面向关联数据的联合式实体识别方法

孙琛琛 申德荣 寇 月 聂铁铮 于 戈
(东北大学信息科学与工程学院 沈阳 110819)

摘 要 文中提出一种基于图的、迭代的联合式实体识别方法.初始时,将多类型的、关联的实体数据对象集合构建实体数据对象关系图,将基于语义路径的相似度和属性相似度结合起来判断数据对象是否匹配;然后,合并匹配成功的数据对象,并对对象图中的相应数据对象结点及其周边执行局部图收缩,这两个操作使对象图的局部语义变得更丰富,促使该局部范围内产生出新的候选匹配对象对,以待后续识别,实现相似度传递,形成一个迭代的识别过程.随着不断迭代,对象图的语义不断丰富,提高了联合式实体识别的准确性.通过实验证明文中提出的方法比已有的联合式实体识别方法和基于对象关系的单类型实体识别方法具有更高的准确性.

关键词 联合式实体识别;相似度传递;基于结构的相似度;实体数据对象关系图
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2015.01739

A Related Data Oriented Joint Entity Resolution Approach

SUN Chen-Chen SHEN De-Rong KOU Yue NIE Tie-Zheng YU Ge
(*College of Information Science and Engineering, Northeastern University, Shenyang* 110819)

Abstract We propose a graph-based iterative joint entity resolution approach. To start off, an entity data object relationship graph is built from the input dataset consisting of multiple classes of related data objects. It hires a hybrid similarity, combining a structure similarity based on semantic paths and an attribute-based similarity, to decide whether two data objects match. Then it merges the matched pair and contracts the neighborhood of the merged pair, which leads to enrichment of semantics of the neighborhood. Enrichment of semantics may help generate some new candidate data object pairs in the neighborhood, which will be resolved later. Generation of new candidate data object pairs is called similarity propagation, making it an iterative process. With the iterative process going on, semantics of the object graph becomes richer and richer, promoting accuracy of entity resolution. The experimental evaluation proves that the proposed approach outperforms existing joint entity resolution approaches and relationship-based single class entity resolution approaches in accuracy.

Keywords joint entity resolution; similarity propagation; structure-based similarity; entity data object relationship graph

1 引 言

实体识别(Entity Resolution,ER)是数据清洗的一个关键方面,对于数据挖掘和数据集成都至关重要[1-4].数据集成和数据挖掘都可能涉及多数据源,不同的数据源有不同的描述实体的方法.由于拼写错误、缩写方式不同、描述格式不同、属性值缺失、

收稿日期:2014-09-30;最终修改稿收到日期:2015-04-07.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316201)、国家自然科学基金面上项目(61472070)资助.**孙琛琛**,男,1987年生,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为实体识别.E-mail:dustinchenchen.sun@gmail.com.**申德荣**,女,1964年生,博士,教授,博士生导师,主要研究领域为分布式数据管理、数据集成.**寇 月**,女,1980年生,博士,副教授,主要研究方向为实体搜索、数据挖掘.**聂铁铮**,男,1980年生,博士,副教授,主要研究方向为数据质量、数据集成.**于 戈**,男,1962年生,博士,教授,博士生导师,主要研究领域为数据库、大数据管理.

实体的某些属性值随着时间推移发生演化(比如年龄、居住地点、工作单位)等,描述同一实体的不同实体数据对象存在差异.实体识别是将一个或多个数据源中描述同一现实世界实体的实体数据对象(简称"数据对象"或"对象")分到同一组的过程.为了实现高质量的数据集成和数据挖掘,需要对数据进行实体识别.传统的实体识别方法基于属性相似度,解决单类型的实体识别问题,待匹配的数据对象之间是独立的[1,3-4].

大数据时代,数据呈现多样性和关联性,在实体识别中体现为识别目标包含多类型的数据对象,数据对象之间存在语义关系,比如引文数据集(涉及文章、作者、会议等)、电影数据集(涉及电影、演员、导演、电影公司等)、消费信息数据集(涉及消费者、商品、厂家、购买记录等)等,称之为关联的数据.使用传统的实体识别方法处理关联的数据,其准确性将会受到限制,因为传统方法无法发掘对象关系,具有明显的局限性.以图 1(b)为例,采用传统的方法:①无法匹配 a_1 和 a_4(属性值不完整);②也无法决定 a_{12} 与 a_2 还是 a_{13} 更匹配(同名不同实体);③需要分别对文章、作者和会议进行实体识别,无法利用对象关系来提高实体识别的准确性和优化实体识别的顺序,比如 c_1 和 c_2 匹配后,a_1 和 a_4,a_2 和 a_5,a_3 和 a_6 的匹配的可能性增加,接下来应当优先对这些数据对象对进行匹配.为了解决这些问题,需要发掘对象关系.目前,已经有一些研究者提出利用关联的数据中的对象关系进行实体识别,其中有一些方法涉及多类实体的联合识别,称之为联合式实体识别.已有的工作都存在一定的局限性,Bhattacharya 等人[5-7]利用引文集中作者的共现关系来识别作者,可以解决第①类问题,解决不了其他两类问题;Dong 等人[8]根据对象关系构建待匹配对象对的依赖图来进行联合式识别,它们可以解决第①类问题,部分地解决第③类问题,但解决不了第②类问题.这两个方法利用对象间直接关联关系,没有充分发掘对象关系,称为基于上下文的联合式实体识别方法.Kalashnikov 等人[9-11]默认只有一类实体未被识别而其他类型实体已被识别,从而构建实体关系图,利用实体间关系可以部分地解决第②类问题,但解决不了第①和③类问题,作为单类型的实体识别方法,无法通过一些数据对象的匹配结果来促进与其关联的其他数据对象的匹配.

