

文章编号:1005-0523(2013)03-0082-07

图解析方式的复合本体映射策略研究

凌仕勇¹, 龚锦红²

(华东交通大学 1. 软件学院; 2. 电子与电气工程学院, 江西 南昌 330013)

摘要:针对本体复杂上下文关系,考虑不同输入模式,构建通用环境下的本体解析图描述。在结合本体结构基础上,建立多层本体结构图描述关系;引入上下文关联的相似度传播建立结构传播因子和实例传播算法;引入快速匹配算法用于本体的快速映射;最终提出一种复合本体映射策略,利用重用迭代得到本体的映射结果。最后通过算法复杂度的分析和一些对比测试证明了策略的可行性和有效性。

关键词:复合本体映射;本体图解析;相似度传播;快速匹配算法

中图分类号:TP311

文献标志码:A

本体映射在许多应用领域都起到重要的作用,如模式集成、语义WEB、数据仓库、电子商务、智能体通信、WEB服务组合、目录映射等。很多前期研究往往侧重于本体的概念、属性、实例等特定模式,而忽视本体上下文结构的分析。近来,出现了大量基于英文语料库 Wordnet 和中文语料库 Hownet 的词汇相似研究,这些研究语料的词汇相似度计算是构成本体概念相似度计算的基础。但是,对于本体本身的研究主要集中于单纯的概念、属性以及实例,如QOM^[1],ASMOM^[2],RiMOM^[3]。COMA是由德国莱比锡大学开发的一套集本体解析,匹配,测试和评估的开源映射系统(<http://sourceforge.net/projects/coma-ce/>),COMA解析各种格式输入文件(XML,RDF,OWL)成一种内部结构,通过这种结构与匹配算法,输入输出,数据库(默认mysql,可以自定义算法接入系统接口)进行交互。提供一套工作流用于定义输入文本的格式,匹配算法,相似度计算的算法(外部接口库 align-4.2-ontowrap.jar),输出文件的格式(xml,html)。新版COMA++^[4]支持不同的输入如XML Schema,RDF,OWL,引入了概念间的关系,但本身没有引入基于语料库的词汇计算,不支持本体结构的相似度传播,并且在映射时为了顾及一般化的多对多映射而采用单向或双向完全遍历的方法,使得复杂度大为增加。S-Match^[5]是由意大利特伦托大学 Fausto Giunchiglia 教授主持开发的一套开源语义映射算法(<http://sourceforge.net/apps/trac/s-match/wiki/>),S-Match将待映射的两个文件解析成树形结构,然后计算两个树节点之间的语义关系,进而找出映射关系。其语义计算分2个步骤:首先解析两树的标签属性,类似于本体的概念名称,通过元素级别的匹配库计算标签概念的语义关系;其次解析两树的节点,类似于本体的概念属性,通过节点级别的匹配库计算节点间的语义关系。其词汇计算的核心是通过 WordNet 词汇库进行语义相似度的计算。S-Match充分考虑了语义库的支持和计算方法,但对应本体结构和映射策略并没有给予过多的支持。

对于输入模式的解析上,除了需要考虑本体的概念,属性外,还需要考虑本体之间的上下文结构关系,本体所拥有的实例。本文将本体或其它的输入模式解析成一种图描述的结构,这种结构不依赖于任何特

收稿日期:2013-02-25

基金项目:江西省自然科学基金项目(20122BAB201040);江西省教育厅青年科学基金项目(GJJ11115);华东交通大学科研项目(12RJ04)

