基于属性模式的实体识别框架

何峰权，李建中

（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘　要： 很多领域都面临实体识别问题，但现有解决框架缺乏通用性。提出了一种基于属性模式的领域无关的实体识别框架。属性的模式代表属性与实体的一种关系，将模式分为四种类型分别处理，针对类型特点提出了更为通用的相似度计算方法。系统根据模式类型决定相似度计算策略，使系统具有更强的扩展性。该框架可以有效综合利用各类属性的特点进行实体识别，结果优于一般的基于属性特征或基于实体关系的方法。

关键词： 实体识别；属性模式；扩展性；框架

中图分类号： TP319.9　　**文献标识码：** A　　**文章编号：** 2095－2163(2014)01－0065－04

An Entity Resolution Framework based on Attribute Patterns

HE Fengquan，LI Jianzhong

（School of Computer Science and Technology，Harbin Institute of Technology，Harbin 150001，China）

Abstract: Entity resolution problem involves wide range of fields，nevertheless the existing frameworks lack versatility. A domain－independent entity recognition framework based on the attribute pattern is proposed in the paper. Attribute pattern represents the relationship between entity and attribute，which is divided into four types. The paper proposes a more general similarity calculation method against attribute pattern features. Similarity calculation strategy is determined according to attributes' pattern and format automatically，so the system has strong scalability. The framework can effectively utilize characteristics of various types of attributes，and it is proved more efficient than any approach based on attributes characteristics or based on entity relationship.

Key words: Entity Resolution；Attribute Patterns；Scalability；Framework

0　引　言

实体识别就是判别来自一个数据源或多个数据源的描述是否指向同一个实体。此问题由来已久，现已提出很多方法。解决实体识别问题所利用的信息可分为两类，属性特征信息和关系信息。基于属性特征的方法最简单、使用得也最多，但却因属性信息有限，在某些情况下并不足以提供高置信度的判断结论。越来越多的方法开始利用属性的关系或规则进行实体识别，但利用这种关系的方式却各不相同，导致缺乏通用性。对每个实体识别问题都需要重新设计解决方案也必将是低效的，因而需要开展研究，予以改进。

本文将不同属性与实体的关系模式概括为四种类型，通过模式类型决定相似度计算策略，再根据属性的格式决定基本的相似度计算函数。系统将多个属性的相似度组织成向量的形式表示，通过监督学习的方法形成判决器，最后在实体关系图上完成迭代划分。

1　相关研究

文献[1,2]研究了相似函数选择和阈值确定问题。通过发现相似函数和阈值的冗余，去除不合适的相似函数和阈值设置。为了有效整合多种方法的优点，文献[3]提出了一种按有监督学习的结果聚类分配权重的方法，为权重分配提供了新的思路，但选择作为聚类的特征是经验性的，是否可以推广尚未确定。文献[4]设计了一个领域无关的实体识别系统，可以通过学习的方式对数据的格式进行转化，以满足识别系统进行比较的需要。文献[5]研究了利用合作者集合的相关性的方法，实验证明其优于一般的非整体分析的方法。

2　基于属性模式的实体识别框架介绍

系统结构如图1所示，主要分为以下几个部分：

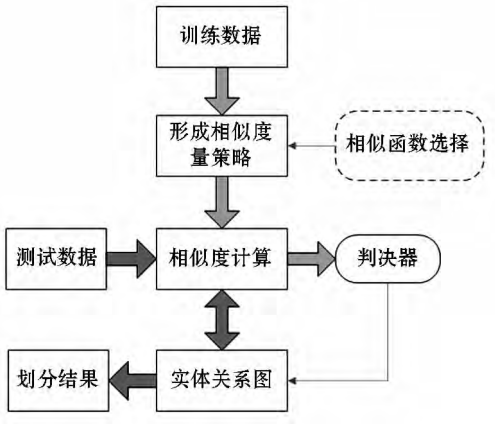

图1 基于属性模式的实体识别系统框架图

Fig. 1 The ER framework based on attribute patterns

收稿日期：2013－06－24
基金项目：国家自然科学基金(61003046)。
作者简介：何峰权(1988－)，男，四川内江人，硕士研究生，主要研究方向：数据质量管理、实体识别；
李建中(1950－)，男，黑龙江哈尔滨人，教授，博士生导师，主要研究方向：数据库研究、传感器网络、数据质量管理等。