paper {title, {authors}, venue},
author {name},
venue {name}

(a) 下面(b)中关联数据集的关系模式

c_1{{"Distributed db"}, {a_1, a_2, a_3}, ve_1},
c_2{{"Distributed database"}, {a_4, a_5, a_6}, ve_2},
c_3{{"Cloud computing"}, {a_7, a_8}, ve_3},
c_4{{"Cloud storage"}, {a_9, a_{10}}, ve_4},
c_5{{"Distributed computing"}, {a_{11}, a_{12}}, ve_5},
c_6{{"Computer Architecture"}, {a_{13}, a_{14}}, ve_6}
a_1("Blair"), a_2{"Bill Lee"}, a_3{"Mike Leo Bush"},
a_4{"J. B. "}, a_5{"B. Lee"}, a_6{"M. Lee Bush"},
a_7{"J. Blair"}, a_8{"M. L. Bush"}, a_9{"Mike Bush"},
a_{10}{"John Black"}, a_{11}{"John Black"}, a_{12}{"Bill L. "},
a_{13}{"Bill Logan"}, a_{14}{"Will"},
ve_1{"VLDB"}, ve_2{"Very large DB"}, ve_3{"CIKM"},
ve_4{"EDBT"}, ve_5{"ICDE"}, ve_6{"ISCA"}

(b) 关联数据集示例

{⟨c_1, c_2⟩, ⟨c_3⟩, ⟨c_4⟩, ⟨c_5⟩, ⟨c_6⟩, ⟨a_1, a_4, a_7⟩, ⟨a_2, a_5, a_{12}⟩,
⟨a_3, a_6, a_8, a_9⟩, ⟨a_{10}, a_{11}⟩, ⟨a_{13}⟩, ⟨a_{14}⟩, ⟨ve_1, ve_2⟩, ⟨ve_3⟩,
⟨ve_4⟩, ⟨ve_5⟩, ⟨ve_6⟩}

(c) 正确的联合式实体识别结果

图 1 联合式实体识别示例

多类型的、关联的数据集是关系型数据,可以映射为实体数据对象关系图[2,12],其中数据对象对应结点,数据对象间语义关系对应结点之间的语义链接,映射后完整保留了数据对象的关系模式(模式信息),而且拥有了图的拓扑结构,可以更方便地、深入地发掘数据对象间的语义关系.在对象图中,数据对象之间的相似度不仅可以通过属性衡量,还可以通过图结构来衡量,从而提高数据对象匹配的准确性,有利于解决第②类问题.数据对象间的路径和链接的模式信息结合起来可以更准确地计算数据对象间基于结构的相似度.对不同类型的数据对象进行联合匹配,可以提高实体识别的准确性,当某些数据对象匹配成功后会引起局部图结构的变化,匹配结点的融合,相关链接的收缩,使得局部语义更丰富,促进局部范围内其他数据对象的匹配.比如当识别出图 1(b)中的 c_1 和 c_2 后,a_1 和 a_4,a_2 和 a_5,a_3 和 a_6 的匹配的可能性增加,后续应该对这些数据对象对进行匹配.这有助于解决第①、③类问题,因为赋予了对象图动态性,随着联合式实体识别的进行,对象图不断演化,结点信息更全面,结构逐渐变得更紧密,语义愈渐丰富,为后续的实体识别提供更多的依据;同时也赋予了联合式实体识别迭代性,同一对数据对象随着局部结构的变化,可能被多次匹配,直到匹配成功或局部结构不再变化.可见,利用对象图进行联合式实体识别可以弥补已有方法的缺点.

本文的主要贡献如下:

(1)以集合划分理论为基础，提出基本的实体识别模型，并在此基础上提出联合式实体识别模型.证明了有限的对象集合的实体识别和联合式实体识别的解是唯一存在且有限步可达的.

(2)提出一种自底向上的、基于图的、迭代的联合式实体识别方法(Graph-Based iterative Joint Entity Resolution approach，GBi-JER)，核心思想是充分发掘逐渐收敛的对象图，并通过迭代地利用动态变化的对象关系(包括直接和间接关系)，提高实体识别的准确性.GBi-JER 包括联合式匹配、联合式合并和相似度传递 3 个主要模块.

(3)提出一种基于语义路径的相似度算法.根据关系模式，赋予不同的链接类型不同的单向关联权值，然后将单向关联权值与双向的随机游走模型结合起来计算结构相似度.

(4)通过实验验证了 GBi-JER 的有效性和各模块的作用.

本文第 2 节首先通过示例描述问题，然后定义基本的实体识别模型，并在此基础上提出联合式实体识别模型；第 3 节总述基于图的、迭代的联合式实体识别方法 GBi-JER，并介绍了实体数据对象关系图和初始化工作；第 4 节阐述联合式匹配，并提出一种基于语义路径的相似度算法；第 5 节介绍联合式合并与相似度传递，联合式合并包括数据对象合并和局部图收缩，局部图收缩会引起相似度传递；第 6 节介绍相关工作；第 7 节是实验与分析，通过与相关工作对比及自身对比，评价本文提出的 GBi-JER；最后第 8 节总结全文.

2　联合式实体识别模型

本节阐述联合式实体识别模型，2.1 节通过示例描述问题；2.2 节介绍基本的实体识别模型，2.3 节在基本的实体识别模型基础上提出联合式实体识别模型.

2.1　问题描述

实体类型是指实体的关系模式类型，比如作者是一种实体类型而文章是另一种实体类型.一个实体数据对象(简称“数据对象”或“对象”)描述真实世界的一个实体，可能存在多个对象对应同一实体.对象类型与实体类型是一致的.对象包括多个属性，分为简单属性和引用属性两类：简单属性的值是 String、Int、Real 等简单类型；引用属性的值是对另外一个对象的引用.图 1(a)中，title 是 paper 的简单属性；paper 还包含指向 author 和 venue 的引用，是引用属性.

