

多策略自适应大规模本体映射算法

蒋 猛, 禹明刚, 王智学

(陆军工程大学 指挥控制工程学院, 南京 210007)

摘 要: 大数据背景下大规模本体映射的时间复杂度较高, 效率和精度较低。为此, 提出一种基于模块化和局部置信度的多策略自适应大规模本体映射算法。对本体内部进行聚类 and 模块化, 基于信息检索策略发现模块间高相似度的相关子本体, 计算相关子本体间各映射策略下的局部置信度, 在组合映射结果时基于局部置信度对相应策略的权值进行自适应调整。在此基础上, 利用启发式贪心策略提取映射结果并基于映射规则矫正结果。实验结果表明, 与 Falcon、ASMOV 方法相比, 该算法具有较高的查全率、查准率与 F-measure 值。

关键词: 大数据; 大规模本体映射; 模块化; 局部置信度; 自适应

开放科学(资源服务)标志码(OSID):



中文引用格式: 蒋猛, 禹明刚, 王智学. 多策略自适应大规模本体映射算法[J]. 计算机工程, 2019, 45(3):14-19.

英文引用格式: JIANG Meng, YU Minggang, WANG Zhixue. Multi-strategy adaptive large-scale ontology mapping algorithm[J]. Computer Engineering, 2019, 45(3):14-19.

Multi-strategy Adaptive Large-scale Ontology Mapping Algorithm

JIANG Meng, YU Minggang, WANG Zhixue

(College of Command and Control Engineering, The Army Engineering University of PLA, Nanjing 210007, China)

[Abstract] Large-scale ontology mapping in the context of large data has high time complexity, low efficiency and accuracy. Therefore, a multi-strategy adaptive large-scale ontology mapping algorithm based on modularity and local confidence is proposed. Clustering and modularizing the inner part of the system, discovering the correlated sub-ontologies with high similarity between modules based on information retrieval strategy, calculating the local confidence under each mapping strategy among the correlated sub-ontologies, and adjusting the weight of the corresponding strategy adaptively based on the local confidence when combining the mapping results. On this basis, heuristic greedy strategy is used to extract mapping results and correct them based on mapping rules. Experimental results show that compared with Falcon and ASMOV methods, the proposed algorithm has higher recall, precision and F-measure value.

[Key words] big data; large-scale ontology mapping; modularity; local confidence; self-adaption

DOI:10.19678/j.issn.1000-3428.0052715

0 概述

近年来, 企业在日常业务及建设发展中产生大量跨部门、跨领域的信息共享与业务协同需求。由于业务的本质是发生在系统间的以数据为形式的对话, 因此实现信息共享与业务协同的关键是用统一有序的标准数据对混乱的数据环境进行梳理, 确保信息拥有者和使用者对数据具有共同认知, 最终实现良好的信息共享。

本体映射是解决异构本体间知识共享, 实现跨区域互操作和数据标准化的有效途径^[1-2]。目前, 大数

据背景下的本体映射主要面临 2 个难题。首先是大数据会产生大本体。在日常业务中, 随着领域数据的大量产生, 本体规模急剧膨胀。现有系统处理大规模本体映射时, 搜索空间的指数级剧增导致映射效率和映射质量大幅降低。其次是数据异构催生领域本体多样性, 传统方法已无法满足大量异构本体对于映射效率以及映射精度的需求。

目前, 在大规模本体映射研究领域, 存在 3 种典型方法:

一种是基于数据场的大规模本体映射方法^[3], 该方法首先从待映射本体中提取与映射任务相关性

基金项目: 国家自然科学基金(61802428)。

作者简介: 蒋 猛(1977—), 男, 博士研究生, 主研方向为大数据计算; 禹明刚(通信作者), 讲师; 王智学, 教授、博士生导师。

收稿日期: 2018-09-20 修回日期: 2018-11-06 E-mail: yumingang8989@163.com

最大的子本体,然后在子本体间进行映射,大量实验结果表明其本体规模仍然较大。

另外一种方法在进行本体对相似度计算前首先计算其置信度,直接放弃低于某置信度阈值的本体对的相似度计算,然后针对保留下来的本体对采用 Sigmoid 等方法确定多策略的权值分配^[4-5]。设本体规模为 n ,判断本体对间匹配与否的时间复杂度为 $O(t)$,则 $O(n^2t)$ 为本体映射总时间复杂度,该方法缩短了 $O(n^2t)$ 中的时间 t ,然而 $O(n^2)$ 本身仍然较大,因此,其映射效率提升有限。

最后一种方法基于模块化思想,将一个大规模本体按照特定规则分解成若干子本体,子本体通常隶属于某特定、具体的领域,且该子本体保持语义的相对独立性和完整性,在此基础上,计算子本体间的相似度^[6-7]。该方法压缩本体规模 n ,以实现时间复杂度及映射精度的提升,但其仍需针对每个子本体

对计算每种映射策略下的相似度,因此,在时间复杂度上仍有提升空间。

本文针对大数据背景下的本体映射问题,提出一种基于模块化和局部置信度的多策略自适应大规模本体映射算法。基于距离和信息量对本体内部进行聚类,将本体划分为若干模块。利用信息检索策略发现模块间高相似度的相关子本体,分别计算不同映射策略下相关子本体间的局部置信度,基于该置信度对各策略进行动态自适应组合并获取映射结果,最终基于领域规则校正映射结果。

1 大规模本体映射框架

本文结合大数据背景下的企业本体映射需求,提出基于模块化和局部置信度的大规模本体映射算法,框架如图 1 所示。其中,空心圆圈和实心圆圈代表 2 类不同的本体。

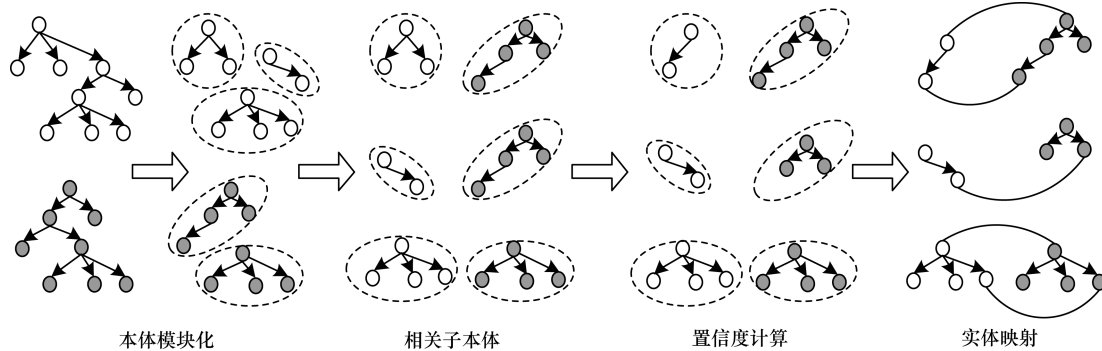


图 1 基于模块化和局部置信度的大规模本体映射框架

基于模块化和局部置信度的大规模本体映射共分 4 个阶段:

1) 本体模块化,通过对本体内部的实体间的相似度计算,对实体进行聚类,即将本体划分成若干规模适中的子本体集合。

2) 相关子本体发现,基于信息检索搜索参考点,提取出相关度高的子本体集。

3) 置信度计算,针对不同映射策略分别计算相关子本体集之间子本体的局部置信度,动态自适应组合各策略并进行综合相似度计算。

4) 实体映射及校正,提取映射结果并基于映射规则校正该结果。

在该方法中,将每个待映射本体的候选匹配实体规模缩减为满足置信度阈值的相关子本体数目,可以压缩映射规模,降低计算复杂度,结果校正环节消除了错误映射,可以确保映射精度。