作者简介:凌仕勇(1974—),男,高级工程师,硕士,主要研究方向为语义WEB、并行运算、网络安全。

一个固定比例的相似度传播值。GMO^[7]则考虑全局的相似度传播。两者都没有考虑概念间传播度的差异性,如图1中的节点1和节点5,节点1的概念空间比节点5大,所以对于节点1的子类相似的可能性要远小于节点5的子类相似的可能性,所以不同概念的相似度传播是有差异的。本文采用文献[8]提出的语义传播因子 SF 的方法,由定义1的语义传播因子 SF 所示, $C(c)$ 为节点 c 的子节点集合, $P(c)$ 为节点 c 的父节点集合。当节点对的父节点对的信息量越大,节点对匹配的可能性越大,其相似度传播因子也越大;当节点对的子节点对的信息量越小,节点对匹配的可能性越大,其相似度传播因子也越大。

定义1 语义传播因子 SF

$$SF = \begin{cases} e^{-\frac{(ic(c_1) + ic(c_2))}{2}} & c_1 \in C(c) \cap c_2 \in C(c') \\ e^{\frac{(ic(c_1) + ic(c_2))}{2} - 1} & c_1 \in P(c) \cap c_2 \in P(c') \end{cases}$$

另外,考虑到概念图中的节点,一般来说,概念的实例主要集中在叶子节点,非叶节点很少拥有大量的实例,所以在拥有大量实例的本体中,可以通过对概念实例的评估来增强概念相似度匹配的准确性。应用到相似度传播中,采用“向上传播”的方法^[4],将概念实例相似度从叶子节点传播到父节点。如定义2所示,利用 e_1 所有实例的相似度最大值和 e_2 所有实例的相似度最大值来计算实例相似度传播值,其中 n 和 m 分别为 e_1 和 e_2 实例的个数。

定义2 实例相似度传播

$$sim(e_1, e_2) = \frac{\sum_{i=1}^n \max_{j=1, \dots, m} (sim(inst_i, inst_j)) + \sum_{j=1}^m \max_{i=1, \dots, n} (sim(inst_i, inst_j))}{n + m}$$

1.3 快速稳定匹配

在 COMA^[9]中,采用 Selection 算子对模式 S1 和 S2 匹配的结果集进行单向 LargeSmall ($|S1| > |S2|$), 单向 SmallLarge ($|S1| \leq |S2|$) 或双向选择 (Both)。对于 Selection 算子的选择策略,提供了选择最大 N 个记录的 MaxN 操作,选择最大几率的 MaxDelta 操作和阈值 Threshold 的操作。这些策略都需要对模式 S1 和 S2 进行遍历,其算法复杂度为 ri (单向), 或 $2ri$ (双向)。

本体 1:1 映射实际上可以看作是对二分图的最大匹配^[10-11], 二分图指的是这样一种图:其所有的顶点分成两个集合 M 和 N , 其中 M 或 N 中任意两个在同一集合中的点都不相连。二分图匹配是指求出一组边,其中的顶点分别在两个集合中,并且任意两条边都没有相同的顶点,这组边叫做二分图的匹配,而所能得到的最大的边的个数,叫做最大匹配。为了解决二分图的最大匹配,采用 Hopcroft Karp^[12]算法。在增加匹配集合 M 时,每次寻找多条增广路,这样迭代次数最多为 $2V^{0.5}$, 所以,时间复杂度就降到了 $O(V^{0.5} \times E)$, V 为解析图顶点数; E 为解析图边数。

二分图匹配本质一种稳定匹配,在后面的迭代映射算法中的收敛条件中,采用稳定匹配均值作为迭代收敛条件。例如,有概念 Old1, Old2, New1, New2, $sim(Old1, New1)=0.9$, $sim(Old1, New2)=0.8$, $sim(Old2, New1)=0.6$, $sim(Old2, New2)=0.2$, 按稳定匹配算法,有匹配 $[Old1, New1]$, $[Old2, New2]$, 其匹配值为 $0.9+0.2=1.1$, 而非稳定匹配 $[Old1, New2]$, $[Old2, New1]$ 的匹配值 $0.8+0.6=1.4$ 。尽管后者匹配值大,但前者是稳定匹配。