实体识别是将给定对象集合中描述同一实体的对象划分到同一个组的过程.当前只考虑图 1(b)中所有作者对象集合 $O_a=\{a_1,a_2,a_3,a_4,a_5,a_6,a_7,a_8,a_9,a_{10},a_{11},a_{12},a_{13},a_{14}\}$，$O_a$中的所有对象之间独立，$O_a$的实体识别就是对 O_a进行集合划分的问题，经过实体识别算法处理后的结果为 $P_{ER}(O_a)=\{\langle a_1,a_4,a_7\rangle,\langle a_2,a_5,a_{12}\rangle,\langle a_3,a_6,a_8,a_9\rangle,\langle a_{10},a_{11}\rangle,\langle a_{13}\rangle\}$，其中一对尖括号内的对象子集表示划分后的组，每个组内的对象描述同一作者，每个作者只被唯一的对象组描述.在上述的集合划分过程中，只利用对象的属性信息，如作者名.

联合式实体识别不同于传统的实体识别，要对多类型的、关联的数据对象进行联合识别，识别过程中要考虑不同对象间关系，接下来举例说明.图 1(a)描述了(b)中关联数据集的关系模式，包括文章(paper)、作者(author)和会议(venue) 3 类对象及对象关系：作者与文章之间的“写作”、“被写作”关系，文章与会议之间的“发表在”、“发表”关系.文章、作者和会议都有各自的属性，为了简化表达，图 1 中作者和会议只给出一个属性，现实世界中对象的属性通常是更复杂的.图 1(b)的关联数据集包括 6 个文章对象，14 个作者对象和 6 个会议对象.图 1(c)是(b)中对象的正确的实体识别结果.在已知图 1 中(a)、(b)的情况下，如何求得图 1(c)是一个联合式实体识别的问题.联合式实体识别利用对象关系提高实体识别的准确性和优化实体识别的顺序.比如 c_1和 c_2匹配后，a_1和 a_4，a_2和 a_5，a_3和 a_6的匹配的可能性增加，后续应该优先识别这些对象对.

2.2　基本的实体识别模型

基本的实体识别模型由定义 1～4、定理 1～2 和一对匹配函数和合并函数组成.基本的实体识别模型的共同条件：给定一个包含 n 个单类型对象的集合 $O_s=\{o_i\}$，对象之间相互独立.

定义 1.　合法划分.如果 O_s的一个划分 $P(O_s)$满足条件：$P(O_s)$中每个对象组只描述同一实体，则 $P(O_s)$是 O_s的一个合法划分.

对象集合 $O_s=\{o_i\}$可以看作是一个划分 $O_s=\{\langle o_i\rangle\}$.显然，$O_s=\{\langle o_i\rangle\}$本身也是一个合法划分.

定义 2.　划分优化.$P_1(O_s)$，$P_2(O_s)$是 O_s的两个合法划分，如果对于 $\forall p_i\in P_1(O_s)$，$\exists p_j\in P_2(O_s)$使得 $p_j\subseteq p_i$，且这些包含关系中至少有一个是真包含，那么划分 $P_1(O_s)$比划分 $P_2(O_s)$更优.

定义 3. ER 划分. $P_{in}(O_s)$是 O_s的一个合法划分,对 $P_{in}(O_s)$进行一次划分后得到 O_s的另一个合法划分 $P_{out}(O_s)$,且 $P_{out}(O_s)$比 $P_{in}(O_s)$更优,则这次划分是 O_s的一次 ER 划分.

一个对象集合的 ER 划分中,只包含一个对象的组称为单例组,否则为非单例组. 对象分为简单对象和组合对象:初始时每一个对象都是简单对象;如果两对象匹配成功,将它们合并,组成一个组合对象. 简单对象对应 ER 划分中的单例组;组合对象对应 ER 划分中的非单例组.

定义 4. 实体识别. 对O_s进行迭代的 ER 划分直到得到最终识别结果 $P_{ER}(O_s)$,满足条件:① $P_{ER}(O_s)$中每个组只描述同一实体,② $P_{ER}(O_s)$中不存在两个组描述同一实体. 此时,称为 O_s被识别完成.

基本的实体识别模型通过一对匹配函数和合并函数来实现实体识别. 每次只比较 O_s中两个对象 o_i和 o_j. 匹配函数 m 是定义在 $O_s \times O_s$上的布尔函数,它决定给定的两个对象是否匹配或者说是否描述同一实体. 如果 m 判定两个对象 o_i和 o_j匹配,则 $m(o_i,o_j)=\text{true}$ 记作 $o_i \approx o_j$;否则,$m(o_i,o_j)=\text{false}$ 记作 $o_i \neq o_j$. 匹配函数独立地判定两个对象是否匹配,即两个对象是否匹配只决定于自身的属性相似度. 合并函数 u 将两个匹配的对象合并为一个组合对象. 如果 $o_i \approx o_j$,$u(o_i,o_j)$将合并 o_i和 o_j,合并结果记作$\langle o_i,o_j \rangle$;否则,$u(o_i,o_j)$是非法的,$\langle o_i,o_j \rangle$是无意义的. 对象合并会产生信息增益,促进新的匹配对出现,比如 o_1,o_2都与 o_3不匹配,但$\langle o_1,o_2 \rangle \approx o_3$. 合并后的组合对象要重新与其他对象进行匹配,因此,使用一对匹配函数和合并函数可以构建出一个迭代的实体识别算法.

基本的实体识别模型满足定理 1 和定理 2.

定理 1. 有限的单类型对象集合的实体识别的解唯一存在.

证明. (1)存在性. 对 R_s进行迭代的 ER 划分,每次划分结果都符合定义 4 中条件①,直到连续的两次 ER 划分的结果相同,说明此时的划分结果已经符合条件②;(2)唯一性. 假定存在 O_s的两个不同划分结果都是 O_s实体识别的解,记作 $P_{ER1}(O_s)$和 $P_{ER2}(O_s)$. 根据 ER 划分的定义可知,$P_{ER1}(O_s)$和 $P_{ER2}(O_s)$都一定符合条件①,而 $P_{ER1}(O_s)$和 $P_{ER2}(O_s)$是不同的划分,那么其中至少有一个不符合条件②,产生矛盾. 因此,假设不成立,最多有一个解. 结合存在性可知,O_s的实体识别的解是唯一的. 证毕.