2 大规模本体映射算法

2.1 本体模块化

本体模块化依据特定策略将待映射本体分解为若干规模适中的本体簇,簇满足以下特点:簇内具备高凝聚度,簇间具备低邻近度。本体内部概念之间

的相似度计算是本体模块化最核心的工作,该计算有 2 种方法,分别基于信息量和距离,本文对 2 种方法进行综合加权。

基于信息量的方法认为概念间的共享信息量决定了概念之间的相似度:

$$sim_i(c_1, c_2) = \frac{2IC(MSCS(c_1, c_2)) + \delta}{IC(c_1) + IC(c_2) + \delta} \quad (1)$$

其中, $MSCS(c_1, c_2)$ 表示 2 个概念 (c_1, c_2) 间的共享信息量, $IC(c) = -\lg p(c)$ 为概念 c 的信息量, δ 为正实数。

基于距离的方法认为,概念本身以及离 2 个概念最近的公共祖先所处的位置共同决定了概念之间的相似度:

$$sim_d(c_1, c_2) = \frac{d(MSCS(c_1, c_2))}{d(c_1) + d(c_2)} \quad (2)$$

其中, $d(c)$ 为概念 c 在本体树中的深度, $d(MSCS(c_1, c_2))$ 代表 2 个概念最近的公共祖先的位置。

综上,本文将本体概念相似度定义为:

$$sim(c_1, c_2) = \alpha sim_d(c_1, c_2) + (1 - \alpha) sim_i(c_1, c_2) \quad (3)$$

其中, α 为权重参数(实验中一般取 $\delta = 1, \alpha = 0.5$)。考虑到部分概念间相似度较低的事实,出于降低时

间复杂度的目的,本文只对父子、兄弟概念间的相似度进行计算。

在相似度结果的存储上,本文借助哈希表保存2个概念及其相似度:

$$links = \{(c_i, c_j, sim(c_i, c_j))\} \quad (4)$$

本文通过改进经典复杂网络算法 AHSCAN^[8] 以实现本体模块化。首先依据相似度值对 *links* 排序,将 *C* 中每个概念初始化为一个簇,依次取出 *links* 中的最大值并合并2个概念所在的簇,然后从哈希表中剔除该 *links*。多次迭代,直到 $|links| = 0$,该过程终止。具体算法如下:

算法1 本体模块化算法

输入 本体 *O* 的概念,属性集合 *D*, *links*

输出 本体概念及属性子集 *P_D*

P_D = ∅;

Sort(*links*); //对 *links* 排序

For each class *c_i* in *C* //初始化

p_i = cluster(*c_i*);

P_D = *P_D* ∪ *p_i*;

classToCluster.put(*c_i*, *p_i*);

end for

while (|*links*| > 0) //概念聚类

link = *links*.get(0); //相似度最大值

p_s = classToCluster.get(link. *c₁*);

p_t = classToCluster.get(link. *c₂*);

If (*p_s* ≠ *p_t*)

If (|*p_s*| + |*p_t*| < β) //合并

for each class *c* in *p_t*

p_s.add(*c*);

classToCluster.put(*c*, *p_s*);

P_D.remove(link);

end for

end if

links.remove(link);

end if

end while

for each property *p_i* in *P* //属性处理

p_i = classToCluster.get(*p_i*.domain);

p_i.add(*p_i*);

end for

return *P_D*;

设本体规模为 *n* (其中,概念和属性数分别为 *n₁* 和 *n₂*),最坏情况下每个概念需与其余所有概念进行相似度计算, *links* 规模为 $n_1(n_1 - 1)/2$,实际中假设每个概念仅需与 *m* 个概念进行相似度计算,则 *links* 实际规模为 $mn_1/2$,据此得到如下的排序时间复杂度:

$$O\left(\frac{mn_1}{2} \cdot \ln\left(\frac{mn_1}{2}\right)\right) \quad (5)$$

此外,初始化、概念聚类、属性处理的时间复杂

度分别为 $O(n_1)$ 、 $O\left(\frac{mn_1}{2}\right)$ 、 $O(n_2)$,因此,算法1的

总时间复杂度为: $O\left(\frac{mn_1}{2} \cdot \ln\left(\frac{mn_1}{2}\right)\right) + O(n_1) +$

