doi:10.3969/j.issn.1672 - 4348.2015.01.009

一种新的自动化本体映射技术

江荔[1], 薛醒思[2]

(1. 福州职业技术学院 计算机系, 福建 福州 350108; 2. 福建工程学院 信息科学与工程学院, 福建 福州 350118)

摘要: 为了更好地组合不同的相似度度量结果以提高本体映射结果的质量,提出一种新的基于调谐值度量和单纯降序提取算法的自动化本体映射技术。该技术首先通过调谐值来度量各种相似度矩阵的可靠性,并为每一个相似度矩阵赋予权重以集成不同的相似度矩阵,然后通过单纯降序提取算法结合阈值的策略提取最终的本体映射结果。实验采用 2012 年本体映射评价竞赛的测试数据集,同 2012 年本体映射评价竞赛的参与者的比较表明该文提出的方法是有效的。

关键词: 本体映射技术; 调谐值度量; 单纯降序提取算法

中图分类号: TP182　　**文献标志码**: A　　**文章编号**: 1672 - 4348(2015)01 - 0049 - 05

A novel automatic ontology aligning technology

Jiang Li[1], Xue Xingsi[2]

(1. Computer Department, Fuzhou Polytechnic, Fuzhou 350108, China;
2. College of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China)

Abstract: To improve the quality of ontology alignment through better aggregating the alignments obtained by various similarity measures, a novel automatic ontology aligning technology based on the tuning measure and naive descending extraction algorithm was proposed. Firstly, the tuning value was utilized to measure the reliability of each similarity matrix, and each similarity matrix was given a weight to aggregate them. Then both naive descending extraction algorithm and threshold strategy were used to extract the final ontology alignment. The testing cases from the ontology alignment evaluation initiative in 2012 were adopted as the data group. The comparison with the participants' algorithm of ontology alignment evaluation initiative in 2012 shows the effectiveness of the proposal.

Keywords: ontology aligning technology; tuning measure; naive descending extraction algorithm

本体是一种最新的信息交换参考模型,它是迄今为止用于获取最准确的语义规范化描述的技术。但是,现实中往往由于本体设计者的主观性,对于同一个领域中的同一对象可能会有不同的描述方法,这就导致了术语和概念描述不一致的异质本体的产生。异质本体中存在的语义异质问题是获取语义层面交互的障碍,而本体映射技术是业内公认的可以有效解决本体间存在的语义异质问题的技术。

现有的本体映射技术都是通过组合不同的相似度度量结果来提高本体映射结果的质量。虽然存在很多种系统拓扑结构以组合相似度度量结果(如串行组合或迭代计算的方式),但是最常用的还是由 COMA + + 提出的并行组合的系统体系结构。[①]在并行组合的系统中,所有的相似度度量技术彼此间独立地运行以获取不同的映射结果。这

① Aumueller D, Do H H, Massmann S. Schema and ontology matching with COMA ++, Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, 2005: 906 - 908。

收稿日期: 2014 - 11 - 10
基金项目: 福建省教育厅科技项目(JA13227);福建省教育厅科技项目(JB12314)
第一作者简介: 江荔 (1980 -),女(汉),湖北天门人,讲师,硕士,研究方向为本体匹配技术和智能计算。

些映射结果用相似度矩阵(该矩阵的行和列分别表示源本体和目标本体中的实体,矩阵的元素是相应实体间的相似度值)来表示,因此每个相似度度量技术都会产生一个相似度矩阵。然后通过相似度集成的方法将不同的相似度矩阵的值组合为单一的相似度矩阵,最后通过阈值选择可信度高的映射元素以生成最后的本体映射结果。针对上述过程,本文提出了一种新的自动化本体映射技术,该技术可以自适应地集成不同的相似度度量的结果并提取最终的本体映射结果。

1 自适应的相似度集成技术

经典的相似度集成的方法是 MIN/MAX、WEIGHTED 和 AVERAGE 等[1],近期又出现了 NONLINEAR[2] 和 OWA[3] 等方法。由于不同的相似度度量技术都有其优势和不足,因此本文在 COMA++ 的 Direction-Both 选择策略和 Stable Marriage 属性[4]的基础上提出了一种新的自适应的相似度集成方法,该方法可以分辨哪一个相似度度量产生的结果更为可靠,从而为可靠的映射结果值赋予更高的集成权重。