定理 2. 有限的单类型对象集合的实体识别的解是有限步可达的.

证明. O_s的实体识别的解的可达性定理 1 已经证明. 现证明其解是有限步的. 分析在最坏的情况下,O_s的对象都描述同一个实体,每次 ER 划分都只识别出一对对象,$2 \leq i \leq n$,那么,需要 $size=\sum i \times (i-1)/2$ 次比较来识别出所有的对象. 其他情况下,比较次数都要比 $size$ 少. 因此,O_s的实体识别的解是有限步可达的. 证毕.

2.3 联合式实体识别模型

为了定义联合式实体识别模型,需要将基本的实体识别模型中的相关定义推广到联合的情况. 联合式实体识别模型由定义 5~9,定理 3~4 和一对联合式匹配函数和联合式合并函数组成. 联合式实体识别模型的共同条件:给定一个多类型的、关联的对象集合 $O_{mr}=\{(O_t)\}_{t=1}^{T}$,$O_t$是第 t 类对象的子集,$SR_{mr} \subseteq O_{mr} \times O_{mr}$是不同对象间的语义关系的集合.

定义 5. 联合式划分. O_{mr}的联合式划分为 $JR(O_{mr})=\bigcup_{t=1}^{T} P(O_t)$,$\exists i,j \in [1,T]$,$P(O_i)$受$P(O_j)$影响,即划分过程受 SR_{mr}中语义关系的影响和约束.

定义 6. 合法联合式划分. 如果 O_{mr}的一个联合式划分$JP(O_{mr})$中任意的 $P(O_t)$都是 O_t的一个合法的划分,那么 $JP(O)$是 O_{mr}的一个合法联合式划分.

定义 7. 联合式划分优化. $JP_1(O_{mr})$,$JP_2(O_{mr})$是 O_{mr}的两个合法联合式划分,如果 $JP_1(O_{mr})$中存在 $P_1(O_{t_1})$都比 $JP_2(O_{mr})$中 $P_2(O_{t_1})$更优,且 $JP_2(O_{mr})$中不存在$P_2(O_{t_2})$比$JP_1(O_{mr})$中的$P_1(O_{t_2})$更优,那么联合式划分 $JP_1(O_{mr})$比联合式划分 $JP_2(O_{mr})$更优.

定义 8. 联合式 ER 划分. $JP_{in}(O_{mr})$是 O_{mr}的一个合法联合式划分,对 $JP_{in}(O_{mr})$进行一次联合式划分后得到O_{mr}的另一个合法联合式划分$JP_{out}(O_{mr})$,且 $JP_{out}(O_{mr})$比 $JP_{in}(O_{mr})$更优,这次联合式划分是 O_{mr}的一次联合式 ER 划分.

定义 9. 联合式实体识别(Joint ER,JER). 对 O_{mr}进行迭代的联合式 ER 划分直到 O_{mr}中所有 O_t都被识别完成. 此时,称为 O_{mr}被识别完成.

联合式实体识别模型通过一对联合式匹配函数和联合式合并函数实现联合式实体识别. 联合式匹配函数 jm 是定义在$(O_{mr} \times O_{mr} \times SR_{mr})$上的布尔函数,非独立地判定两个对象是否匹配,即两条对象是

否匹配不仅决定于自身的属性相似度，还取决于它们与其他对象的语义关系，两个对象的匹配情况要受其他对象匹配结果的影响和约束. 联合式合并函数 ju 不仅要将两个匹配的对象合并为一个组合对象，还要处理相应的对象语义关系，新的组合对象继承原来对象的语义关系，丰富了与该对象关联的对象的语义关系，促进新的匹配对出现. 比如图 1(b)中 c_1 和 c_2 匹配后，a_1 和 a_4，a_2 和 a_5，a_3 和 a_6 各自的语义关系更紧密，匹配的可能性增加. 语义关系继承产生不同对象间的相似度传递，提高了联合式实体识别的准确性，这是联合式实体识别的最大优势. 容易发现，对象合并和语义关系继承使得联合式实体识别模型是迭代的.

联合式实体识别模型满足定理 3 和定理 4.

定理 3.　有限的多类型、关联对象集合的联合式实体识别结果唯一存在.

定理 4.　有限的多类型、关联对象集合的联合式实体识别的解是有限步可达的.

定理 3、定理 4 很容易通过定理 1、定理 2 的结论推导出，因此，不再赘述.

3　基于图的迭代的联合式实体识别方法——GBi-JER

本节将介绍基于图的迭代的联合式实体识方法(Graph-Based iterative Joint Entity Resolution，GBi-JER)，它借助实体数据对象关系图实现了联合式实体识别模型. 3.1 节介绍实体数据对象关系图；3.2 节阐述 GBi-JER 方法的整体流程；3.3 节介绍 GBi-JER 的初始化工作.

3.1　实体数据对象关系图

为了充分发掘多类型对象间语义关系来进行联合式实体识别，将关系型数据集构建实体数据对象关系图. 将关系型数据映射到实体数据对象关系图，可以充分地保留关系型数据的关系模式[2,12]，这一特点在联合式实体识别过程发挥重要的作用. 后文提到的模式信息即关系型数据的关系模式，关系模式中的对象关系也称为对象语义关系.

定义 10.　实体数据对象关系图. 一个实体数据对象关系图是一个有向图 $G=(O,L)$，其中 O 是对象集合，$L\subseteq O\times O$ 是不同的对象间语义链接的集合；存在一个对象类型映射函数 φ：$O\to N$ 和语义链接类型映射函数 ψ：$L\to\Gamma$，对于每个对象 $o\in O$ 有 $\varphi(o)\in N$，对于每条语义链接 $l\in L$ 有 $\psi(l)\in\Gamma$，N 为对象类型的集合，Γ 为链接类型的集合.