$O\left(\frac{mn_1}{2}\right) + O(n_2) = O(n \ln n)$ 。相比已有3种典型

方法,该算法时间复杂度明显降低。

2.2 相关子本体

相关子本体指在本体模块化的基础上,发现异构本体之间的高相关性子本体。相关子本体获取的前提是参考点的确定,参考点指一批具有较高可信度的匹配实体对,通过考察参考点在子本体集中分布状况从而获取相关的子本体。本文将本体映射问题转化为对信息检索的求解问题,结合文献[9-10]的研究成果,利用开源全文搜索引擎 Lucene (<http://Lucene.apache.org>) 实现对相关子本体的搜索。

文献[11]设计了动态缩小搜索空间算法,首先找出可靠性高的一批参考点,然后对其临近节点进行相似度的迭代计算,计算过程如图2所示。

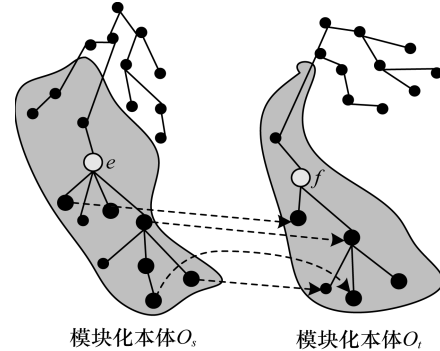


图2 基于参考点的相关子本体计算过程示意图

在图2中,实心圆圈表示本体概念,虚线箭头为映射关系, *O_s*、*O_t*为不同领域经模块化后的本体集, *e*、*f*为2个参考点。以本体 *O_s*、*O_t*为例,给出如下基于信息检索的参考点发现算法:

算法2 基于信息检索的参考点发现算法

步骤1 为 *O_s*、*O_t*的概念(属性)建立虚拟文档,进行分词、正规化等预处理并建立倒排索引。name和 comments 两个域分别存储概念(属性)的 URI 和所有自然语言描述: URI + label + comment + other description。

步骤2 针对 *O_t*中每个概念(属性)的词汇构建查询语句,实现对 *O_s*概念(属性)的倒排索引,保存查询结果中相似度最大的 Top-*K* 个结果。设2个子本体 $O_i \in O_1, O_j \in O_2$,则两者间相似度为:

$$sim(O_i, O_j) = \frac{2N(O_i, O_j)}{\sum_{O_{2K} \in O_2} N(O_i, O_{2K}) + \sum_{O_{1K} \in O_1} N(O_{1K}, O_j)} \quad (6)$$

$N(O_i, O_j)$ 返回 O_i 和 O_j 间参考点个数,若 2 个子本体间的相似度大于某一阈值 μ , 则将它们认定为相关子本体。

步骤 3 交换本体 O_s 和 O_t , 重复执行步骤 2。

步骤 4 取步骤 2 与步骤 3 结果的交集,即为所求参考点,若 2 个参考点有重复实体,则取相似度较大者。

相关子本体的获取是置信度计算及实体映射的基础,同时,置信度计算及实体映射均在相关子本体间展开。

2.3 置信度计算与实体映射

在本体模块化及相关子本体获取后,需进行置信度的计算。不同映射策略对本体本身的特征要求不同,例如,基于实例的映射策略对映射本体的实例规模以及实例集的分布状况都有所限制。而现有的基于全局固定权值的映射策略忽略了不同策略对于本体特征的差异性和适用性,因此,本文引入局部置信度的概念,以实现多种映射策略的动态自适应组合和权值分配。