1.1 调谐值度量

每个相似度度量技术产生的映射结果可以表示为一个相似度矩阵,该矩阵的每一个行对应一个本体中每一个实体,每一列对应另外一个本体中的每一个实体,矩阵的元素表示两个本体中相应的实体间的相似度值。因此在参考映射结果为1:1的情况下,两个实体如果最终能够映射上的话,在相似度矩阵中的相应元素的值应当比同这两个实体的行或列相同的其他实体对对应的元素的值大,即意味着这两个实体彼此间都是最佳映射。在此基础上,一个相似度矩阵的调谐值度量可以定义如下:

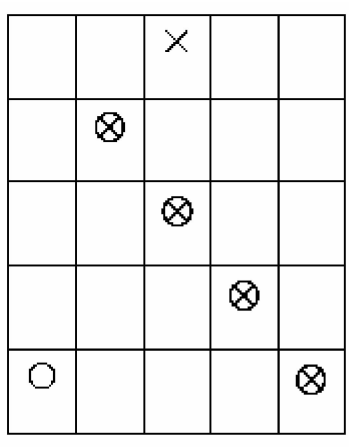
$$h = \frac{s\_\max}{\min(|O_1|, |O_2|)} \qquad (1)$$

其中 $s\_\max$ 是在相似度矩阵中对应的行和列中拥有唯一最高值的实体映射对的数量,$|O_1|$ 和 $|O_2|$ 分别是本体 O_1 和 O_2 中的实体数量。

图1给出了相似度矩阵的调谐值计算的例子。本体 O_1 和 O_2 中的每一个实体对的相似度如图1(a)所示,图1(b)中"×"表示相似度矩阵在该位置的相似度值是该行中的最大值,"○"表示相似度矩阵在该位置的相似度值是该列中的最大值,"⊗"表示相似度矩阵在该位置的相似度值在该行该列中都是最大值。因此,图1(a)中的相似度矩阵的 $s\_\max$ 就是图1(b)中的"⊗"个数,该矩阵的调谐值 h 为 4/5 = 0.8。

O_2 / O_1	e_{21}	e_{22}	e_{23}	e_{24}	e_{25}
e_{11}	0.11	0	0.22	0.1	0.1
e_{12}	0.22	1	0.2	0.2	0.2
e_{13}	0.18	0.09	0.36	0.09	0.18
e_{14}	0.11	0.22	0.11	0.9	0.1
e_{15}	0.3	0.2	0.1	0.1	1

(a)本体 O_1 和 O_2 的相似度矩阵



(b)相似度矩阵中行列最大值

图1 相似度矩阵的调谐值计算示例

Fig. 1 Calculation example of tuning value for similarity matrix

1.2 相似度集成过程

由于调谐值度量可以用于表示某个相似度度量获得的相似度矩阵的重要性和可靠性,因此调谐值度量可以作为权重以集成不同的相似度矩阵(即不同的相似度度量结果)。给定两个实体 e_{1i} 和 e_{2j},其最终的相似度值可定义如下:

$$\mathrm{FinalSim}(e_{1i},e_{2j}) = \frac{\sum_{k=1}^{num} \mathrm{tune}_k \times \mathrm{sim}_k(e_{1i},e_{2j})}{\mathrm{num}} \quad (2)$$

其中 tune_k 是第 k 个相似度矩阵的调谐值，num 是相似度矩阵的数量，$\mathrm{sim}_k(e_{1i},e_{2j})$ 是第 k 个相似度矩阵中 e_{1i} 和 e_{2j} 的相似度值。

在本文的工作中，相似度集成的过程如下：1）将不同的相似度度量的映射结果转化为相应的相似度矩阵；2）计算每一个相似度矩阵的调谐值，并通过公式（2）将不同的相似度矩阵中的对应元素的值集成起来，获取最终的相似度矩阵；3）将最终的相似度矩阵转化为本体映射结果。

此外，如果某个相似度矩阵有很高的调谐值，说明该相似度矩阵对应的相似度度量方法是可靠的，因此该矩阵中每行或每列相似度值最低的元素被认为是噪音。在集成相似度矩阵元素值的时候，可以先过滤掉一定比例的噪音元素（即将其置 0）。将矩阵元素的度值进行排序之后，可过滤掉的相似度值的数量由以下公式计算：

$$\mathrm{noiseNum} = \min(l-1, \mathrm{tune} \times 1) \quad (3)$$

其中 noiseNum 是相似度矩阵中可过滤的噪音元素的数量，l 是相似度矩阵的行或列的长度（在本文中取相似度矩阵的行长度和列长度的较小者），tune 是相似度矩阵的调谐值。