实体数据对象关系图简称数据对象关系图或对象图. 对象图中的对象也称为结点或点，语义链接简称链接. 对象图的两点间存在双向关系，后文中，对象图中将用无向链接表示双向链接. 如果两对象类型 μ_1，μ_2 之间存在有向的语义关系 γ_0，那么记作 $\mu_1\gamma_0\mu_2$；γ_0 的逆向关系为 γ_0^{-1}，记作 $\mu_2\gamma_0^{-1}\mu_1$. 引文网络是一个对象图，以图 1 为例，包括文章、作者和会议 3 类对象，存在的关系类型有：文章与会议之间“发表在”和“发表”，文章与作者之间“被写作”和“写作”. 在初始的对象图中所有的对象都是简单对象；在联合式实体识别过程中会生成组合对象.

定义 11.　语义路径. 对象图中的语义路径(semantic path，sp)由对象和语义链接有序组成的，不允许经过重复的对象和链接. 对象 o_0 有序地经过 $l_0,o_1,l_1,\cdots,o_{w-1},l_{w-1}$ 到达 o_w 的语义路径记作 $sp(o_0,o_w)=(o_0,l_0,o_1,l_1,\cdots,o_{w-1},l_{w-1},o_w)$，简写为 $sp(o_0,o_w)=(o_0,o_1,\cdots,o_{w-1},o_w)$.

定义 12.　模式图. 给定一个对象图 $G=(O,L)$，有模式图 $G_{sch}=(N,\Gamma)$，G_{sch} 是一个有向图，N 是对象类型的集合，$\Gamma\subseteq N\times N$ 是对象类型间的语义边的集合，语义边表示对象类型间的语义关系.

图 2 为图 1(a)对应的模式图.

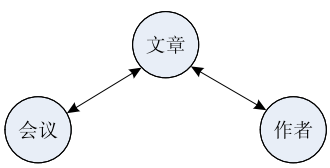

图 2　模式图示例

定义 13.　模式路径. 给定模式图 $G_{sch}=(N,\Gamma)$，$\mu_i\in N,i\in[0,w]$，$\gamma_j\in\Gamma,j\in[0,w-1]$，对象类型 μ_0 和 μ_w 间的一条模式路径表示一种复合的语义关系，记作 $schp(\mu_0,\mu_w)=(\mu_0,\gamma_0,\mu_1,\cdots,\mu_{w-1},\gamma_{w-1},\mu_w)$ 当每两个对象类型间只有一种语义关系时，可简写为 $schp(\mu_0,\mu_w)=(\mu_0,\mu_1,\cdots,\mu_w)$.

每一条语义路径都存在对应的模式路径，给定一条语义路径 $sp(o_0,o_w)=(o_0,o_1,\cdots,o_{w-1},o_w)$，其模式路径记作 $schp(sp(o_0,o_w))=(\varphi(o_0),\varphi(o_1),\cdots,\varphi(o_{w-1}),\varphi(o_w))$.

3.2　GBi-JER 概述

GBi-JER 迭代地发掘逐渐收敛的对象图，充分利用对象关系(包括直接和间接关系)，进行联合式实体识别. GBi-JER 具体流程及构成如图 3 所示，上半部分由有向实线连接、表示算法流程，下半部分由

有向虚线连接、表示源数据集和对象图在每个功能模块中的参与情况，ABS(Attribute-Based Similarity)表示基于属性的相似度，SBS(Structure-Based Similarity)表示基于结构的相似度，GBi-JER 包括3个主要功能模块：联合式匹配模块、联合式合并模块和相似度传递模块以及一个初始化模块，涉及两个数据结构：一个对象图和一个候选队列. 整个方法的输入是一个脏的关联数据集，输出是一个干净的关联数据集.

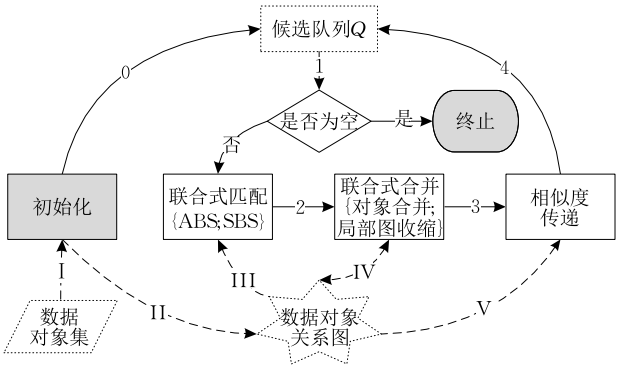

图 3 GBi-JER 工作流程

首先，初始化工作，包括对象图初始化和候选队列对初始化. 将关系型的数据集(由图3中平行四边形表示)根据对象图定义和数据集的关系模式构建原始的对象图，图3中虚线II指向的七边形表示对象图. 使用 Canopy[13] 技术对对象分块，块内生成初始的候选匹配对象对，将它们依次插入到候选队列 Q 中，如图3中实线0所示.