本文采用基于实例、结构和属性的 3 种策略组合方式,在每种策略下分别进行局部置信度的计算,若两实体基于该策略的局部置信度高于某特定阈值,则认定其在该策略下存在映射。

在基于实例的策略中,实例匹配局部置信度计算如下:

$$M_{\text{credibility}}^{\text{instance}} = \left(1 - \frac{a}{|I_s| + |I_t| + a}\right) \times \frac{\min\{|I_s|, |I_t|\}}{\max\{|I_s|, |I_t|\}} \quad (7)$$

其中, $|I_s|$ 与 $|I_t|$ 分别为源本体与目标本体的实例规模, $M_{\text{credibility}}^{\text{instance}}$ 正相关于本体对的实例数量之和,即本体对规模越大,其匹配的置信度就越大。同时, $M_{\text{credibility}}^{\text{instance}}$ 负相关于两本体实例数目的差值,即本体对的实例数目差距越大,则匹配的置信度就越小。调节参数 $a > 1$, 以确保 $M_{\text{credibility}}^{\text{instance}} \in (0, 1)$ 。若 $M_{\text{credibility}}^{\text{instance}}$ 低于预设阈值,在该本体对映射中将抛弃基于实例的映射策略;反之,若 $M_{\text{credibility}}^{\text{instance}}$ 高于预设阈值,则以 $M_{\text{credibility}}^{\text{instance}}$ 为依据进一步确定该策略的权重,然后计算该策略下本体对的相似度。

在基于结构的策略中,结构匹配局部置信度计算如下:

$$M_{\text{credibility}}^{\text{structure}} = \frac{\text{Common}(|\text{depth} \cup \text{sub}|)}{\min(|\text{Nonleaf}_{o_1}|, |\text{Nonleaf}_{o_2}|)} \quad (8)$$

其中, $\text{Common}(|\text{depth} \cup \text{sub}|)$ 为 2 个本体中深度和子概念数均相同的概念数目, $\min(|\text{Nonleaf}_{o_1}|, |\text{Nonleaf}_{o_2}|)$ 为 2 个本体中非叶子概念的总数的最小值。

在通常情况下,通过计算某元素在树结构中所处的深度及密度来衡量其在结构树中的位置。同时,可通过提取本体的特征信息来进一步判断参与映射的本体对间是否存在更为详细的结构信息。式(8)中分母表示非叶子概念或非叶子属性的两实体中的较小者。

在基于属性的策略中,属性匹配局部置信度计算如下:

$$M_{\text{credibility}}^{\text{property}} = \frac{\text{SameRange}(|c|) \cup |\text{depth}|}{\min(|\text{prop}_{o_1}|, |\text{prop}_{o_2}|)} \quad (9)$$

其中, $\text{SameRange}(|c|) \cup |\text{depth}|$ 为待映射本体中所共有的属性数目, $\min(|\text{prop}_{o_1}|, |\text{prop}_{o_2}|)$ 为待映射本体中属性数目的最小值。

属性匹配局部置信度计算与结构匹配局部置信度计算形式上类似,其局部置信度由基于属性的深度位置和属性所描述的概念来共同确定。

相比基于全局固定权值的相似度计算方法,局部置信度的引入规避了大量非必要计算,且局部置信度能够较好地展现某映射策略在面对具体映射任务时的可信程度。基于局部置信度,对由各策略获取的相似度进行加权,得综合相似度 $\text{sim}(e_i, e_j)$ 为:

$$\text{sim}(e_i, e_j) = \frac{\sum_k M_{\text{credibility}}^k \cdot \text{sim}_k(e_i, e_j)}{\sum_k M_{\text{credibility}}^k} \quad (10)$$

其中, $k \in \{\text{概念, 实例, 结构}\}$, $M_{\text{credibility}}^k$ 和 $\text{sim}_k(e_i, e_j)$ 分别为基于 k 策略的局部置信度和相似度。