(1) 相似度度量策略形成模块。该模块通过属性的模式和数据格式自动地选择相似度函数,形成相似度度量策略。

(2) 相似度计算模块。该模块按照选择的相似度函数计算实体对的相似度。

(3) 判决器模块。该模块在训练阶段统计实体对的相似度分布情况,在实体划分阶段辅助判断。

(4) 实体关系图。实体划分阶段在实体关系图上迭代进行,每次完成实体合并以后,重新计算经过调整的实体对的相似度,直到所有相似边都处理完毕,实体划分结束。

3 系统各部分的实现

实体识别问题模型可描述为,假设 $A = [a_1 a_2 \cdots a_n]$ 是一个关系表的属性集合,r 表示 R 中一条记录,$r[a_i]$ 表示记录 r 的属性 a_i 的值。需要识别的对象称为实体,实体集合为 $O = \{o_1 o_2, \cdots, o_{|o|}\}$,记录中的每个属性是实体的一个描述,一条记录描述的实体表示为 $r.o$。实体识别的过程一般是通过比较两条记录的相似度 $sim(r_i, r_j)$,判断记录指代的是否为同一实体,即 $r_i.o$ 和 $r_j.o$ 是否相同。

3.1 相似度计算策略的形成

为了实现系统的通用性,相似度计算策略必须领域无关地进行。为此分析了实体与属性间的关系,按其特点进行了分类。利用各属性的模式可以确定相似度计算的方法。

实体的属性对于实体有特定的含义。如汽车作为实体,汽车的名称、颜色是汽车的固有特征,不会发生变动,而汽车的价格却可能发生变化,车牌号是与汽车一一对应的标示。论文作者作为实体,作者名称是作者的固有特征,而论文名、论文发表的会议、合作者、时间等都可能发生变化,如图2所示。

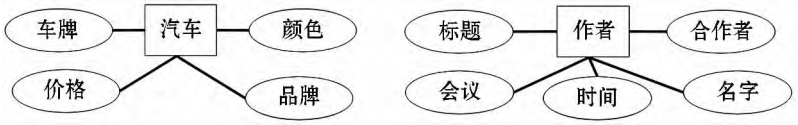

图2 实体和属性举例

Fig. 2 Samples of entities and attributes

定义1 属性的模式 实体识别中,实体属性与实体数量的对应关系称为属性的模式。

类似关系数据库中实体之间的数量对应关系,本文中的模式也分为4类。不同的模式决定了该属性在实体识别中的作用,如表1所示。表中,Agree/DisAgree 表示属性值是否一致,Yes/No 表示肯定的是/否判断。

表1 属性的模式

Tab. 1 Attribute patterns

属性模式	实体:属性	Agree	DisAgree
I	1:1	Yes	No
II	N:1	Possible yes	No
III	1:M	Yes	Unknown
IV	N:M	Possible yes	Unknown

实体与属性间的关系判断过程如下:

(1) if $r_i[a_k] = r_j[a_k]$ then $r_i.o = r_j.o$, else $r_i.o \neq r_j.o$。属性值相同则一定是同一实体,不同则一定不是同一实体。如表示实体身份的 *ID*,相当于主键的作用,实体识别问题正是缺少这样的属性值。

(2) if $r_i[a_k] = r_j[a_k]$ then $r_i.o \sim r_j.o$, else $r_i.o \neq r_j.o$。属性值相同可能是同一实体,不同则不是同一实体。比如汽车的颜色,每辆汽车的颜色是固定的一种,但不是独有的,其他汽车也可能有这样的颜色。此类属性是实体的固有特征,可以通过其不同来区分两个实体。

(3) if $r_i[a_k] = r_j[a_k]$ then $r_i.o = r_j.o$, else $r_i.o \neq r_j.o$。属性值相同则是同一实体,不同则无法判断是否同一实体。比如研究者撰写的论文,一个人会写多篇论文,且论文是没有重复的。此类属性为实体特有,且实体还可能拥有多个此属性值,主要在其相同时发挥作用。

(4) if $r_i[a_k] = r_j[a_k]$ then $r_i.o \sim r_j.o$, else $r_i.o \neq r_j.o$。属性值相同可能是同一实体,不同则无法判断。比如研究者参加会议,一个研究者可能参加很多会议,这个会议也有很多不同的研究者参加。类似的属性还有研究者的合作者。