2　映射结果提取

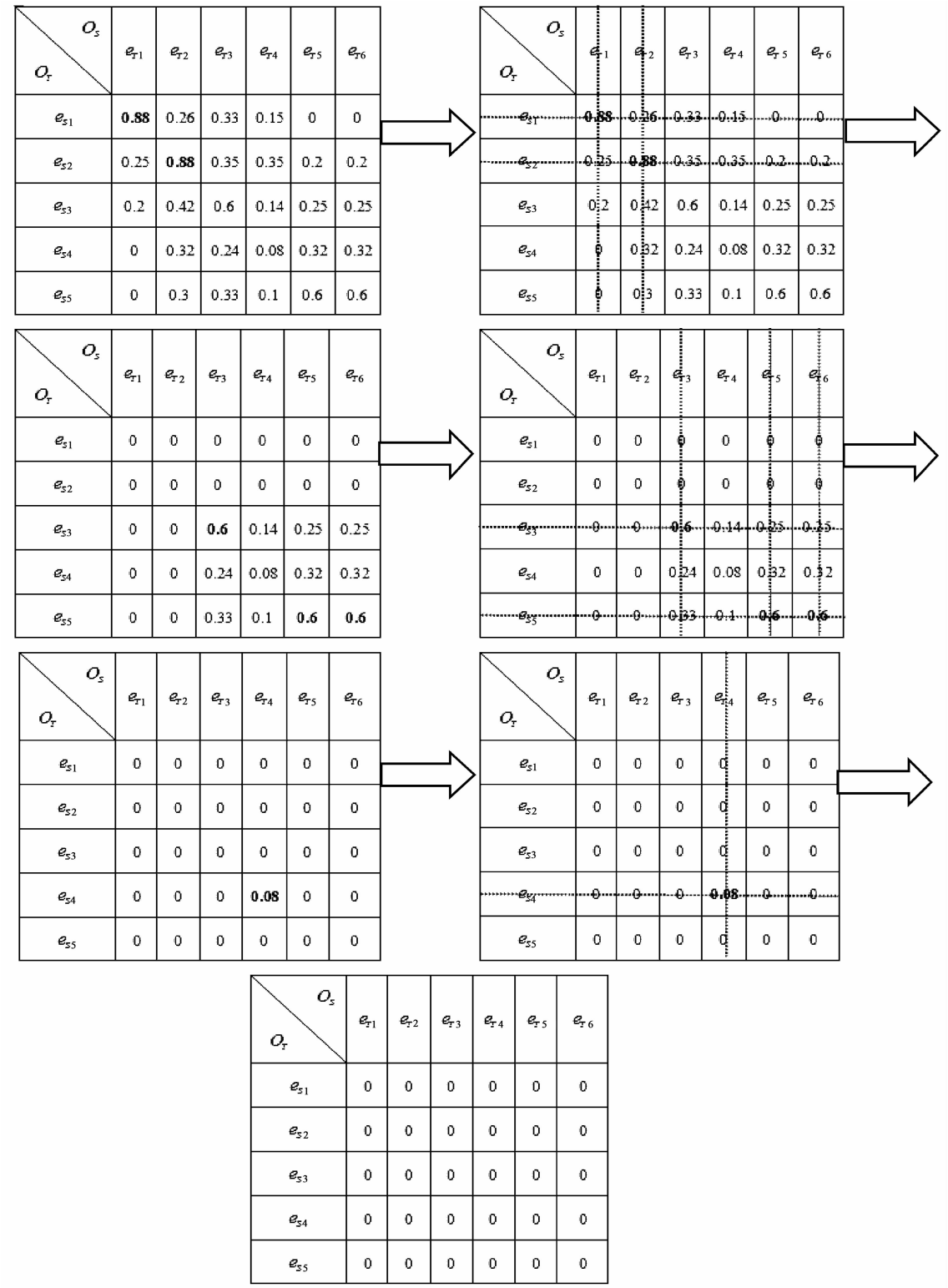
各种相似度度量获得的映射结果集成之后，将得到一个最终的相似度矩阵，该矩阵的第 i 行和第 j 列分别代表源本体 O_s 和 O_T 中的实体 e_{Si} 和 e_{Tj}，矩阵中第 i 行和第 j 列的值表示实体 e_{Si} 和 e_{Tj} 的相似度值。然而，最终获得的相似度矩阵可能具有大量的噪声，如何从相似度矩阵中提取出比较可靠的实体对应关系也是一个关键问题。由于矩阵中的值表示实体对的相似度，从一定程度上反应了实体对等价的可信度，即相似度值越大说明实体对等价的可能性越高，反之说明实体对等价的可能性越小。目前使用得最多的是阈值策略，即通过设置阈值过滤掉相似度值较小的实体对应关系，保留相似度值较大的实体对应关系作为最终的结果。由于单纯降序提取算法可以提取出比阈值策略更好的映射结果，因此本文采用基于单纯降序提取算法的方法来提取最终的映射结果。单纯降序提取算法步骤如下：1）将矩阵中所有相似度值按降序排列；2）记录下最大值在矩阵中的位置；3）把与最大值同行同列的相似度值设置为 0；4）重复以上三个步骤直到矩阵中所有相似度值都为 0。最终提出的映射结果为步骤（2）中记录下的所有位置所对应的实体对。图 2 中给出了通过单纯降序提取算法提取本体映射结果的示例。