$\text{sim}_k(e_i, e_j)$ 的计算采用文献[12]的方法。提取映射结果时,传统方法一般借助于二分图最大指派算法,即采取稳定婚姻匹配法或者匈牙利算法,依据相似度值获取映射关联,但这类方法的实际效率并不理想。本文采用启发式的贪心策略来获取映射结果。具体过程为:

1) 对映射本体对集合中的映射按照 $\text{sim}_k(e_i, e_j)$ 值进行排序, $\text{sim}_k(e_i, e_j)$ 最大者加入到结果集中。

2) 对剩余映射进行如下迭代操作:若该映射中实体与现有映射结果集中实体无交集,则将其纳入映射结果集;若存在交集,则跳过。

3) 在映射完成后进行校正工作,进一步提高映射质量。

2.4 映射校正

对上文得到的综合相似度进行融合,得到本体映射初步结果。但是,该过程中的误差会导致结果中出现不良映射。映射校正的目的是消除不良映射,确保映射结果的精度^[13],具体举措包括在映射中增加基本的语义规约或借助现有相关知识库对映射过程进行辅助。本文通过对大量映射结果进行分析研判,提出如下映射规则:

规则 1 If $C_1^i = C_1^j$ and $C_2^m = C_2^n$, then $C_1^i \neq C_2^m \rightarrow C_1^j \neq C_2^n$ 。

规则 2 If $C_1^i = C_1^j \cup C_1^k$ and $C_2^m = C_2^n \cup C_2^o$, then $C_1^i \neq C_2^m \rightarrow C_1^j \neq C_2^n$ 。

规则 3 If $Rand(P_1^i) \neq Rand(P_2^m)$ and $Domain(P_1^i) \neq Domain(P_2^m)$, then $C_1^i \neq C_2^m$ 。

规则 4 If $Super(C_1^i) \neq Super(C_2^m)$ and $SubClass(C_1^i) \neq SubClass(C_2^m)$, then $C_1^i \neq C_2^m$ 。

利用上述规则对映射结果进行校验,以剔除映射结果集中的不良映射。需要注意的是,映射规则的构建需要技术人员和领域专家的共同参与,且是一个不断迭代、调整、完善的过程。

3 实验结果与分析

本次的本体映射实验在 MyEclipse 平台上采用 Java 语言开发完成。其中,硬件环境配置为:酷睿 i7 CPU,主频 3.60 GHz,内存 4 GB,硬盘 950 GB;软件环境配置为:java 编译环境 jdk8,采用 MyEclipse 9.0 版本。整个实验过程在 Windows 7 系统下进行。

为验证本文算法的有效性,选取 OAEI (Ontology Alignment Evaluation Initiative) 的标准测试集 Benchmarks 进行实验。作为目前公开的本体匹配研究领域权威测试数据集,Benchmarks 具备较全面的测试数据类型,且规模较大,能够对本体映射方法的性能进行较好地测试。Benchmarks 共包含 110 个本体,分为 3 类:1xx,2xx,3xx,分别代表信息完整本体、部分语义缺失本体、现实本体(如 jobs、food、conference 等),分别用于评估算法对信息完整本体的映射能力、处理语义缺失数据的稳定性、处理现实本体的应用能力。实验在 1xx、2xx、3xx 与参考本体之间进行映射,并将标准测试结果与该映射结果进行比较。Benchmarks 数据集的本体分类如表 1 所示。

表 1 Benchmarks 数据集分类

分组	测试集编号	数目	描述
1xx	101-104	3	词汇和结构都相似
	201-210	18	词汇不相似,结构相似
2xx	221-247	18	词汇相似,结构不相似
	248-266	67	词汇和结构都不相似
3xx	301-304	4	现实世界的本体

