基于遗传规划和主动学习的本体实例匹配*

孙煜飞[1,2]，马良荔[2]，解嘉宇[1]
（1. 中国人民解放军 91635 部队，北京 102249；2. 海军工程大学 电子工程学院，武汉 430033）

摘 要：针对现有实例匹配方法存在的准确率和学习效率不高的问题，提出了一种新的基于遗传规划和主动学习的链接规则学习方法，并用于本体实例匹配。设计了更合理的链接规则表示，并针对链接规则的特点，对遗传规划的初始种群产生、适应度函数和进化算子进行了详细设计。提出了一种考虑样本相关性的主动学习采样策略，使得稀有样本被优先训练。实验结果表明，该方法不仅学习效率更高，而且能够学习出高质量的链接规则，取得了较好的本体实例匹配结果。

关键词：本体匹配；实例匹配；链接规则；遗传规划；主动学习
中图分类号：TP399 **文献标志码**：A **文章编号**：1001-3695(2018)05-1380-06
doi:10.3969/j.issn.1001-3695.2018.05.021

Genetic programming and active learning based ontology instance matching

Sun Yufei[1,2], Ma Liangli[2], Xie Jiayu[1]
(1. *Unit* 91635 *of PLA*, *Beijing* 102249, *China*; 2. *College of Electronic Engineering*, *Naval University of Engineering*, *Wuhan* 430033, *China*)

Abstract: Aim at existing instance matching methods' problem of low precision and learning efficiency, this paper proposed a novel genetic programming and active learning based linkage rule learning method, which was applied to ontologies instance matching. According to the characteristics of a more reasonable expression of linkage rule, this method carried on a detailed design of the initial population, fitness function and evolutionary operator of genetic programming. Meanwhile, it proposed a correlation-aware active learning sampling strategy, which preferred the rare training samples. The experimental results show that the proposed method is not only has higher learning efficiency, but also can learn linkage rules of high quality, which achieve good ontology instance matching results.

Key words: ontology matching; instance matching; linkage rule; genetic programming; active learning

0 引言

本体作为一种领域知识概念化和模型化的方法，被认为是解决语义异构问题的关键途径。随着本体应用的快速发展、本体数量大幅增长，这些本体描述的内容存在重复和关联，但在本体模式上却表现各异。本体匹配旨在发现异构本体中语义关联的实体（概念、属性和实例），并建立语义关联实体间的匹配关系[1]。它对于消除本体异构、实现本体集成和数据融合等具有重要作用。目前，本体匹配已成为语义 Web 领域的一个研究热点。

实例匹配是本体匹配的重要方面，目前受到越来越多的关注[2]。一个本体实例通常包含若干描述该实例的属性及其属性值，基于属性值相似度计算的方法通过比较两个实例属性值之间的异同来识别实例匹配。不同的属性值具有不同数据特征，且不同的属性值对于识别实例匹配的作用不等。因此，利用何种相似度计算方法对两个属性值进行比较、各相似度值如何进行组合以及如何根据上述相似度值判断两个实例是否匹配等问题显得尤为困难。

通常情况下，基于属性值相似度计算的方法可以表示成链接规则[3]（linkage rule），用来声明两个实例匹配应满足的条件。Silk[4]通过人工制定链接规则，后来陆续提出了使用机器学习[5]和主动学习[6]的方法，从已知实例匹配中学习出链接规则，它的一个突出特点是支持属性值的变换。Knofuss[7]利用遗传算法学习链接规则，提出了一种称为伪 F 测度值的适应度函数，因此不需要训练数据，是目前唯一的无监督学习方法，但是该方法被证明在现实世界本体上的应用效果较差[8]。Raven[9]能够学习出线性和布尔型的链接规则，缺陷是不能表示复杂的链接规则。EAGLE[10]结合遗传规划和主动学习方法，能够学习出较复杂的链接规则。COALA[11]研究了未标记样本间的结构信息，并提出了一种考虑样本相关性的主动学习策略，该方法需要进行大量计算，时间复杂度较高。在数据库领域，MARLIN[12]利用 SVM 学习线性规则，动态确定各属性对的权值，该方法的缺陷是它不能表示复杂的链接规则。文献[13]提出利用遗传规划解决记录链接问题，但是该方法的链接规则中运算符由算术表达式组成，不符合现实世界真实情况，难以理解。

鉴于以上研究工作中存在的问题，本文提出了一种新的基于遗传规划和主动学习的链接规则学习算法，并用于实例匹配。设计了一种更合理的链接规则表示法，并针对链接规则的特点，对遗传规划初始种群产生、适应度函数、进化算子进行了详细设计。最后提出了一种考虑样本相关性的主动学习策略，

收稿日期：2016-11-30；**修回日期**：2017-02-01 **基金项目**：总装预研基金资助项目

作者简介：孙煜飞（1988-），男，工程师，博士，主要研究方向为数据集成、本体映射（sharesorrows@163.com）；马良荔（1968-），女，湖北武汉人，教授，博士，主要研究方向为软件可靠性；解嘉宇（1977-），男，副研究员，硕士，主要研究方向为装备管理.

