大量冗余计算;另外还忽略了一个重要现象,即小的概念总是隶属于一个大的概念,在大概念不一致的情况下,即使小概念相似,仍然不能被确认为是一个正确的匹配,也正是因为这个原因,导致本体匹配中存在很多错误匹配,致使精确率下降。

2)忽略概念划分粒度差异。由于本体类似于现有网络的网页,大多由用户自由创建,那么在概念的划分粒度上存在差异是不可避免的。比如对人的划分可以是:人→黄种人→中国人→北京人,也可以是:人→中国人→北京人。两种不同的划分方式都是正确的,然而正是因为这样的差异,使得本体的树形结构存在差异,从而极有可能导致在本体匹配的过程中产生错误匹配或者不完全匹配。

目前,基于树形结构对本体做匹配的工具中,比较出名的是由伊利诺伊大学芝加哥分校 ADVIS 实验室(ADVances in Information Systems Research Laboratory)开发的工具 AgreementMaker^[1]。该工具在利用本体的树形结构信息中,提出了两个算法,分别是 DS^[2](Descendant's Similarity Inheritance)和 SSC^[2](Sibling's Similarity Contribution)。这两个算法分别从本体树的横纵两个方向出发计算概念的相似度。

2 改进的本体匹配算法

2.1 概念相似度计算简介

编辑距离方法:编辑距离就是用来计算从原串(s)转换

收稿日期:2010-10-08;修回日期:2010-11-29。

作者简介:张玉芳(1965-),女,上海人,教授,主要研究方向:数据挖掘、网络入侵检测、并行计算; 李川(1986-),男,重庆合川人,硕士研究生,主要研究方向:本体匹配、自然语言处理; 熊忠阳(1962-),男,重庆人,教授,博士生导师,博士,主要研究方向:数据挖掘、网格技术、并行计算。

到目标串(t)所需要的最少的插入、删除和替换的数目。将编辑距离运用于相似度计算的公式如下^[3]:

$$SS(s1, s2) = \frac{1}{\frac{ed(s1, s2)}{s1.len + s2.len - ed(s1, s2)}} \quad (1)$$

其中: $SS(s1, s2)$ 表示两个字符串 $s1$ 和 $s2$ 之间的相似度; $ed(s1, s2)$ 表示字符串 $s1$, 和 $s2$ 之间的编辑距离; $s1.len$ 和 $s2.len$ 分别表示字符串 $s1$ 和 $s2$ 的长度。由于本文着重关注本体的结构方面的信息, 所以并没采用特别复杂而效果显著的概念相似度计算方法, 而以简单的编辑距离方法来计算概念相似度。编辑距离代表着一类计算本体相似度的方法, 它们主要通过考查概念的字符相似度来确定概念的相似度。

基于词典 WordNet 计算概念相似度的方法^[4]:

$$SW(s1, s2) = -\lg \frac{(len(s1, s2) + 1)}{2 \times depth} \quad (2)$$

其中: $len(s1, s2)$ 为两个词所在的最短的同义词集间的路径长度, $depth$ 为分类树的高度。当两个词语完全同义时, 就有 $len(s1, s2) = 0$, $depth = 1$, 此时 $SW(s1, s2) = 1$ 。该方法代表了另一类计算概念相似度的方法, 它们通过概念的语言学特征计算相似度。

现有的本体匹配工具大多是通过结合这两类方法来取得更加准确的概念相似度。

2.2 改进的本体匹配算法

本体匹配的整个算法流程分为以下几个步骤。

1) 解析本体。解析本体的任务在于将本体文件解析为按照概念的层次结构形成的树形结构, 树的每一个节点为本体中的一个类, 或者属性, 而本体树的类节点具有共同的祖先节点 Class, 属性节点具有共同的祖先节点 Property。同时, 解析的过程中去掉类名或者属性名的公有前缀后缀。本文的实验过程中采用了由 HP 实验室开发的操作本体的开源框架 Jena^[5] 来解析本体文件。

2) 自上而下根据树形结构的匹配。在该步骤中, 以表示本体文件中的类的节点为例。

① 初始化匹配集合, 找出属于源本体的节点集合 N_s 和属于目标本体的节点集合 N_t , 其中 N_s 的元素 E_{ns} 满足如下条件: E_{ns} 的直接父节点为源本体的类根节点 ($Class_s$), 关系表示为 $Parent(E_{ns}) = Class_s$; 类似的, N_t 的元素 E_{nt} 满足条件 E_{nt} 直接父节点为目标本体的类根节点 ($Class_t$), 即 $Parent(E_{nt}) = Class_t$ 。然后, 通过匹配集合 N_s 和 N_t 中的元素即可完成本体树的第一级概念的匹配任务。