接着，进入迭代的联合式实体识别，包括3个循环的步骤：(1)如图3中实线1所示，先判断此时 Q 是否为空，如果 Q 已经为空，则终止；否则，从 Q 队首取一对候选匹配对，将它输入到联合式匹配模块中，如果匹配成功，则进入下一模块；否则，重复上述操作. 虚线Ⅲ由对象图指向联合式匹配模块，表示在判断两个对象是否匹配要用到对象图. 联合式匹配模块综合属性相似度和结构相似度衡量相似度，因此更加准确. 本文提出一种基于语义路径的相似度算法来衡量对象的结构相似度，将在第4节介绍. 图1(b)中，a_{12} {"Bill L."}与 a_2 {"Bill Lee"}，a_{13} {"Bill Logan"}哪个更相似，仅利用属性相似度无法解答，而基于语义路径的相似度算法可以解决这个问题，后文将给出答案；(2)图3中实线2表示匹配的对象被输入到联合式合并模块，带双向箭头的虚线Ⅳ表示双向的数据流动. 匹配的对象将被合并，同时对该对象对在对象图中对应的两结点及其周边执行局部图收缩：将两结点合并为新的结点，将与两结点相关的链接进行收缩处理. 这两个操作丰富了对象图的局部语义：对象合并使得对象的信息增益；局部图收缩使得局部图结构更加紧密，对象间语义关系更加丰富；(3)图3中实线3表示下一步将进行相似度传递，虚线Ⅴ从对象图指向相似度传递模块表示相似度传递将用到对象图. 对象合并和局部图收缩改变了匹配对周边的图结构，使与它们关联的对象的属性相似度、尤其是结构相似度增加，引起周边对象的重新计算，生成新的候选匹配对. 新的候选匹配对被插入到 Q 中合适位置，如图3中实线4所示. 以图1(b)为例，当 c_1 和 c_2 匹配成功后，a_1 和 a_4，a_2 和 a_5，a_3 和 a_6 的匹配的可能性增加，因此将它们插入到 Q 中合适位置，以待后续匹配. 图3中实线1、"否"、2、3、4形成一个闭合的有向环，GBi-JER 的3个步骤形成了一个迭代的联合式实体识别过程，终止条件为 Q 为空，此后对象图将不再发生变化. 给定一个多类型的、关联的对象集合，存在一个唯一的、最小的对象图与之对应，称为干净的对象图. GBi-JER 接收一个多类型的、关联的对象集合，输出一个干净的对象图作为联合式实体识别的结果. GBi-JER 是灵活的，不同的基于属性相似度算法和基于结构的相似度算法都可以嵌入到其中，发挥其优势.

3.3 初始化工作

GBi-JER 的初始化工作包括对象图初始化和候选队列初始化. 初始化工作是个冷启动的过程.

首先，对象图初始化，将关系型数据集根据对象图定义和数据集的模式信息构建出多个小型的、原始的子对象图. 比如，将图1(b)中 c_1，a_1，a_2，a_3，ve_1 构建成一个原始的子对象图，包括3类5个对象，2类4条语义链接，如图4所示. 随着联合式实体识别的进行，经历对象合并和局部图收缩，多个原始的

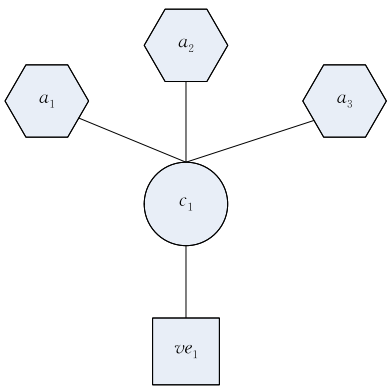

图 4 原始的对象图示例

子对象图将不断地融合在一起，逐渐聚合成大型的对象图；在此过程中，对象图的结点数目逐渐减少，密度逐渐增加，语义逐渐丰富. 构建候选对象图的代价为 $O(i+j)$，i 是对象总数，j 是链接总数. 链接可用于结构相似度计算，对象图可以比依赖图[8,14]更好的引导相似度传递.

其次，候选队列初始化. 使用 Canopy 的分块技术[13]对对象分块，块内生成初始的候选匹配对象对，称这些候选匹配对象对为 I 类候选匹配对. GBi-JER 使用一个优先队列 Q 作为候选队列来维持候选匹配对的识别顺序，Q 的每个结点存储一对候选匹配对和相应的优先级. 将初始的候选匹配对依次插入到 Q 中，并给予相同的、较低的优先级，将关联的对象产生的候选匹配对放在邻近的位置，以便 GBi-JER 及时地识别出关联的实体，提高识别效率. 后文相似度传递过程中，候选匹配对的优先级将可能发生变化，这将是对实体识别顺序的优化.

4　联合式匹配

4.1 节介绍 GBi-JER 的联合式匹配函数；4.2 节介绍基于语义路径的相似度算法.

给定一个对象图 $G(O_{mr},L_{mr})$，$O_{mr}=\{O_t\}_{t=1}^{T}$ 是一个多类型的、关联的对象集合，O_t 是第 t 类对象子集，$L_{mr}\subseteq O_{mr}\times O_{mr}$ 是不同的对象间的语义链接的集合.

4.1　联合式匹配函数

$o_i,o_j\in O_t$. 定义一个混合的对象相似度函数 sim_{hyb}，

$$sim_{\text{hyb}}(o_i,o_j)=(1-\delta)\times sim_{\text{abs}}(o_i,o_j)+\delta\times sim_{\text{sbs}}(o_i,o_j) \tag{1}$$

其中，sim_{abs} 是 o_i 和 o_j 的基于属性的相似度，sim_{sbs} 是 o_i 和 o_j 的基于结构的相似度，δ 是两种相似度的权值分配系数，可由用户根据两种相似度的重要性进行设定. sim_{abs} 采用已有的基于属性的相似度算法[15]；sim_{sbs} 采用后文将提出的一种基于语义路径的相似度算法. sim_{sbs} 也可以采用任何基于结构的相似度的算法[16].

GBi-JER 的联合式匹配函数 $jm_{\text{GBi-JER}}$，首先，计算 o_i 和 o_j 的对象相似度 $sim_{\text{hyb}}(o_i,o_j)$；然后，将 $sim_{\text{hyb}}(o_i,o_j)$ 与给定的匹配阈值 $\theta_{m\text{-hyb}}$ 比较，如果 $sim_{\text{hyb}}(o_i,o_j)\geqslant\theta_{m\text{-hyb}}$，那么 $jm_{\text{GBi-JER}}(o_i,o_j)=\text{true}$，$o_i\approx o_j$；否则，$jm_{\text{GBi-JER}}(o_i,o_j)=\text{false}$，$o_i\neq o_j$. 其中 $\theta_{m\text{-hyb}}$ 由领域专家给出或通过实验测定.