如图 2 所示，最终抽取出了 6 个实体对应关系，它们分别为：$(e_{S1},e_{T1},0.88,=)$，$(e_{S2},e_{T2},0.88,=)$，$(e_{S3},e_{T3},0.6,=)$，$(e_{S5},e_{T5},0.6,=)$，$(e_{S6},e_{T6},0.08,=)$，$(e_{S4},e_{T4},0.6,=)$。事实上，本体中实体间的相似度值越大，实体之间存在对应关系的可能性也越大。然而单纯降序提取算法直到矩阵中所有相似度值都为 0 时才终止，这将会提取出一些相似度值很小的实体对应关系，而这些实体对应关系很有可能属于噪声。因此，本文融合了阈值策略和单纯降序提取算法，设置一个阈值参数，当相似度矩阵中所有值都小于阈值时，终止算法的执行。假设阈值取为 0.5，即小于 0.5 的相似度认为是不可信的，则图 2 的相似度矩阵抽取结果的过程在第 5 步就将终止，最终将抽取出 5 个实体对应关系，它们分别为 $(e_{S1},e_{T1},0.88,=)$，$(e_{S2},e_{T2},0.88,=)$，$(e_{S3},e_{T3},0.6,=)$，$(e_{S5},e_{T5},0.6,=)$，$(e_{S5},e_{T6},0.6,=)$。

3　实验配置与结果

实验所采用的相似度度量技术为：SMOA 距离（基于词汇的相似度度量）、基于 WordNet[5] 的相似度度量（基于语言学的相似度度量）和基于本体结构的相似度度量（similarity flooding，SF）算法。实验所采用的本体映射结果的质量度量标准为：查全率（recall）、查准率（precision）和 f 度量（f-measure）[5]。

实验采用本体映射领域公认的 2012 年本体映射评价竞赛的测试数据集。表 1 中给出了 2012 年本体映射评价竞赛参与者和本文提出的方法的运行结果的均值，其中符号 R、P 和 F 分别表示获取的映射结果的查全率、查准率和 f 度量的值。



O_s / O_T	e_{T1}	e_{T2}	e_{T3}	e_{T4}	e_{T5}	e_{T6}
e_{s1}	**0.88**	0.26	0.33	0.15	0	0
e_{s2}	0.25	**0.88**	0.35	0.35	0.2	0.2
e_{s3}	0.2	0.42	0.6	0.14	0.25	0.25
e_{s4}	0	0.32	0.24	0.08	0.32	0.32
e_{s5}	0	0.3	0.33	0.1	0.6	0.6

O_s / O_T	e_{T1}	e_{T2}	e_{T3}	e_{T4}	e_{T5}	e_{T6}
e_{s1}	**0.88**	0.26	0.33	0.15	0	0
e_{s2}	0.25	**0.88**	0.35	0.35	0.2	0.2
e_{s3}	0.2	0.42	0.6	0.14	0.25	0.25
e_{s4}	0	0.32	0.24	0.08	0.32	0.32
e_{s5}	0	0.3	0.33	0.1	0.6	0.6

O_s / O_T	e_{T1}	e_{T2}	e_{T3}	e_{T4}	e_{T5}	e_{T6}
e_{s1}	0	0	0	0	0	0
e_{s2}	0	0	0	0	0	0
e_{s3}	0	0	**0.6**	0.14	0.25	0.25
e_{s4}	0	0	0.24	0.08	0.32	0.32
e_{s5}	0	0	0.33	0.1	**0.6**	**0.6**