② 计算概念相似度。本文在计算概念相似度的环节中, 采用编辑距离方法。综合计算概念的 Name, Label, Comment 三个特性的编辑距离相似度而形成概念的整体相似度。具体方法如下: 假定 Se_n, Se_l, Se_c 分别为 Name, Label 和 Comment 的编辑距离相似度, 采用式(1) 计算得到, 那么整体相似度 S 的计算公式可以表示为:

$$S(n_s, n_t) = \alpha \cdot Se_n + \beta \cdot Se_l + \gamma \cdot Se_c \quad (3)$$

其中:

$$\alpha = \frac{Se_n}{Se_n + Se_l + Se_c} \quad (4)$$

$$\beta = \frac{Se_l}{Se_n + Se_l + Se_c} \quad (5)$$

$$\gamma = \frac{Se_c}{Se_n + Se_l + Se_c} \quad (6)$$

这样, 可以最大化优势项而最小化劣势项。当两个概念 n_{si} 与 n_{tj} 的相似度 $S(n_{si}, n_{tj})$ 满足如下条件, 那么算法就认为概念 n_{si} 与 n_{tj} 是一对合法的匹配: $S(n_{si}, n_{tj}) \geq Threshold$ 且 $S(n_{si}, n_{tj}) = \text{Max}(S(n_{si}, n_{t1}), S(n_{si}, n_{t2}), S(n_{si}, n_{t3}), \dots, S(n_{si}, n_{tm}))$ 。其中 m 为集合 N_t 的元素个数, $j \leq m$, $Threshold$ 为预先设置好的阈值。

在本文中对于属性节点的相似度计算与对于类节点的相似度计算也采用式(3) 进行计算。

③ 选取匹配对的子节点集合。假设概念 n_{si} 与 n_{tj} 已然被认为是一对合法的匹配, 那么在实验过程中就可以缩小匹配的范围为两棵 n_{si} 与 n_{tj} 为根节点的子树。从而可以得到新的集合 N_s, N_t , 它们的元素分别满足条件 $Parent(E_{ns}) = n_{si}$, $Parent(E_{nt}) = n_{tj}$ 。一旦确定新的 N_s 与 N_t , 再次运用概念相似度的计算方法即可取得新的 N_s 与 N_t 之间的合法匹配。通过反复执行 ③, 直到没有可以被选取的 N_s 与 N_t , 则自上而下的根据本体树形结构的本体匹配步骤结束。

3) 二次匹配。之所以引入二次匹配, 主要是由两个待匹配的本体文件对概念的划分粒度不同, 必然导致待匹配的本体的树形结构存在差异, 那么单单严格按照本体树对本体作匹配就显得不够。在二次匹配中, 算法认定在之前基于树形结构匹配中产生的匹配对为合法匹配, 然后充分利用现有的已经被确定为合法匹配的匹配对, 筛选出没有被匹配的概念, 对这些尚未被匹配的概念再做一次匹配运算。在存在源本体文件的一个概念 n_s 与目标本体的一个概念 n_t 满足: $S(n_s, n_t) \geq Threshold$ 且 n_s 与 n_t 没有被匹配, 则指定筛选范围为以 n_s 与 n_t 的祖先节点 n_s' 与 n_t' 为根节点的两棵子树, 而 n_s' 与 n_t' 满足条件: n_s' 与 n_t' 已经被认为是合法匹配。在这一过程中, 所有类节点的根节点 $Class_s$ 和 $Class_t$ 与所有属性节点的根节点 $Property_s$ 和 $Property_t$ 被认为是(虚拟的)合法匹配。然而在二次匹配中, 由于没有可靠的树形结构信息可以参考, 所以本算法将这些尚未被匹配的概念分源本体和目标本体分别归为两个集合直接计算概念相似度进行匹配。由于在之前基于树形结构的匹配过程中产生的合法匹配对所包含的所有类和属性都不属于二次匹配考虑的范畴, 所有二次匹配的运算量并不算大。