1.4 复合映射策略

1.4.1 复合映射路径

基于前述的本体图解析,从3个层次和两个特征对异构本体进行映射分析。对于给定的本体,首先根据本体各节点元素及其上下文结构关系构建顶点和边,进而构建本体解析图;根据一定的策略注入本体的节点元素,对各节点元素进行遍历,得到节点的上下文关系,即图的结构特征;然后依据节点内容特征和相似度传播特征对各节点元素进行相似度的计算;最后是建立基于重用思想上的迭代映射。

1.4.2 复合映射框架

如图2所示,Matcher是最底层的匹配,采用基于词汇路径距离和基于wordnet的语义相似度的叠加值,完成节点内容特征层次上的相似度计算。MultiMatcher是多个Matcher的聚合,MultiMatcher基于某个节点进行遍历,针对遍历后的节点内容特征进行相似度计算,完成结构层次上的概念计算。Strategy是多个MultiMatcher的聚合,对解析图节点进行树遍历,基于多个MultiMatcher完成某个策略的映射。最终,多个MultiMatcher和多个Matcher共同完成本体解析图的结构特征和内容特征的多级映射策略,IterStrategy基于集合的交,并,差重用思想,对多个策略进行稳定匹配迭代映射。

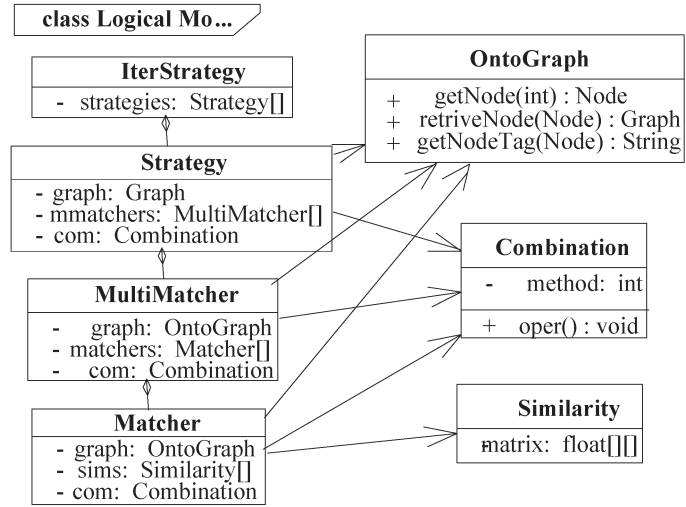


图2 复合映射策略
Fig.2 Composite mapping strategy

在本复合策略中,因涉及到多个Matcher的多个结果集合,所以需要用到对结果集进行合并的类Combination,合并方法一般有平均值法,最大值法,大多数法等等。Combination类采用集合相似度SSim的计算方法,集合相似度是两个集合间最优映射的顶

点对相似度的总和。 $SSim(S_a, S_b) = \frac{\sum_{e \in (S_a, S_b)} weight(e)}{\max\{|S_a|, |S_b|\}}$, 其中: $M(S_a, S_b)$ 是 S_a 和 S_b 两个集合最大权匹配得到的边的集合; e 是最大匹配中的边; $\max\{|S_a|, |S_b|\}$ 求出 S_a 和 S_b 两个集合中所包括的最大节点的个数。

1.4.3 结果重用

在策略迭代IterStrategy中,我们引入了重用算子 $\cap$ — $\cup$,分别代表集合的交,差,并操作,用于迭代映射的结果集重用。 $\cap$ 操作intersect返回两个不同结果集的共同映射对。—操作Diff返回映射对出现在一个结果集而没有出现在另外一个结果集中。 $\cup$ 操作Union则合并两个结果集的映射对。