使得稀有样本被优先训练,学习效率更高。实验结果表明,该方法能够在少量训练集条件下,学习出高质量的链接规则,取得较好的实例匹配结果。

1　问题描述

首先用一个简单的例子说明实例匹配和链接规则。如图 1 所示,paper 和 contribution 为两个与论文有关的数据源,它们各包含若干实例,每个实例对应一个 uri 和若干属性值(这里只列出了几个重要属性)。注意,每个数据源原本应为 RDF 三元组描述,为了简化起见,在此用二维表格表示。

paper

uri	title	author	page	published on	date
uri_1	semantic web primer	Tim Berners Lee	12	american science	2000-1-1
uri_2	Ontology matching	Jerome Euzenat	16	TKDE	2013-2-6
uri_3	linked data	Christian Bizer	26	journal of web semantics	2011-8-5
…	…	…	…	…	…

contribution

uri	hasTitle	hasAuthor	Page	publication	Date
uri'_1	ontology matching	Pavel Shvaiko	16	TKDE	2010/11/9
uri'_2	ontology matching	Jerome Euzenat	16	IEEE transactions on knowledge and data engineering	2013/2/6
uri'_3	ontology matching: a survey	Jerome Euzenat	20	ISWC	2008/8/5
…	…	…	…	…	…

图1　实例匹配示例

从图 1 中可以看出,paper 数据源中 uri_2 与 contribution 数据源中 uri'_2 是匹配的,而与 uri'_1、uri'_3 是不匹配的。这可以通过制定一个链接规则来表示匹配应满足的条件,一个可行的链接规则是:如果 title 属性值与 hasTitle 属性值的 Jaccard 相似度不小于 0.7(条件 1);并且 author 属性值与 hasAuthor 属性值的编辑距离相似度和 date 属性值与 Date 属性值的日期相似度的较小值不小于 0.8(条件 2),则这两个实例是匹配的。这可以表示为树型结构,如图 2 所示。图中灰色矩阵为运算符,白色矩阵为终止符。

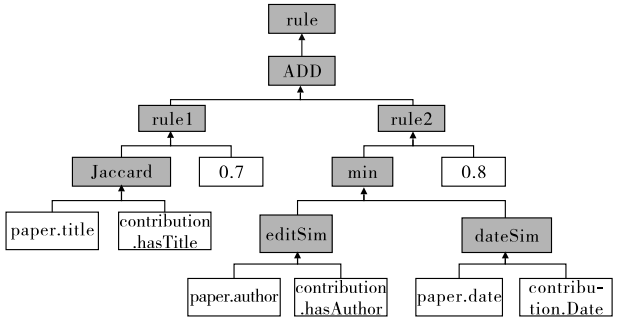

图2　可行的链接规则

参考文献[5],给出实例匹配和链接规则的形式化描述如下。

定义 1　实例匹配。给定两个数据源 S 和 T,实例匹配旨在发现 S 和 T 中指称现实世界相同对象的具有不同 URI 的实例:

$$M=\{(s,t)\mid s\in S,t\in T,s\sim_R t\} \tag{1}$$

其中,关系 $\sim_R$ 表示实例之间的匹配。

定义 2　链接规则。一条链接规则 r 试图根据两个实例的属性值相似度判断它们是否匹配:

$$r:S\times T\rightarrow[-1,+1] \tag{2}$$

$r(s,t)\rightarrow+1$,表示 s 和 t 匹配,即 $s\sim_R t$;否则为不匹配。

定义 3　链接规则学习。给定一组训练数据,包括匹配实例集 R^+ 和不匹配实例集 R^-,其中 $R^+\subseteq M$、$R^-\subseteq(S\times T)\setminus M$,可以从中学习出链接规则,该链接规则能够最大满足给定训练数据,同时具有很强的泛化能力:

$$R^+\cup R^-\rightarrow r \tag{3}$$

2　基于遗传规划的链接规则学习

遗传规划[14](genetic programming,GP)是在遗传算法的基础上发展起来的一种群体智能优化算法,它利用选择、交叉和变异三个基本遗传算子模拟自然界物竞天择的进化过程。与遗传算法不同的是,GP 对可行解采用树型结构进行编码,相对于遗传算法的定长编码具有更好的适用性。同时,由于其强大的启发式搜索寻优能力,目前已广泛应用于风险评估、作业调度、预测等任务中。