3 实验及结果分析

3.1 本体匹配竞赛简介

伴随着语义网的发展, 本体匹配领域的研究也取得了一些成果, 许多科研工作者将自己设计的本体匹配算法加以完善, 开发出了许多本体匹配工具。为了衡量各种本体匹配工具的优劣, 以便用户择优使用, 在 2005 年 10 月, 国际本体匹配竞赛^[6] (Ontology Alignment Evaluation Initiative, OAEI) 诞生, 并在加拿大举行了第一届竞赛。现在, OAEI 已经作为国际语义网会议 (International Semantic Web Conference, ISWC) 的一部分。每一届的 OAEI 都会提供测试数据, 以 OAEI 提供的 benchmark 的测试数据为例, benchmark 里面的测试数据由专家做了精心修改, 对本体匹配工具的各方面性能进行考查。

每一届本体匹配竞赛之后, OAEI 都会将参赛的工具的匹配结果公布出来, 以供科研工作者参考、学习。

3.2 实验结果及分析

本文中的实验采用 Java 为程序开发语言, 在 Eclipse 开发平台上利用开源框架 Jena 解析和操作本体文件, 利用国际上

广泛采用的 Alignment API^[7]产生本体匹配文件并与标准匹配文件比较,做评估。测试数据采用 OAEI2008 的 benchmark 提供的测试数据。对比实验数据直接从 OAEI 网站中获取现有的参赛工具的匹配结果。评价指标采用查准率 P 、查全率 R 和 $F\text{-Measure}$ 值。

$$P = \text{正确的匹配对} / \text{找到的匹配对}$$

$$R = \text{正确的匹配对} / \text{应有的匹配对}$$

$$F\text{-Measure} = 2PR / (P + R)$$

为了充分体现树形结构在本体匹配中的作用,对比实验首先是与只使用了编辑距离的本体匹配工具 EDNA 进行对比。对比结果如表 1 所示。

表 1 EDNA 与本文方法的比较

匹配方法	精确率	查全率	$F\text{-Measure}$
EDNA	0.643 824	0.832 059	0.725 937
本文方法	0.959 941	0.940 294	0.950 016

由表 1 可以看到在利用了本体树形结构信息之后,本体匹配的效果得到了很大的提升。

同时,对比实验还包括与不完全利用了本体树形结构信息的 AgreementMaker 的匹配结果做了比较。比较结果如表 2 所示。

表 2 AgreeMaker 与本文方法的比较

匹配方法	精确率	查全率	$F\text{-Measure}$
AgreeMaker	0.984 412	0.916 765	0.949 385
本文方法	0.959 941	0.940 294	0.950 016

由表 2,可以看到,与 AgreementMaker 相比,本文的方法精确率偏低,而查全率较高。在通过 $F\text{-Measure}$ 指标的衡量下,本文的方法只是略高于 AgreementMaker。然而,与本文的方法不同的是,AgreementMaker 内部整合了多种本体匹配算法,如 BSM、PSM^[8]、DSI、SSC 等,它的工作原理就是将各种算法运行的结果进行迭代匹配,综合多中算法的匹配结果,能够被多种算法同时发现的匹配成为正确的匹配的可能性自然而然加大,由此增加了 AgreementMaker 的精确率。于此同时,本文的方法将会耗时更少,原因如下:1) AgreementMaker 自下而上匹配顺序导致了冗余计算;2) 多种算法的迭代匹配必然增加计算的时间开销;3) SSC 方法的原理决定其计算量必然大于本文的基于子树范围的匹配原则。因为,SSC 在选取匹配对象时,将会对源本体树的所以第 i 层子孙节点与目标本体树的所有的第 i 层子孙节点做匹配,而忽略这些子孙

节点是否具有相同的父亲节点,也就是 SSC 的节点选取方法不满足以子树为范围的选取原则。所以对这些节点做笛卡尔乘积式的相似度计算必然存在过多的冗余计算。经过以上分析,可以看出本文的方法在保证了匹配效果的前提下可以有效地减少计算量,提高了本体匹配的效率。

4 结语

本文的方法在充分利用本体的树形结构特征的同时,根据本体的 Name, Label, Comment 采用编辑距离计算相似度,通过实验证明了方法的可行性。进一步的工作将会从本文中并未涉及到本体匹配中运用十分成功的基于 RDF 图的匹配^[9]以及基于结构信息的反馈机制,如 SF 方法^[10],等方面入手,完善本文提到的方法。