1.4.4 收敛条件

对于每一种复合策略的映射,都可以得到一个映射结果集合。考虑到映射算法本身属于二分图匹配的的稳定算法,故将两次映射结果集合中的相似度均值误差作为迭代收敛条件。

2 复杂度分析

整个映射系统包括本体解析图构建,多个Strategy,每个Strategy包括多个MultiMatcher,以及每个MultiMatcher包括多个Matcher。在此定义:整个映射系统复杂度 c_{is} ,本体解析图构建复杂度 c_g ;Strategy复杂度 c_s ,MultiMatcher复杂度 c_{mm} ,Matcher复杂度 c_m ,以及结果集的合并操作复杂度 c_c ,相似度计算复杂度 c_s 和结果集匹配复杂度 c_h 。

$$c_m = c_c + c_s; c_{mm} = n_m \times c_m + c_c; c_s = n_{mm} \times c_{mm} + c_c; c_{is} = c_g + n_s \times c_s + c_c;$$

其中: n_m , n_{mm} , n_s 分别表示为每个MultiMatcher包括的Matcher个数,每个Strategy包括的MultiMatcher个数,整个系统包括的Strategy个数。故有: $c_{is} = c_g + n_s \times \{n_{mm} \times [(n_m \times (c_c + c_s) + c_c)] + c_c + c_h\}$ 。

对于本体图的解析,采用对图进行树遍历的方法,故本体图解析的复杂度为树的遍历复杂度 $O(n \log(n))$ 。基于字符串距离的依赖于字符串的长度 l ,为 $O(l)$ 。而基于wordnet的语义距离则是基于语义词典,其计算复杂度为 $O(l)$ 。在计算相似度传播时,考虑到传播的深度带来的复杂性,一般只考虑节点的

直接父节点,直接子节点和直接兄弟节点。故复杂度为这些节点集合的计算复杂度,设节点集合大小为 s ,结构相似度的计算复杂度为 $O(s^2)$ 。Hopcroft Karp 快速匹配算法的复杂度为 $O(V^{0.5} \times E)$,合并操作实际是集合的操作,故为 $O(s^2)$ 。

综合以上,系统复杂度

$$c_i = O(n \log(n)) + O(1) * \{ O(1) * [O(1) * (O(s^2) + O(s^2))] \} + O(s^2) + O(V^{0.5} \times E) = O(n \log(n)) \times O(s^2) + O(V^{0.5} \times E)$$

考虑到一般结构集大小 s 一般远小于节点数,故系统复杂度为 $O(n \log(n)) + O(V^{0.5} \times E)$ 。相对于一般系统如 GLUE 系统的复杂度 $O(n^2)$,系统运行复杂度大大降低,其主要原因在于首先采用了基于树的解析图结构,而不是对本体概念的遍历(其复杂度为 $O(n^2)$);其次采用了快速匹配算法 Hopcroft Karp,将算法复杂度将为 $O(V^{0.5} \times E)$;最后,尽管经过了多次迭代和多种策略,但正是由于应用了多种策略,使得迭代相当少就可以达到很好的精度。

3 测试分析

本系统分为系统性能对比测试和实验评估。系统性能考虑对两个本体映射的解析特征结构,映射对个数以及映射时间进行比较,并在同一型号 PC 机和主流系统进行对比。实验评估考虑的是对本体映射的质量进行比较。

3.1 系统性能测试

系统性能测试将采用 OAEI2010^[13]作为测试集,其中 101#为基准集,分别选用 103#,203#,224#,233#,246#,265#,301#为测试集。表 2 给出了一部分测试结果,包括本体的类,对象属性,数据属性以及实例个数、构建解析图后的顶点和边的个数、以及测试集相对于基准集的映射时间。相对来说,由概念及概念间关系解析出的顶点和边数越少,解析时间越少,相应映射时间也少。233#解析后的边数为 0 说明概念间没有任何相对关系如继承关系,ISA 关系,属性关系等等,概念为孤立的节点。