链接规则学习本质上是一个全局优化问题,它试图从满足要求的解空间中寻找最佳的可行解,且这些可行解具有树型结构,因此可以利用 GP 来解决链接规则学习问题。基于遗传规划的链接规则学习算法(GPlink)的描述如下:

输入:种群大小 N、最大进化代数 T、交叉概率 p_c、变异概率 p_m、精英保留比率 p_e、初始种群 P^0。
输出:末代种群。
1 for($t=1;t\leqslant T;t++$)
2　计算种群 P^{t-1} 中每一个体的适应度;
3　while($|P'|\leqslant N$) do
4　　从 P^{t-1} 中选择两个个体 r_1 和 r_2;
5　　以交叉概率 p_c 对 r_1 和 r_2 进行交叉操作;
6　　以变异概率 p_m 对 r_1 和 r_2 进行变异操作;
7　　将产生的两个子代加入新种群 P';
8　end while
9 用 P^{t-1} 中适应度最高的 p_eN 个体替换掉新种群 P' 中适应度最差的 p_eN 个体;
10 end for
11 return P^t

种群中个体必须为有效的链接规则(详见 2.1 节),算法首先产生初始种群(详见 2.2 节),在每一轮迭代中,计算当前种群中每个个体的适应度(详见 2.3 节),作为个体被选择的依据。然后完成复制、选择、交叉和变异四个基本遗传操作(详见 2.4 节)。末代种群作为算法的输出。

2.1　有效的链接规则

实际上,图 2 已经给出了有效链接规则的表示。在此给出其形式化说明,首先定义如下记号。

a)property(简称 p),获取实例的属性值。

b)comparison(简称 com),comparison 的输入必须是两个属于不同实例的 property,输出为[0,1]内的相似度值。

c)comparison operator(简称 comOp),对若干个 comparison 值或 comOp 值进行组合,输出为[0,1]内组合相似度值。comOp 允许嵌套,即一个 comOp 的输出可以作为另外一个 comOp 的输入。常见的组合函数有 max(最大值)、min(最小值)、

weighted mean(加权平均值),当组合函数为 weighted mean 时,通常还需指定各输入参数的权值。

d)由于 com 和 comOp 的输出都为区间数[0,1],所以被合称为相似度算子(记为 m)。

e)aggregation operator(简称 aggOp),对若干布尔值进行操作,输出仍为布尔值。aggOp 允许嵌套,即一个 aggOp 的输出可以作为另外一个 aggOp 的输入。常见的组合函数有 ADD(且)、OR(或)等。

定义 4 原子链接规则。一条链接规则称为原子链接规则(atomicRule),当且仅当它由一个相似度算子 m 和一个阈值 θ 组成(当 m 返回值大于等于 θ 时,匹配成立)。

定义 5 有效链接规则。一条有效链接规则(validRule,简称 rule)可以由下式得到:

a)rule = atomicRule;

b)rule = aggOp(rule$_1$,rule$_2$,…,rule$_n$)

其中 b)称为复合链接规则。

2.2 初始种群的产生

标准 GP 算法采用 ramped-half-and-half 方法产生初始种群[14],其中一半个体由单个终止符组成,另一半个体由深度为最大树深(事先确定的参数)的满树组成。针对链接规则的特点,采用类似的方法产生初始种群,第一部分中个体按照下列步骤产生:

a)从数据源 S 和 T 的属性匹配 Mp 中随机选择一组属性对(p_1,p_2),作为待比较的属性对。

b)随机选择一个相似度度量方法 com,作为上述属性对的相似度计算方法。

c)重复步骤 a)和 b)。

d)随机选择一种组合方法 comOp,并产生两个(0,1]区间内的随机数 w_1 和 w_2,作为上述两个 com 的组合方法。

e)产生一个(0,1]区间内的随机数 θ_1,与 comOp 一起组成链接规则 rule1。

f)重复上述步骤 a)~e),生成链接规则 rule2。

g)随机选择一种组合方法 aggOp,组合 rule1 和 rule2,生成最终链接规则。

第二部分中个体前两个步骤同 a)b),最后产生一个(0,1]区间内的随机数 θ'_1,生成最终链接规则。

第一部分个体为包括所有元素的满树,第二部分个体为原子链接规则,两部分个体数目相等,这样可以快速区分各属性对的重要程度。由此产生的初始种群中个体表现形式如图 3 所示。

值得注意的是,上述步骤中假设已知两个数据源的属性匹配,这样可以大大减少算法的搜索空间。如果属性匹配未知,则可以通过本体匹配技术发现属性匹配,或者随机选择两个属性。