4.2　基于语义路径的相似度算法

本节介绍一种基语义路径的相似度算法，它属于基于结构的相似度.

结合对象图和模式图可知，模式图中的一条语义边对其两端的两个对象类型的重要程度通常是不同，除非这是一条一一映射的语义边. 现将一条双向的语义边分折成两条单向的语义边，并赋予它们相应的权值，称为单向关联权值，如图 5 所示. 赋予不同的语义边不同的单向关联权值，表示不同强弱的语义关系. 单向关联权值通过有向边的头尾结点的对象类型之间的对应关系确定，比如每篇文章只对应一个会议，因此 ω_{damp}（文章，会议）$=1/1$，而每个会议平均对应 x 篇文章，x 可取分数，因此 ω_{damp}（会议，文章）$=1/x$，图 5 示例中平均每个会议对应 $x=56.8$ 篇文章. 单向关联权值可以通过领域专家给出或者通过实验得到. 本文采用基于实验的方法得到单向关联权值，随机抽取一定比例的给定数据集的数据，得到样本的实体识别结果后构建对象图，统计关联的对象类型之间的平均对象对应数目，语义边的头类型的对象总数作分子，语义边的尾类型的对象数的总和作分母，即

$$\varpi_{\text{damp}}(\mu_i,\mu_j)=\frac{\sum_{\varphi(o_k)=\mu_i}|o_k|}{\sum_{\varphi(o_k)=\mu_i}\ \sum_{\varphi(o_m)=\mu_j,o_k\to o_m}|o_m|} \tag{2}$$

其中，分子部分表示类型为 μ_i 的对象数，分母部分表示与类型为 μ_i 的对象对应的类型为 μ_j 的对象数.

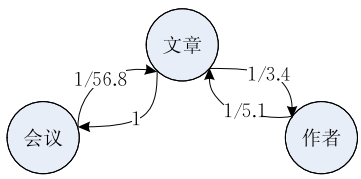

图 5　模式图中单向关联权值示例

图上两个结点的相似度通过结构来衡量，通常有两类方法[16-17]：基于相邻结点的方法和基于路径求和的方法. 基于相邻结点的方法只计算直接相邻的结点，不能利用间接的关系；基于路径求和的方法利用了结点间直接和间接的关系. 后者比前者考虑了更丰富的结构信息，因此更准确. 本文提出的基于语义路径的相似度算法属于基于路径求和的方法.

本文采用随机游走模型[18]来计算对象图中两点的结构相似度. 给定对象图 G，o_s，o_t 是 G 上的两个对象结点，命中时间（Hitting time）H_{o_s,o_t} 表示一次从 o_s 到达 o_t 的随机游走的期望步数，命中时间是非对称的；往返时间（Commute time）$C_{o_s,o_t}=H_{o_s,o_t}+$

H_{o_t,o_s}表示一次从o_s到达o_t再返回到o_s的随机游走的期望步数,往返时间是对称的.命中时间对应单向随机游走模型,往返时间对应双向随机游走模型.由于相邻对象间不同方向的语义链接的单向关联权值不同,因此不能使用单向的随机游走模型而需要使用双向的随机游走模型来计算对象的相似度,即从o_s沿某条路径$sp(o_s,o_t)$到达o_t再沿着$sp(o_s,o_t)^{-1}$($sp(o_s,o_t)$的逆路径)返回到o_s.

定义 14. 模式路径的关联度.给定模式图$G_{sch}=(N,\Gamma)$,$\mu_i\in N$,$i\in[0,w]$,存在模式路径$schp(\mu_0,\mu_w)=(\mu_0,\cdots,\mu_i,\cdots,\mu_w)$,其关联度(connection)为

$$con(schp(\mu_0,\mu_w))=\prod_{i=0}^{w-1}\varpi_{\text{damp}}(\mu_i,\mu_{i+1}) \quad (3)$$

定义 15. 两对象沿着某语义路径的关联度.两对象$o_0,o_w\in O_{mr}$之间的一条长度为$w+1$语义路径$sp(o_0,o_w)=(o_0,\cdots,o_i,\cdots,o_w)$,$0\leq i\leq w$,那么语义路径$sp(o_0,o_w)$的关联度为

$$con(sp(o_0,o_w))=pr_{\text{ht}}(sp(o_0,o_w))\times con(schp(sp(o_0,o_w))) \quad (4)$$

其中,$pr_{\text{ht}}(sp(o_0,o_w))$表示沿着$sp(o_0,o_w)$的随机游走概率,$con(schp(sp(o_0,o_w)))$是$sp(o_0,o_w)$的模式路径的关联度,表示该路径的重要程度.那么,对象o_0,o_w沿着$sp(o_0,o_w)$的关联度为

$$\begin{aligned}con(o_0,o_w)_{sp(o_0,o_w)}&=con(o_0,o_w)_{sp(o_0,o_w)^{-1}}\\&=\frac{1}{2}\times(con(sp(o_0,o_w))+con(sp(o_0,o_w)^{-1}))\end{aligned} \quad (5)$$

从式(5)可知,对象o_0,o_w沿着$sp(o_0,o_w)$或$sp(o_0,o_w)^{-1}$的关联度具有对称性.

对象图中两对象的结构相似度可以通过对两者沿着所有语义路径的关联度求和得到,但是,过长的路径的关联度很弱,计算代价却很大,因此,为了平衡开销和准确性,本文将限定两结点间路径的长度.对象图中两结点o_i,o_j之间所有路径的集合记作$SP_{\text{all}}(o_i,o_j)$,把长度不大于len的所有路径的集合记作$SP_{len}(o_i,o_j)$.本文通过实验测定len取8.