将映射过程的持续时间作为评价映射效率的重要依据,将查准率 p 、查全率 r 和 F-Measure 作为考察映射结果质量的指标。各指标计算表达式如下:

$$p = |R \cap A| / |A| \quad (11)$$

$$r = |R \cap A| / |R| \quad (12)$$

$$F = 2p \cdot r / (p + r) \quad (13)$$

其中, A 为映射结果中正确的部分, R 为参考映射结果。F-Measure 是建立在 p 和 r 上的一个综合性评

判指标,其值越大表示算法映射效果越优。

表 2 所示为实验中本文算法与同样对本体进行模块化处理的 Falcon 算法^[14]在数据集上的平均运行时间对比。其中,实例、结构、属性的局部置信度阈值分别取 0.8、0.6、0.6。

表 2 2 种算法运行时间比较

算法	1xx	2xx	3xx	总计
Falcon 算法	101	112	78	291
本文算法	92	96	71	259

由表 2 可以看出,在各数据集上,本文算法的映射效率均优于 Falcon 算法。原因如下:

1) 本文算法中的聚类阶段是对 AHSCAN 的优化,降低了本体模块化的时间复杂度。

2) 本文算法借助信息检索技术搜索参考点,与传统参考点提取方法相比压缩了搜索空间,从而缩短了发现相关子本体的时间。

进一步地,将本文算法的实验结果与国内外经典映射方法 OntoDNA^[15]、ASMOV^[16]的映射结果进行比较,结果如图 3、表 3、表 4 所示。

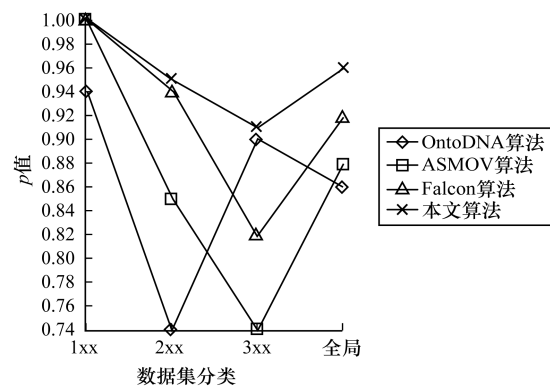


图 3 不同算法的查准率比较

表 3 不同算法的查全率比较

算法	1xx	2xx	3xx	全局
OntoDNA 算法	1.00	0.76	0.78	0.85
ASMOV 算法	1.00	0.82	0.86	0.89
Falcon 算法	1.00	0.87	0.77	0.88
本文算法	1.00	0.85	0.79	0.88

表 4 不同算法的 F-Measure 值比较

算法	F-Measure 值
OntoDNA 算法	0.855
ASMOV 算法	0.885
Falcon 算法	0.900
本文算法	0.914

由图 3、表 3、表 4 可以看出:

1) 对于数据组 1xx,除 OntoDNA 算法外,ASMOV、Falcon 和本文算法均能取得较好效果,能

很好地识别出相互间存有语义关联的实体对。

2) 对于数据组 2xx, 本文算法与 Falcon 算法在 p 值上的测试结果优势明显, ASMOV 结果较好, 而 OntoDNA 的表现较差。在 r 值上, 本文算法为 85%, 与 Falcon 算法存在一定差距, 表明本文算法仍有提升空间。

3) 对于数据组 3xx, OntoDNA 的 p 测试结果表现突出(90%), 但其 r 值仅为 78%。本文算法拥有最高的 p 值, r 值仅劣于 ASMOV。由于 3xx 数据组所含本体中均为实际本体, 因此映射系统在 3xx 上的测试结果能够在很大程度上反映其处理实际数据的能力。因此, 本文算法具备较强的实际数据处理能力。