虽然 com、comOp 和 aggOp 函数可以为任意有意义的函数,在本章实现中只利用了以下函数,分别如表 1~3 所示。

2.3 适应度函数

给定一组训练集,其中包括匹配集 R^+ 和不匹配集 R^-,对于实例集合 A 和 B:

$$A = \{a \mid \exists t \in T, (a,t) \in R^+ \cup R^-\} \tag{4}$$

$$B = \{b \mid \exists s \in S, (s,b) \in R^+ \cup R^-\} \tag{5}$$

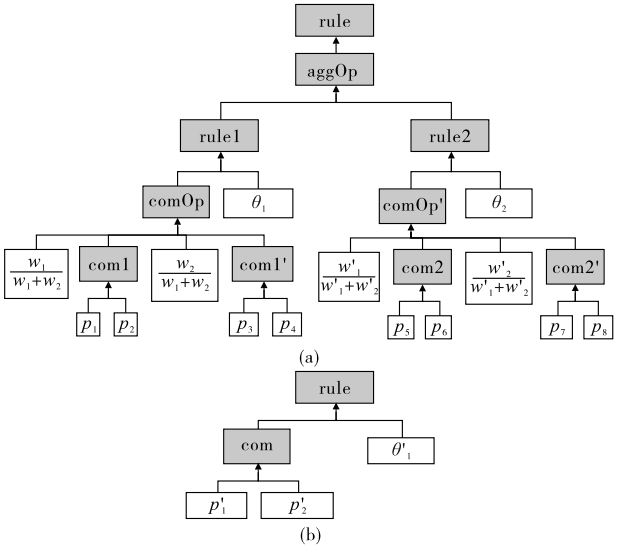

图3 初始种群中个体表现形式

表 1 实现的 comparison 函数

名称	描述
editSim	基于编辑距离的相似度
Jaccard	Jaccard 相似度
triGram	triGram 相似度
dateSim	日期相似度
numericSim	数值相似度

表 2 实现的 comOp 函数

名称	描述
max	取最大值
min	取最小值
weighted mean	取加权平均值

表 3 实现的 aggOp 函数

名称	描述
ADD	交
OR	或

利用链接规则 r 对 A 和 B 进行匹配(值得注意的是,当属性值为另外一个实例时,取属性值为该实例的名称,当属性值为空白节点时,取属性值为该空白节点的所有属性值的并。最后对所有属性值进行分词并转换成小写等规范化处理),得到预测匹配实例集 M^+。利用 F_1-measure 值来评价链接规则 r 的性能:

$$F_1 = \frac{2pr}{p+r}\left(p = \frac{M^+ \cap R^+}{M^+}, r = \frac{M^+ \cap R^+}{R^+}\right) \tag{6}$$

遗传规划在不断迭代过程中,逐渐会产生越来越复杂的个体,这些个体中有些表达式是冗余的,它们对输出结果没有影响,但是会造成算法效率变低。在遗传规划中这种冗余的表达式称为基因内区[15]。为了降低基因内区的消极作用,对节点数较多的个体施加一个惩罚函数,作为个体的适应度函数:

$$F_{\text{fitness}} = F_1 - \lambda \times N \tag{7}$$

其中:N 为个体的节点数;λ 为惩罚因子。

2.4 进化

在遗传规划的进化过程中,除了精英保留策略(复制),通过选择、交叉和变异三个基本遗传算子使得种群中个体不断优化,从而逐步逼近最优解。

1)选择　采用锦标赛选择法,每次随机选择 k 个个体,取其中适应度最大值。锦标赛选择法在遗传规划中是一种常用的选择方法。

2)交叉　传统遗传规划采用子树交叉法,考虑到链接规则的特殊性,采用子树交叉会造成许多无效的链接规则,在此设计了四种特殊的交叉操作,算法随机选择其中一种进行交叉。

a)子树交叉 1,从个体 r_1 中随机选择一个以 com 或 comOp 为根节点的子树,然后从个体 r_2 中随机选择一个以 com 或 comOp 为根节点的子树,对两者进行子树交叉,如图 4 所示;b)子树交叉 2,从个体 r_1 中随机选择一个以 rule 为根节点的子树,然后从个体 r_2 中随机选择一个以 rule 为根节点的子树,对两者进行子树交叉,如图 5 所示;c)子树合并 1,从个体 r_1 中随机选择一个以 comOp 为根节点的子树,然后从个体 r_2 中随机选择一个以 comOp 为根节点的子树,对两者进行子树合并,然后,每个输入参数有 50% 的概率被删除,如图 6 所示;d)子树合并 2,从个体 r_1 中随机选择一个以 aggOp 为根节点的子树,然后从个体 r_2 中随机选择一个以 aggOp 为根节点的子树,对两者进行子树合并,然后,每个输入参数有 50% 的概率被删除,如图 7 所示。