按照属性的模式类型可以给出更通用的相似度计算方法。类型 I 和类型 II 中,属性与实体对应关系简单,直接比较属性值是否精确匹配,即可对实体是否匹配做出精准判别。类型 III 和类型 IV 中,属性与实体间是多对一和多对多的关系,即一实体含有的此类属性值能构成一个集合。简单的精确匹配已经不能充分利用属性值包含的信息。但在通常情况下,集合内的属性值之间是彼此联系的,比如论文、会议的领域性,合作者间的群体性等。利用这类关系信息可以发现更多实体间的相似性。杰卡德相似度可以用来衡量集合之间的联系。但仍在预期随着实体集合的扩大,能有更多相关性得到发现,式(1)就可以保证当其中一个集合扩大后,两集合的相关性不会因此减小。具体公式为:

$$S_{ja}(S_1, S_2) = \frac{|S_1 \cap S_2|}{\min(|S_1|, |S_2|)} \tag{1}$$

若实体识别的对象是论文作者,则论文名是第 III 类属性。若两篇论文的领域比较接近,则其更可能是同一实体。文章的标题含有多个单词,需要过滤掉其中的非关键字,而标题的关键字集合为 T,可通过式(1)计算论文题目之间的关联性。

计算第 IV 类属性的相关度主要分为两部分,第一部分是求得单值之间的相关性,第二部分是求取集合的相关性。

会议是实体的第 IV 类属性。若两研究者参加的是同一领域的会议,则其更可能为同一实体。若会议包含的作者集合为会议 c_i 和 c_j 间的相关性 $s_c(c_i, c_j)$,其计算如式(2)所示,最终作者 a_1 和 a_2 参加的会议集合间的相关性 $f_C(C_1, C_2)$,其计算则如式(3)所示:

$$s_c(c_i, c_j) = \begin{cases} 1, c_i = c_j \\ s_{ja}, c_i \neq c_j \end{cases} \tag{2}$$

$$f_C(C_1,C_2)=\frac{\sum_{c_i\in C_1,c_j\in C_2}s_c(c_i,c_j)}{\min(|C_1|,|C_2|)} \quad (3)$$

同理,合作者也是实体的第 IV 类属性,可利用以上方法计算其相关度。按照属性的模式选择对应的相似度计算函数,可以简化问题模型,使系统更具通用性。

相同类型的属性对实体识别的贡献由属性值的值域范围决定。

定义 2 相遇度　属性的相遇度是指实体拥有该属性一个确切值的概率。一般情况下,相遇度与属性值集合的大小成反比,即属性值的范围越大,该属性值相对于实体的相遇度越低。

相遇度可以粗略衡量属性对实体识别的贡献。以论文作者为实体识别的对象而言,作者包含的属性包括作者参加的会议和论文的合作者。假设作者数量为 n,会议数量为 m,作者与会议的相遇度为 $1/(n\times m)$,同理,作者与合作者的相遇度为 $1/(n\times n)$,一般 n≫m。两个作者拥有共同的合作者的情形比两作者参加过同一会议的情形更可能代表其是同一作者。所以合作者的相遇度比会议更小,更适用于识别问题。

属性的重要性可由属性的模式和相遇度共同决定。不同模式的属性在实体识别问题中的优先级排序为: I > II > III > IV,同一种模式的属性可通过相遇度大小进行比较。

要生成相似度计算策略,需要知道属性的类型以及属性的模式,属性的模式可以通过训练数据发现。属性的类型需要通过输入指定,根据属性的类型可以确定基本的比较函数。如基本的数据类型有字符串型(String),数值型(number),日期(Date),文本(Text),姓名(Name),邮件(Email),地址(Address)等。基本的数据比较已经有很多成熟的函数,因而可以直接调用。

3.2　判决器模块的设计

判决器是框架中的一个重要模块,完成对实体的相似度分布情况的统计操作。可将记录的每一属性当作实体的一个维度,根据属性模式选取一些重要属性进行相似度计算。$f_k(r_i,r_j)$ 表示计算记录 i 和记录 j 的属性 k 的相似度,则实体对的相似度可表示为:

$$F(r_i,r_j)=[f_1(r_i,r_j),f_2(r_i,r_j)\cdots\cdots f_k(r_i,r_j)] \quad (4)$$

判断器是相似空间的一个划分,每个子区域称为判决单元,记为 cell。一种均匀划分的方法如下: 对 k 维相似度空间的每一维均划分为 m 个子区域,则第 i 维的第 n 个区域可表示成 $\left[\frac{n_i-1}{m},\frac{n_i}{m}\right]$,$n_i\in[0,m-1]$ 且 $n_i\in N$,故可将相似度空间划分为 m^k 个 cell。

每个判决单元分别包含正计数器 P 和负计数器 N。在学习阶段,每个实体对的相似向量 F 落在相应 cell 中,若该实体对为同一实体,则该 cell 的正计数器 P 加 1,否则负计数器 N 加 1。P/P + N 即为该区域判断为同一实体的概率,支持度越接近于 1 说明实体对越可能是同一实体。所以判断器是一个 k 维向量到 [0,1] 区间的数学映射。