综上, 相较于其他 3 种算法, 本文算法具备较好的映射能力。

对参与映射的实体规模进行缩减, 在一定程度上解决了传统方法计算量过大的问题。计算实体相似度时, 本文算法充分考虑本体自身特征, 采取多策略自适应组合的模式, 基于规则对映射结果进行校正, 确保了映射精度。实验结果表明, 该算法在确保查全率的基础上, 能够提高映射的查准率和映射质量。

4 结束语

本文提出一种基于模块化和局部置信度的多策略自适应大规模本体映射算法。通过将本体进行模块化和相关性计算, 缩小搜索空间并降低时间复杂度。通过局部置信度计算, 实现对映射策略权值的自适应调整。实验结果表明, 该算法在提高映射效率的同时可以提升映射结果质量。

目前, 机器学习处理大量数据的能力较高, 企业系统长期运行积累了大量数据, 为训练集的构建提供了较好的输入。基于机器学习提高本文大规模本体映射算法的效率与精度将是下一步的研究方向。

参考文献

- [1] 王顺, 康达周, 江东宇. 本体映射综述[J]. 计算机科学, 2017, 44(9): 1-10.
- [2] SHVAIKO P, EUZENAT J. Ontology matching: state of the art and future challenges[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 158-176.
- [3] DARAIO C, LENZERINI M, LEPORELLI C, et al. The advantages of an ontology-based data management approach: openness, interoperability and data quality[J]. Scientometrics, 2016, 108(1): 441-455.
- [4] 蒋湛, 姚晓明, 林兰芬. 基于特征自适应的本体映射方法[J]. 浙江大学学报(工学版), 2014, 48(1): 76-84.
- [5] ALANI H, SAAD S. Schema matching for large-scale data based on ontology clustering method [J]. International Journal on Advanced Science; Engineering and Information Technology, 2017, 7(5): 1790-1797.
- [6] 孙煜飞, 马良荔, 郭晓明, 等. 基于模块化的大规模本体映射方法[J]. 模式识别与人工智能, 2016, 29(5): 410-416.
- [7] 孙煜飞, 马良荔, 周润芝. 一种自适应的多策略本体映射方法[J]. 海军工程大学学报, 2016, 28(2): 75-82.
- [8] YURUK N, METE M, XU X W, et al. AHSCAN: agglomerative hierarchical structural clustering algorithm for networks[C]//Proceedings of IEEE/ACM International Conference on Advances in Social Network Analysis and Mining. Washington D. C., USA: IEEE Press, 2009: 72-77.
- [9] 杨月华, 杜军平, 平源. 基于本体的智能信息检索系统[J]. 软件学报, 2015, 26(7): 1675-1687.
- [10] 李文清, 孙新, 张常有, 等. 一种本体概念的语义相似度计算方法[J]. 自动化学报, 2012, 38(2): 229-235.
- [11] SEDDIQUI M H, AONO M. An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size [J]. Web Semantics: Science, Service and Agents on the World Wide Web, 2009, 7(4): 344-356.
- [12] CALDAROLA E G, RINALDI A M. A multi-strategy approach for ontology reuse through matching and integration techniques[M]. Berlin, Germany: Springer, 2016.
- [13] FARIA D, PESQUITA C, SANTOS E, et al. The agreement maker light ontology matching system[EB/OL]. [2018-08-25]. http://disi.unitn.it/~p2p/RelatedWork/Matching/Feriae_AgreementMakerLight13.pdf.
- [14] HU W, QU Y Z. Falcon-AO: a practical ontology matching system [J]. Web Semantics: Science, Services and Agents on the Word Wide Web, 2008, 6(3): 237-239.
- [15] CHING C K, CHIEN S L. Ontology mapping and merging through OntoDNA for learning object reusability [J]. Educational Technology and Society, 2006, 9(3): 27-42.
- [16] JEAN Y R, KABUKA M R. ASMOV; results for OAEI 2008[EB/OL]. [2018-08-25]. <http://www.dit.unitn.it/~p2p/OM-2008/oeai08paper.pdf>.

编辑 吴云芳