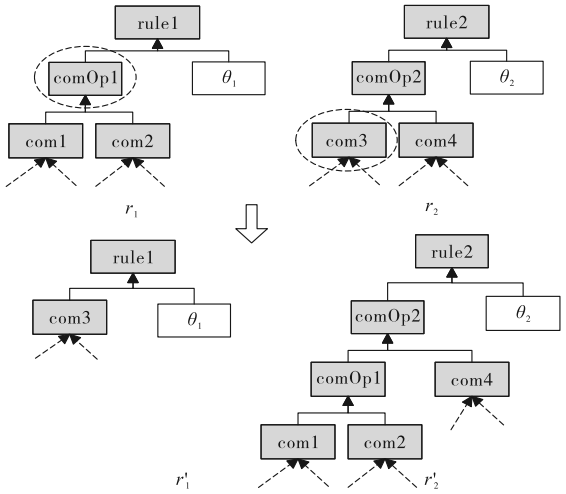

图4　子树交叉1

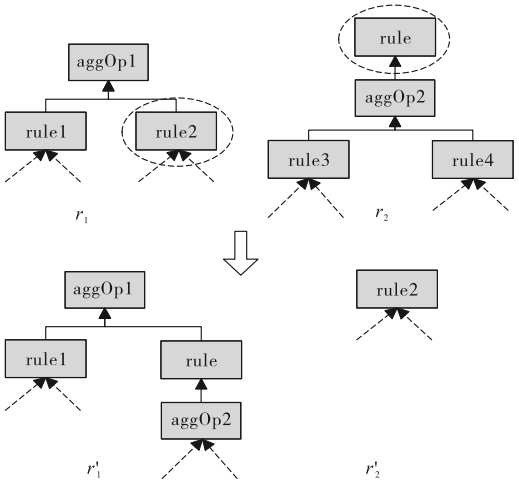

图5　子树交叉2

子树交叉 1 使得 comOp 得以嵌套,子树交叉 2 使得 aggOp 得以嵌套,子树合并 1 使得 comOp 对多个属性值相似度进行组合,子树合并 2 使得 aggOp 对多个规则进行组合。

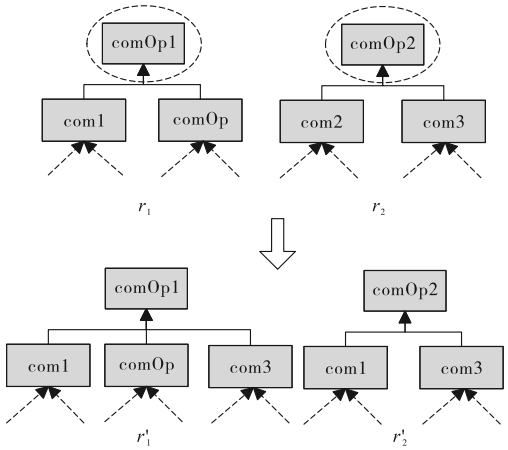

图6　子树合并1

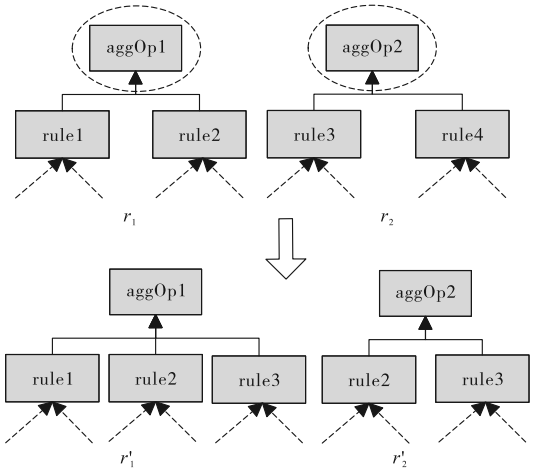

图7　子树合并2

3)变异　变异是产生新个体的主要步骤,能够防止算法早熟收敛。考虑到链接规则的特殊性,设计了以下五种变异操作,算法随机选择其中一种进行变异。

a)从个体 r 中随机选择一个 aggOp 节点,替换成任意其他 aggOp 函数;b)从个体 r 中随机选择一个 comOp 节点,替换成任意其他 comOp 函数;c)从个体 r 中随机选择一个 com 节点,替换成任意其他 com 函数;d)从个体 r 中随机选择一个权值节点,假设该值为 w,然后产生一个区间(0,1]内随机数 d,最后将该值替换成 $(w+d)/2$;e)从个体 r 中随机选择一个阈值节点,假设该值为 θ,然后产生一个区间(0,1]内随机数 d',最后将该值替换成 $(\theta+d')/2$。