参考文献:

- [1] CRUZ I F, ANTONELLI F P, STROE C, et al. Using the AgreementMaker to align ontologies for the OAEI Campaign 2009 [EB/OL]. [2010-09-08]. <http://www.cs.uic.edu/~ifc/webpapers/Cruz-OAEI-2009-new.pdf>.
- [2] CRUZ I F, SUNNA W. Structural alignment methods with applications to geospatial ontologies [EB/OL]. [2010-09-07]. <http://www.cs.uic.edu/~ifc/webpapers/gis-alignments-publish.pdf>.
- [3] 简宁胜.一个本体匹配工具的设计与实现[D].南京:东南大学,2006.
- [4] 郑松元,刘大有.本体匹配算法的研究[D].长春:吉林大学,2009.
- [5] HP Labs Semantic Web Programme, Jena [EB/OL]. [2010-09-08]. <http://jena.sourceforge.net/>.
- [6] OAEI [EB/OL]. [2010-09-08]. <http://oaei.ontologymatching.org/>.
- [7] EUZENAT J, LIG I. Alignment API [EB/OL]. [2010-09-08]. <http://alignapi.gforge.inria.fr/>.
- [8] CRUZ I F, ANTONELLI F P, STROE C. Efficient selection of mappings and automatic quality-driven combination of matching methods [DB/OL]. [2010-11-13]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.156.5563&rep=rep1&type=pdf>.
- [9] 王颖,刘群,王普强,等.一种基于 RDF 图的本体匹配方法[J].计算机应用,2008,28(2):460-462.
- [10] MELNIK S, GARCIA-MOLINA H, RAHM E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching [C]// ICDE 2002: 18th International Conference on Data Engineering. San Jose, CA: [s. n.], 2002: 117-128.
- [5] JARRAR M. Towards automated reasoning on ORM schemes mapping ORM into the DL-RDF description logic [C]// ER 2007: Proceedings of the 26th International Conference on Conceptual Modeling, LNCS 4801. Berlin: Springer-Verlag, 2007: 181-197.
- [6] JARRAR M. Mapping ORM into the SHOIN/OWL description logic [C]// ORM'07: Proceedings of the International Workshop on Object-Role Modeling, LNCS 4805. Berlin: Springer-Verlag, 2007: 729-741.
- [7] NGUYEN T, THANH N. Modeling ORM schemas in description logics [M]// Complex Systems Concurrent Engineering: Collaboration, Technology Innovation and Sustainability. Berlin: Springer-Verlag, 2007: 547-555.
- [8] WINTRAECKEN J J V R. The NIAM information analysis method: theory and practice [M]. Berlin: Springer, 1990.

(上接第 1066 页)

参考文献:

- [1] HALPIN T. Object-Role Modeling (ORM/NIAM) [M]// Handbook on Architectures of Information Systems. Berlin: Springer-Verlag, 2006: 81-103.
- [2] HALPIN T. ORM 2 [C]// On the Move to Meaningful Internet Systems 2005: OTM Workshops, LNCS 3762. Berlin: Springer-Verlag, 2005: 676-687.
- [3] JARRAR M, MEERSMAN R. Ontology engineering - the DOGMA approach [C]// Advances in Web Semantics I: Ontologies, Web Services and Applied Semantic Web, LNCS 4891. Berlin: Springer-Verlag, 2009: 7-34.
- [4] OMG. Semantics of Business Vocabulary and Business Rules (SBVR) [S]. OMG, 2008.

- [5] JARRAR M. Towards automated reasoning on ORM schemes mapping ORM into the DL-RDF description logic [C]// ER 2007: Proceedings of the 26th International Conference on Conceptual Modeling, LNCS 4801. Berlin: Springer-Verlag, 2007: 181-197.
- [6] JARRAR M. Mapping ORM into the SHOIN/OWL description logic [C]// ORM'07: Proceedings of the International Workshop on Object-Role Modeling, LNCS 4805. Berlin: Springer-Verlag, 2007: 729-741.
- [7] NGUYEN T, THANH N. Modeling ORM schemas in description logics [M]// Complex Systems Concurrent Engineering: Collaboration, Technology Innovation and Sustainability. Berlin: Springer-Verlag, 2007: 547-555.
- [8] WINTRAECKEN J J V R. The NIAM information analysis method: theory and practice [M]. Berlin: Springer, 1990.