3.3　实体划分算法

实体划分在实体关系图上进行。实体关系图的顶点表示记录,边表示实体对间的相似度,通过边的操作进行实体划分。

关系图的顶点分为两类,一类是原始顶点,其中只包含一条记录; 另一类是划分过程中新形成的点,称为超点,超点带有表示实体的标签,且包含此实体的记录的集合。边 e 代表的是实体对间存在相似,边的权值为相似向量。原始关系图中仅含原始顶点,当所有实体对的相似向量计算完毕,并建立起原始关系图后,就可开始进行实体划分了。

实体划分算法主要过程为: 从未标记边中选择相似度最大的边,查询判决器,若大于判断阈值,则判为同一实体,合并相关顶点,即 CLUSTER 操作,有关边的相似度则需要进行重新计算; 否则即对边做暂时标记。继续在剩下未标记边中寻找相似度值最大的边,重复此过程。当没有未标记边剩余时,再对标记边进行拆分操作 SPLIT,直到无边剩余。

CLUSTER 操作主要是对顶点进行合并或创建。当边的对象(e. O)与端点标签相同时进行合并,否则就需要新建顶点。具体操作如表 2 所示。其中,边所连接的记录为 x 和 y,记录所在的顶点分别为 u,v。顶点调整过程中,特别当顶点包含的记录增多后,顶点的属性集合增大,此属性的相关度也可能增大,此时需要重新计算有关边的相似度。

表 2　合并操作

Tab. 2 CLUSTER operators

边的对象与顶点标签关系	顶点调整情况	
e. O = e. u. tag = e. v. tag	Merge(U,V)	U = U + V
e. O = e. u. tag ≠ e. v. tag	ADD(U,y)	U = U + y
e. O = e. v. tag ≠ e. u. tag	ADD(V,x)	V = V + x
e. O ≠ e. u. tag , e. O ≠ e. v. tag	New(N)	N = x + y

SPLIT 操作主要是对边进行拆分,仅当边的对象与两端点的标签不一致时,需要创建代表新实体的超点,其他情况则无需修改端点的内容。

如有数据表,其属性为 A = [conference | year | title | author],如表 3 所示。一条记录中包含了多个作者。

表 3　论文的信息

Tab. 3 Records on citations

ID	conf.	year	Title	author		
R1	VLDB	2011	T1	Bob	Peter	
R2	ICDE	2011	T2	Peter	Jean	Smith
R3	VLDB	2012	T3	Bob	Jean	Jim
R4	ICDE	2012	T4	Bob	Jim	Joe

若以作者为实体识别的对象,则根据模式首先按作者姓名的相似度形成待判断的实体对,以无向图的形式表示实体间的关系。图 3 展示了对表 3 中作者的划分过程。图 3(a)表示的是原始实体关系图。首先选择相似度最大的 R3 和

R4 记录对中的 Bob，发现其匹配概率大于判断阈值，则进行合并；同时 Jim 也被判定为同一实体，分别创建了 2 个新点，一个代表 Bob，一个代表 Jim，如图 3(b) 所示。重新计算有关边的相似度，但此时 R1 和 R3、R4 间关于 Bob 的相似度还是没有增加。计算 R1 和 R2 所连接的边 Peter，发现其匹配概率大于阈值，则 R1 和 R2 进行合并，如图 3(c) 所示。这时重新计算 Bob 边的相似度，此时有共同的合作者 Jean，相似度增大，大于判断阈值，因此进行合并。Bob 包含了记录 R1、R2 和 R3，最后 Jean 也被认为是同一实体，如图 3(d) 所示。

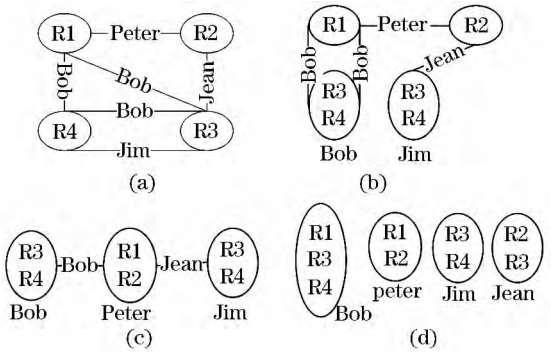

图 3 MaxF 算法实体划分过程

Fig. 3 Process of entity partitions

4 实验结果与分析

为了验证框架的有效性，编程实现了有关算法，并对结果进行了评价。操作系统为 Windows XP，2GB 内存，2.45GHz 主频，开发工具为 Code: : Blocks 8.02，编程语言为 C + +。