3　主动学习采样策略

首先介绍 EAGLE[10] 中基于委员会投票法(query by committee,QBC)的主动学习采样策略。在 GP 算法中,种群中每个个体都可以视为一个分类器,因此种群可以视为委员会,基于 QBC,当所有分类器对某个未标记样本的不一致程度达到最大时,选择该样本并进行标记能够获得最大性能提高。对于未标记样本 (s,t),其中 s、t 分别为属于两个数据源的实例,假设当前种群为 $P=\{I_1,I_2,\cdots,I_n\}$,其中每一个体对样本 (s,t) 的预测标记为 $I_i(s,t)(i=1,2,\cdots,n)$,则样本 (s,t) 被选择的依据是

$$\delta=\underset{(s,t)\in U}{\operatorname{argmax}}\left(n-\left|\{I_i:I_i(s,t)\to+1\}\right|\right)\times\left(n-\left|\{I_i:I_i(s,t)\to-1\}\right|\right) \tag{8}$$

该方法视种群中每一个体为一委员会成员参与投票过程。然而,种群中很多个体的基因是类似的,不满足委员会投票法

中基于参数不同假设的前提。并且在初始迭代中,多数个体的预测置信度较低,如果它们参与投票,则会导致不一致性判断不准确,影响主动学习的采样过程。

下面通过一个例子给出本文主动学习采样策略的思路。

如图 1 所示的两个有关论文的数据源之间进行实例匹配,在大部分情况下,通过比较(paper. title, contribution. hasTitle)两个属性(下文省略数据源名称)就可以确定两个实例是否匹配。在少数情况下,即使(title, hasTitle)相等,两个实例也是不匹配的(图中 uri_2 与 uri'_1),这可以通过增加(author, hasAuthor)属性比较来解决。然而,在极少数情况下,属性对(title, hasTitle)相似,且(author, hasAuthor)相似,两个实例仍不匹配(图中 uri_2 与 uri'_3),因此还需要进一步增加其他属性比较(如 date 与 Date)。由于后两种情况非常稀少,在训练集一定的条件下,通常很难包含这样的训练数据,导致学习出的链接规则无法达到最优的泛化性能。

如果能在训练集中加入这些稀有的、有价值的训练数据,不仅能减少人工标注的工作量,还能学习出泛化能力更强的链接规则。基于此,给出本文主动学习采样策略如下:

对于样本 $x=(s,t)$,假设数据源的属性匹配为 $M_p=\{m_1, m_2, \cdots, m_n\}$,定义实例对$(s,t)$的相似度向量为$(s,t)$在各匹配属性上的相似度值(字符串属性默认用 editSim 计算,数字型属性默认用 numericSim 计算,见表 1):

$$S_x=(\mathrm{sim}_1, \mathrm{sim}_2, \cdots, \mathrm{sim}_n) \tag{9}$$

定义两个样本的距离为对应相似度向量的 Jensen-Shannon 散度:

$$D_{JS}(S_x \parallel S_y)=H\left(\frac{S_x+S_y}{2}\right)-\frac{H(S_x)+H(S_y)}{2} \tag{10}$$

其中,$H(S)$表示信息熵:

$$H(S)=-\sum_{i=1}^{n}\mathrm{sim}_i \log \mathrm{sim}_i \tag{11}$$

假设有标记样本集为 L,未标记样本集为 U,本文的主动学习采样策略为

$$\delta(u)=\arg\max_{u\in U}\arg\min_{l\in L} D_{JS}(S_u \parallel S_l) \tag{12}$$

该策略选择这样的未标记样本 u,它使得与有标记样本集的最小 JS 散度达到最大,即相似度向量与当前有标记样本的相似度向量最不一致时,样本被选择。综上,结合主动学习和遗传规划的链接规则学习算法(ActiveGPLink)如下:

输入:人工标注最大代价 H,每轮采样的样本数 m,有标记样本集 L,未标记样本集 U。
输出:最优链接规则 r。
1 $h=0, L=\varnothing$;
2 while($h\leqslant H$)
3 按照式(3.5)从 U 中选择 m 个样本,记为 E;
4 $L\leftarrow L\cup E, U\leftarrow U\setminus E$;
5 在 L 上运行 GPLink 算法,返回值为 P^t;
6 $P^0\leftarrow P^t, h\leftarrow h+k$;
7 end while
return 种群中适应度最高的个体 r

ActiveGPLink 算法在每轮迭代中,选择 m 个最有价值的未标记样本来更新训练集,并运行一次 GPLink 算法,末代种群作为下次 GPLink 算法的输入。当达到最大标注代价时,ActiveGPLink 算法终止。

4 实验分析

为了验证 ActiveGPLink 算法的有效性,用 Java 实现了该算法,其中 GP 算法基于 ECJ[16] 开发,ECJ 是一个关于进化计算的 Java 开源库。实验运行环境为 Inter Xeon E3-1231 v3 CPU、8 GB 内存、JDK1.7、Windows 7 操作系统。