另外,本系统和另外两个映射系统 COMA^[9],S-Match^[5]进行映射时间对比,从对比可以看出,本系统在性能上得到一定的提升。

表 2 本体图解析和映射性能
Tab.2 Ontology graph parsing and mapping performance

类别	编号							
	101	103	203	224	233	246	265	301
图解析	类	37	37	37	38	36	32	15
	对象属性	26	26	26	26	2	2	0
	数据属性	46	46	46	46	6	7	40
	实例	57	56	57	1	56	32	0
	解析图顶点	101	101	101	101	37	34	55
	解析图边	123	115	123	123	0	22	54
	解析时间	181	141	125	94	46	47	78
映射性能	映射对		119	112	119	45	38	53
	映射时间		166	125	109	141	78	109
	映射时间(COMA)		213	152	155	189	99	156
	映射时间(S-Match)		289	223	210	301	231	352

3.2 实验评估测试

实验评估采用 OAEI2010 作为测试集,从编号 101 到 304 共 50 个数据集。 $F-Measure = (b^2 + 1)PR / (b^2P + R)^{[14]}$, 设当 $b=1$ 时, $F-Measure = 2PR / (P + R)$ 作为评估指标。其中: P 为查准率 $Precision = \text{发现的正确映射数} / \text{发现映射数}$; R 为查全率或称召回率 $Recall = \text{发现的正确映射数} / \text{实际存在的映射数}^{[15]}$; $F-Measure$ 指标综合了信息检索中的准确率和召回率。图 3 比较了 $F-Measure$ 在本文的测试系统 IOMBF (iterative ontology mapping based figure), COMA++ 以及 S-Match 的曲线图,可以看出,本文的测试效果优于 COMA++, 与 S-Match 不相上下。

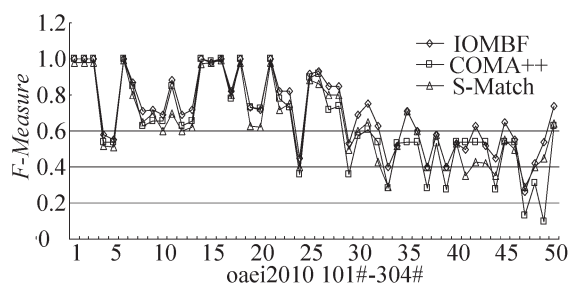


图3 F-Measure 实验比较

Fig.3 F-measure experiment comparison

4 结束语

本文通过对本体映射的分析,构建了一种基于多层解析的图描述结构,并在此基础上通过分析本体上下文结构,引入结构和实例相似度传播和快速匹配算法,设计一种复合匹配策略,用于本体映射。算法分析和实验证明,本系统能在保证性能基础上提高一定映射的质量。本文侧重于充分挖掘本体本身的内在结构,进而提出了相应的策略,对于不同的复合策略,将影响迭代算法的收敛次数,进而影响映射性能映射效果,因此,对于复合策略的研究有待于做进一步的工作。

参考文献:

- [1] EHRIG M, STAAB S. Quick ontology mapping[C]// Rome Proc of the ISWC'04, Italy: Proc of VLDB'01, 2004: 683-697.
- [2] YVES R, MANSUR R. ASMOV results for OAEI 2007[C]// Bexco Proc of the ISWC'07, Korea: Morgan Kaufmann Publishers, 2007: 141-151.
- [3] 唐杰, 梁邦勇, 李涓子, 等. 语义 Web 中的本体自动映射[J]. 计算机学报, 2006, 29(11): 1956-1976.
- [4] DENG L L L N, SMASSMAN N. Instance matching with COMA++[J]. Model Management and Metadaten-Verwaltung, 2007, 35(7): 56-63.
- [5] GIUNCHIGLIA F, SHVAIKO P, YATSKEVICH M. S-match: an algorithm and an implementation of semantic matching[J]. Proceeding of ESWS, 2004, 23(5): 61-75.
- [6] DONG X, HALEVY A, MADHAVAN J. Reference reconciliation in complex information spaces[C]// Tomah in Proc of the 2005 ACM SIGMOD Int Conference on Management of Data, New York: ACM Press, 2005: 85-96.
- [7] Muhlenbein H, Mahnig T. Convergence theory and application of the factorized distribution algorithm[J]. Journal of Computing and Information Technology, 1999, 7(1): 19-32.
- [8] 李熙, 徐德智, 王建新. 本体映射中结构策略改进算法[J]. 计算机工程与应用, 2010, 46(8): 24-28.
- [9] HH DO, RAHM E. COMA-a system for flexible combination of schema matching approaches[J]. in Proceedings of VLDB, 2001, 62(2): 610-621.
- [10] GUSFIELD D, IRVING R W. The stable marriage problem: structure and algorithms[J]. MIT Press Cambridge, 1989, 10(3): 23-35.
- [11] LOVASZ L. Matching theory[D]. New York: Elsevier Science Ltd, 1996: 233-235.
- [12] Blum N. A simplified realization of the hopcroft karp approach to maximum matching in genenal graphs[J]. Inst fur Informatik on Computing, 1999, 2(4): 225-231.
- [13] 宋岚, 雷莉霞, 王洪. 基于本体的智能化语义信息处理系统研究[J]. 华东交通大学学报, 2009, 10(5): 31-34.
- [14] CJVAN RIJSBERGEN. Information retrieval [D]. 2nd edition. Glasgow: Dept. of Computer Science University of

Glasgow, 1999:98-105.

[15] HDOS M, E. Comparison of schema matching evaluations[J]. workshop on Web Databases, 2002, 6(1): 12-18.

Composite Ontology Mapping Strategy of the Graph Parsing

Ling Shiyong¹, Gong Jinghong²

(1. School of Software Engineering, East China Jiaotong University; 2. School of Electrical and Electronic Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: According to complex context relation of ontology and in light of different input schema, this paper proposes ontology graph parsing in general environment. Based on ontology structure, the paper constructs graph description on multiple layer ontology. Then it introduces similarity propagation of structural and instance level in context relation and rapid mapping with rapid match algorithm. Finally, a composite ontology mapping strategy is proposed, which iteratively achieves ontology mapping result. The feasibility of the strategy is verified through algorithm complexity analysis and experiment comparison.

Key words: composite ontology mapping; ontology graph parsing; similarity propagation; rapid match algorithm

(上接第54页)

[8] 中华人民共和国住房和城乡建设部. CJJ37-2012 城市道路工程设计规范[S]. 北京:人民交通出版社, 2012.

[9] 李鹏飞, 过秀成, 李岩. 基于高级交通管理系统数据的城市干道协调控制方法[J]. 东南大学学报, 2012, 28(2): 229-235.

[10] 胡佩峰, 田宗忠, 袁振洲, 等. 交通运行管理中的可变绿波带优化[J]. 交通运输系统工程与信息, 2011, 11(1): 61-72.

Optimal Design of Signal Coordination Control System on Bailuzhou Road Based on Synchro

Guo Jiangang, Lin Wenfan, Chen Bitai, Huang Hainan, Chen Jinshan, Zhang Wenxing

(College of Transportation, Fujian Agriculture and Forestry University, Fuzhou 350002, China)

Abstract: A signal coordination control system on an arterial road can effectively improve the capacity of existing transportation facilities. Taking the Bailuzhou Road in Xiamen city as an research object, this paper proposes the signal coordination control method “with different directions, at different periods, and at half-cycle length” according to the geometric structure characteristics and the survey data. Based on Synchro, the signal coordination control system is established. The simulation results of traffic operational states on this arterial road indicate that the average intersections delay is reduced by 11.3% and 21.6% respectively, and the average arterial delay is reduced by 8.2% and 55.9% respectively, while the total delay of the road network is reduced by 13.3% and 22.0% respectively during the morning and evening peak hours.

Key words: coordination control system; Synchro; optimal design; arterial roads