实验中，采用的数据是 DBLP 中的论文信息。记录包含作者 A，论文名 T，会议 C，发表时间 Y 等，实体识别的对象是论文作者，由于作者名包含多个作者，是一个多值字段，则分为了作者和合作者两部分。总共 1 084 条记录，其中，860 条用作训练数据，剩余 224 条用作测试数据，数据共计 100KB。

实验结果则采用关于实体对的准确率(Precious) 、召回率(Recall) 和调和平均 F - score 进行评价。令实际为同一实体的实体对集合为 S，实验判断为同一实体的实体对的集合为 R，则准确率 $P = \frac{|S \cap R|}{|R|}$，召回率 $R = \frac{|S \cap R|}{|R|}$，调和平均 $F = \frac{2PR}{|P + R|}$。

为了比较各个属性对实体识别的作用效果，采用了限定属性的方法。判断阈值 $\alpha = 0.9$，实验结果如图 4 所示，横坐标字母代表的是计算相似度时使用的属性。其中，仅计算名字和会议的相似度时，其召回率只有 25.6%，准确率为 87.2%。仅计算名字和论文标题时，召回率为 43.7%，准确率为 89.3%。而名字和合作者组合时的召回率为 38.4%，准确率为 94.7%。单个属性对实体识别的贡献来看，合作者的准确率较高，标题的召回率较高。多个属性组合可以明显提高结果召唤率。名字、会议、标题、合作者的组合时准确率为 97.1%，召回率为 91.7%，调和平均为 94.7%，均为最好结果。实验结果说明，文中设计的框架可以有效综合各种属性的信息。

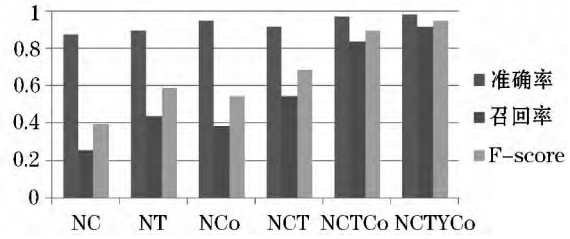

图 4 各属性对实体识别的作用

Fig. 4 Effects of each attibute

特别地，选取上述第 6 组实验中部分具体实例的实验结果，如表 4 所示。当重名的实体数较少时，实验结果更好。原因之一是这类作者的合作者较少，而合作者中一般都是相同的作者，相同实体容易得到识别；又由于发表文章较少，文章标题之间的关键词和会议重合几率都比较低，不同实体间容易区分。

表 4 部分实验实例结果

Tab. 4 Part of samples

Name	Entities	Citations	P	R	F
Guohui Li	4	25	0.964	0.841	0.898
Guo - Hui Li	2	6	1.0	1.0	1.0
John Smith	5	14	0.943	0.911	0.906
Guohua Li	3	4	1.0	1.0	1.0
Scott Waterman	2	4	1.0	1.0	1.0

5 结束语

本文提出了一种基于模式的实体识别方法，针对模式特点的相似度计算方法更具有通用性。以向量表示属性的相似度，通过监督学习形成判决器。实体划分阶段每次选择最相似的实体对，通过查询判断单元进行判断，更新相关实体对的相似向量，并迭代进行实体划分。实验结果表明能自动有效地进行实体划分。现存的问题包括平均划分相似空间的方法不够精细，用户要求的准确率较高时，召回率较低。下一步的研究重点包括判断器的划分方式以及当用户输入较高判断阈值情况下如何提高系统的召回率。

参考文献:

[1] MENESTRINA D, WHANG S E, GARCIA - MOLINA H. Evaluation of entity resolution approaches on real - world match problems [C] //VLDB, 2010: 208 - 219.

[2] WANG Jiannan, LI Guoliang, YU Xu , et al. Entity matching: how similar is similar [C] //VLDB, 2011: 622 - 633.

[3] CHEN Z, KALASHNIKOV D V, MEHROTRA S. Exploiting context analysis for combining multiple entity resolution systems [C] //SIGMOD, 2009: 207 - 218.

[4] TEJADA S, KNOBLOCK C A, MINTON S. Learning domain - independent string transformation weights for high accuracy object identification [C] //Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02) , 2002.

[5] BHATTACHARYAI, GETOOR L. Collective entity resolution in relational Data [C] //TKDD, 2007.