4.1 实验设计

实验数据集选用 OAEI 2010 PR 数据集[17] 和 OAEI 2011 NYT 数据集[18],PR 数据集包括 person1、person2 和 restaurant 三个测试,描述有关人和餐馆的实例数据,NYT 数据集旨在建立 NewYorkTimes 与 DBPedia、Freebase 和 Geonames 之间的链接,分为三种类型 people、organizations、locations,共七个测试(NYT 与 Geonames 只有 locations 链接),实验选取 NYT-DBPedia 的三个测试,因为参与 NYT 的实例匹配系统在该测试上的平均表现最差。OAEI 组织者为每个数据集提供了参考匹配 R,方便评测实例匹配的结果。数据集的基本信息如表 4 所示。

表 4 数据集基本信息

数据集	$\lvert S\rvert$	$\lvert T\rvert$	$\lvert R\rvert$
person1	2 000	1 000	500
person2	2 400	800	400
restaurant	339	2 256	112
NYT-DBPedia-loc.	3 840	1 920	1 920
NYT-DBPedia-org.	6 088	1 949	1 949
NYT-DBPedia-peo.	9 958	4 977	4 977

所有的实例对$(s,t)\in S\times T$组成实验数据集,在参考匹配中,随机选取 10 条作为初始正训练集 L^+,对于两个正训练集 $(a,b)\in L^+$和$(c,d)\in L^+$,生成两个负训练集(a,d)和(c,b)。剩下的作为未标记样本集。实验结果用整个数据集上的 F-measure 值进行度量。每次实验运行五次,取平均值。

共设计了两个实验。

实验 1 对 ActiveGPLink 与参加 OAEI 竞赛的实例匹配系统进行匹配结果比较,包括 ASMOV、CODI、LN2R、ObjectCoref、RiMOM、AgreementMaker、SERIMI 和 Zhishi. links。前面五个系统参加了 OAEI 2010 PR 测试,后面三个参加了 OAEI 2011 NYT 测试。

实验 2 为了说明 ActiveGPLink 主动学习策略的优势,与批量学习、随机采样法和委员会投票法进行比较。

实验参数设置如表 5 所示。

表 5 实验参数设置

参数	说明	取值
H	最大标注代价	100
m	每轮采样数	10
N	种群大小	100
T	最大进化代数	30
k	锦标赛大小	5
P_c	交叉概率	0.6
P_m	变异概率	0.2
P_e	精英保留比率	0.05

4.2 实验结果与分析

图 8 和 9 给出了 PR 数据集和 NYT-DBPedia 数据集上的实验结果。从图 8 和 9 可以看出,ActiveGPLink 方法在六个测试中的四个获得最高 F-measure 值,在剩下两个测试中获得第二 F-measure 值。可见,基于遗传规划的方法能够学习出精确

的链接规则,相对于其他基于相似度的方法,能够取得更好的实例匹配结果。注意这个对比是不公平的,其他方法没有使用训练集,而 ActiveGPLink 是在有训练样本的条件下学习的结果。

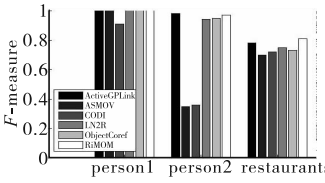

图8　PR数据集实验结果

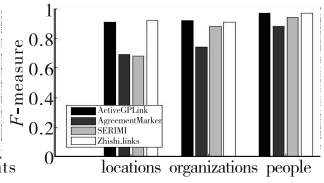

图9　NYT-DBPedia数据集实验结果

为了说明批量学习所需的标注代价,图 10 显示了在 NYT-DBPedia-locations 数据集上,在不同的标注规模上运行 GPLink 算法,用末代种群中适应度最高的个体进行实例匹配的 *F*-measure 值。

从图 10 可以看出,批量学习所需的标注代价远大于主动学习所需的标注代价,当训练数据达到 1 000 时,*F*-measure 值才达到 90% 以上,而主动学习达到这一目标所需的标注代价不到 100。这说明只有当训练数据达到一定规模时,稀有训练数据才会被训练,从而才能学习出有判别力的属性对。

为了说明 ActiveGPLink 的主动学习策略的有效性,将本文主动学习采样策略与随机采样和基于委员会投票法(分别记为 random 和 QBC)采样进行比较,其他保持不变,并对它们进行比较。在每轮迭代后,选取种群中适应度最高的个体进行实例匹配,作为当前迭代的实验结果。图 11 显示了在 NYT-DBPedia-locations 数据集上,两种策略各自运行五次的平均 *F*-measure 值和标准差。