O_s / O_T	e_{T1}	e_{T2}	e_{T3}	e_{T4}	e_{T5}	e_{T6}
e_{s1}	0	0	0	0	0	0
e_{s2}	0	0	0	0	0	0
e_{s3}	0	0	0	0	0	0
e_{s4}	0	0	0	**0.08**	0	0
e_{s5}	0	0	0	0	0	0

O_s / O_T	e_{T1}	e_{T2}	e_{T3}	e_{T4}	e_{T5}	e_{T6}
e_{s1}	0	0	0	0	0	0
e_{s2}	0	0	0	0	0	0
e_{s3}	0	0	0	0	0	0
e_{s4}	0	0	0	0	0	0
e_{s5}	0	0	0	0	0	0

图 2　单纯降序提取算法提取本体映射结果的示例

Fig. 2　Ontology alignment result extracted by naive descending extraction algorithm

可以从表 1 中的数据看出,本文的方法的获取的本体映射结果的 *f*-measure 值优于所有 2012 年本体映射评价竞赛参与者,此外本文提出的方法的运行时间为 8 s,排名第 2 位。从实验结果上可以看出本文提出的方法是有效的。

表 1　本文的方法同 2012 年本体映射评价竞赛参与者的比较

Tab. 1　Comparison between the proposed algorithm and the participants algoithm of Ontology Alignment Evaluation Initiative in 2012

系统	*R*	*P*	*F*	运行时/s	系统	*R*	*P*	*F*	运行时/s
MapSSS	0.77	0.99	0.87	35	WikiMatch	0.54	0.74	0.62	750
YAM + +	0.72	0.98	0.83	120	ServOMap	0.43	0.88	0.58	18
AROMA	0.64	0.98	0.77	8	LogMap	0.45	0.73	0.56	28
AUTOMSv2	0.54	0.97	0.69	80	MaasMatch	0.57	0.54	0.56	38
WeSeE	0.53	0.99	0.69	650	MEDLEY	0.50	0.60	0.54	85
Hertuda	0.54	0.90	0.68	9	ServOMapLt	0.20	1.00	0.33	7
HotMatch	0.50	0.96	0.66	20	ASE	0.54	0.49	0.51	40
Optima	0.49	0.89	0.63	380	本文提出的方法	0.88	0.94	0.91	8

4　结论

为了更好地组合不同的相似度度量结果以提高本体映射结果的质量，本文提出了一种新的基于调谐值度量和单纯降序提取算法的自动化本体映射技术。该技术首先通过调谐值来度量各种相似度矩阵的重要性和可靠性，并为每一个相似度矩阵赋予权重以集成不同的相似度矩阵，然后通过单纯降序提取算法结合阈值的策略提取最终的本体映射结果。实验采用 2012 年本体映射评价竞赛的测试数据集，同 2012 年本体映射评价竞赛的参与者的比较表明该文提出的方法是有效的。

参考文献：

[1] Alsayed A. Management of XML data by means of schema matching[D]. Magdeburg:Otto-von-Guericke-University,2010.

[2] Ji Q, Haase P, Qi G. Combination of similarity measures in ontology matching using the OWA operator[M]//Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice. Berlin:Springer, 2011:281 - 295.

[3] Gusfield D, Irving R W. The stable marriage problem: structure and algorithms[M]. Cambridge: MIT Press,1989:156 - 157.

[4] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM,1995,38(11):39 - 41.

[5] Van Rijsbergen C J. Information Retrieval[M]. London: Butterworth,1975:48 - 49.

（责任编辑：肖锡湘）

（上接第 14 页）

参考文献：

[1] 姚燕，王玲，田培. 高性能混凝土[M]. 北京：化学工业出版社，2006.

[2] 冯乃谦. 高性能混凝土与超高性能混凝土的发展与应用[J]. 施工技术，2009，38(4)：1 - 6.

[3] 王成启，张悦然. 矿物掺合料对海工自密实高性能混凝土耐久性影响[J]. 混凝土，2014(1)：56 - 61.

[4] 李清富，薛延信，张海洋. 不同掺合料对高性能混凝土工作性能影响的试验研究 [J]. 混凝土，2012(1)：58 - 61.

[5] 李北星，马立军，关爱军，等. 箱梁 C55 高性能混凝土的抗裂性与耐久性研究[J]. 武汉理工大学学报，2010，32(14)：40 - 44.

[6] 王明芳，孙玉永. 高性能混凝土抗盐蚀耐久性试验研究[J]. 工业建筑，2012，42(6)：127 - 130.

（责任编辑：陈雯）