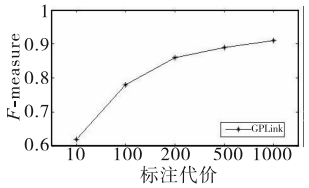

图10　批量学习所需的标注代价

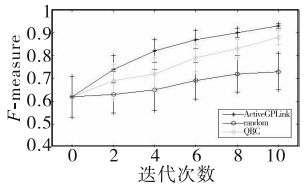

图11　各采样策略的学习曲线对比

从图 11 可以看出,随机采样策略的 *F*-measure 值提升缓慢,迭代 10 次后甚至比本文方法迭代两次的结果还差,说明样本中存在大量的冗余,新增样本无法有效提升链接规则学习的结果。基于委员会投票法的采样策略相比随机采样具有明显优势,但是比本文方法稍差,说明由种群中个体组成委员会存在一定的局限性。本文主动学习策略在前几次迭代中,*F*-measure 值得到迅速上升,说明有价值的样本被优先选择并训练,通过遗传规划学习出的链接规则能够获得更好的实例匹配结果。最后,本文主动学习策略的标准差小于随机采样的标准差,说明主动学习策略相对于随机采样波动较小,算法更稳定。

5　结束语

本文提出了一种新的基于遗传规划和主动学习的链接规则学习方法,并用于本体实例匹配;设计了更合理的链接规则树型表示方法,并针对该链接规则表示方法的特点,对遗传规划的初始种群产生、适应度函数和进化算子进行了重新设计;设计了一种新的主动学习策略,优先选择有价值的样本进行训练,有效地减少遗传规划所需的人工标注代价。本文方法不仅学习效率更高,还能学习出高质量的本体实例匹配链接规则,取得了较好的本体实例匹配结果,能够应用于大规模实例匹配。后续将进一步对可用函数集进行扩展,并应用于更多真实数据集。

参考文献:

[1] Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges[J]. IEEE Trans on Knowledge and Data Engineering,2013,25(1):158-176.

[2] 胡伟,柏文阳,瞿裕忠. 语义 Web 中对象共指的消解研究[J]. 软件学报,2012,23(7):1729-1744.

[3] Ngomo A N. On link discovery using a hybrid approach[J]. Journal on Data Semantics,2012,1(4):203-217.

[4] Volz J, Bizer C, Gaedke M, *et al*. Discovering and maintaining links on the Web-ISWC of data[C]//The Semantic Web. Berlin: Springer-Verlag,2009:650-665.

[5] Isele R, Bizer C. Learning expressive linkage rules using genetic programming[J]. Proceedings of the VLDB Endowment, 2012, 5(11):1638-1649.

[6] Isele R, Bizer C. Active learning of expressive linkage rules using genetic programming [J]. Web Semantics Science Services & Agents on the World Wide Web,2013,23(4):2-15.

[7] Nikolov A, d'Aquin M, Motta E. Unsupervised learning of link discovery configuration[C]//The Semantic Web: Research and Applications. Berlin: Springer-Verlag,2012:119-133.

[8] Ngomo A C N, Lyko K. Unsupervised learning of link specifications: deterministic vs. non-deterministic[C]//Proc of the 8th International Conference on Ontology Matching. 2013:25-36.

[9] Ngomo A C N, Lehmann J, Auer S, *et al*. Raven-active learning of link specifications[C]//Proc of the 6th International Workshop on Ontology Matching. 2011:25-37.

[10] Ngomo A C N, Lyko K. EAGLE: efficient active learning of link specifications using genetic programming[C]//The Semantic Web: Research and Applications. Berlin: Springer-Verlag,2012:149-163.

[11] Ngomo A C N, Lyko K, Christen V. COALA: correlation aware active learning of link specifications[C]//The Semantic Web: Semantics and Big Data. Berlin: Springer-Verlag,2013:442-456.

[12] Bilenko M, Mooney R J. Adaptive duplicate detection using learnable string similarity measures[C]//Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press,2003:39-48.

[13] De Carvalho M G, Laender A H F, Goncalves M A, *et al*. A genetic programming approach to record deduplication[J]. IEEE Trans on Knowledge & Data Engineering,2012,24(3):399-412.

[14] Koza J R. Genetic programming: on the programming of computers by means of natural selection[M]. Cambridge,MA:MIT Press,1992.

[15] 卢少华,云庆夏,夏安邦. 影响遗传规划基因内区的因素分析[J]. 系统工程理论与实践,2003,23(2):101-105.

[16] Luke S, Panait L, Balan G, *et al*. ECJ: a Java-based evolutionary computation research system[EB/OL]. 2015-11-23. http://cs.gmu.edu/~eclab/projects/ecj/.

[17] Euzenat J, Ferrara A, Meilicke C, *et al*. Results of the ontology alignment evaluation initiative [C]//Proc of the 5th International Conference on Ontology Matching. 2010:85-117.

[18] Euzenat J, Ferrara A, Hage W R, *et al*. Final results of the ontology alignment evaluation initiative [C]//Proc of the 6th International Conference on Ontology Matching. 2011:81-110.