定义 16. 基于语义路径的相似度.两个对象$o_i,o_j\in O_{mr}$,那么o_i,o_j基于语义路径的相似度为

$$sim_{\text{path}}(o_i,o_j)=\sum_{sp\in SP_{len}(o_i,o_j)}con(o_i,o_j)_{sp} \quad (6)$$

接下来,解决引言中提出的第②个问题,图1(b)中a_{12}与a_2还是a_{13}更匹配,以此来展示基于语义路径的相似度在联合式实体识别中的作用.

初始的时候,GBi-JER无法确定a_{12}与a_2还是a_{13}更匹配,它们的属性相似度非常接近.然而,当实体识别进行到一定阶段,如图6所示,a_{12},a_2,a_{13}周边的对象已经完成匹配,a_2已与a_5合并成为组合对象$\langle a_2,a_5\rangle$,此时,分析语义路径发现,a_{12}与$\langle a_2,a_5\rangle$存在语义路径($a_{12},c_5,\langle a_{10},a_{11}\rangle,c_4,\langle a_3,a_6,a_9\rangle,\langle c_1,c_2\rangle,\langle a_2,a_5\rangle$)和($a_{12},c_5,\langle a_{10},a_{11}\rangle,c_4,\langle a_3,a_6,a_9\rangle,c_3,\langle a_1,a_4,a_7\rangle,\langle c_1,c_2\rangle,\langle a_2,a_5\rangle$),而$a_{12}$与$a_{13}$之间不存在语义路径,由此,GBi-JER判定$\langle a_2,a_5\rangle$与$a_{12}$匹配,而$a_{12}$与$a_{13}$不匹配.

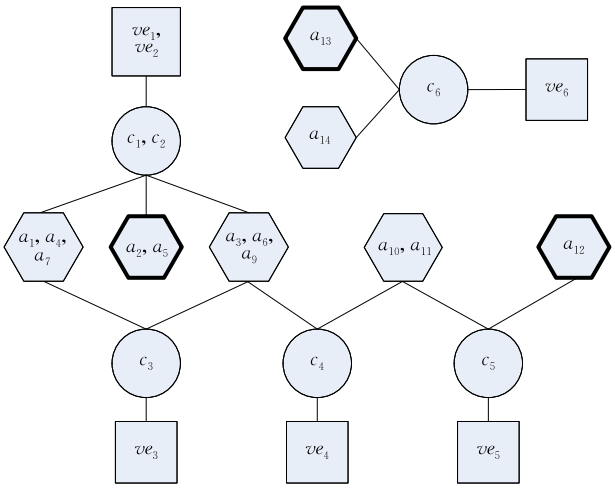

图6 利用基于语义路径的相似度的实体识别示例

5 联合式合并与相似度传递

GBi-JER合并匹配成功的对象对,对该对象对在对象图中对应的两结点及其周边执行局部图收缩.对象合并可能产生对象的信息增益,局部图收缩能够传递对象的相似度,引发已匹配对周边对象的重新计算.这两个操作使得对象图具有动态性.

首先,通过一个示例来了解联合式合并与相似度传递的梗概.以图1(b)中$c_1,a_1,a_2,a_3,ve_1,c_2,a_4,a_5,a_6,ve_2$为例,两个会议$ve_1,ve_2$已经匹配,已合并为一个组合对象$\langle ve_1,ve_2\rangle$,这些对象构成的对象关系子图如图7(a)所示,此时两篇对应的文章被判定为候选匹配对.接着$jm_{\text{GBi-JER}}(c_1,c_2)=\text{true}$,$ju_{\text{GBi-JER}}$将之合并为$\langle c_1,c_2\rangle$,结点$\langle c_1,c_2\rangle$将继承$c_1$和$c_2$的原有语义链接$l_1,l_2,l_3,l_4,l_5,l_6,l_7,l_8$,此时$\langle c_1,c_2\rangle$与$\langle ve_1,ve_2\rangle$出现两条相同的语义链接$l_7$和$l_8$,出现冗余,需随机删除一条$l_8$.当执行完上述操作后,$\langle c_1,c_2\rangle$关联对象的周边结构发生变化,$a_1$和$a_4$,$a_2$和$a_5$,$a_3$和$a_6$成为3对新候选匹配对,将3对新候选匹配对加入到候选队列Q中的合适位置.如此过程,循环往复直到Q为空.

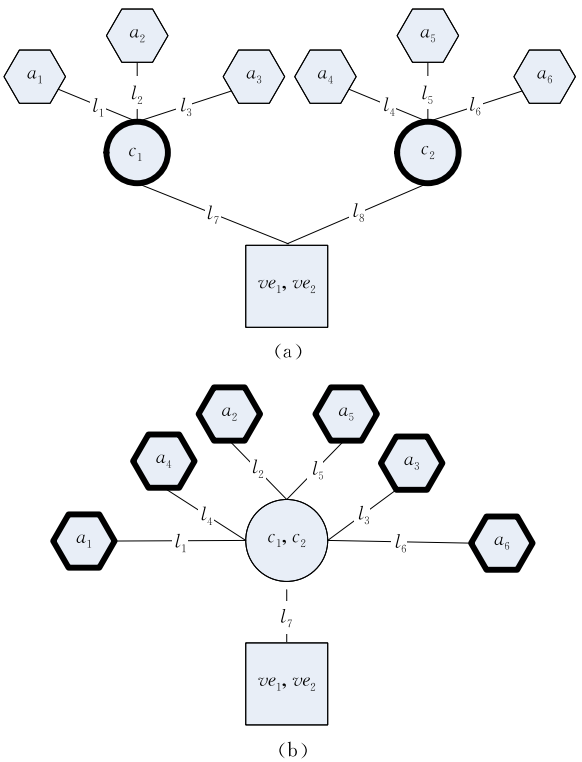

图 7　联合式合并示例

接下来，先介绍联合式合并的两个重要步骤：数据对象合并和局部图收缩；然后阐述局部图收缩带来的影响：相似度传递.

5.1　数据对象合并

对象合并能产生信息增益，消除假值（真值发现），提高联合式实体识别的精确性. 对象合并原则：对象合并前后的信息总量不能发生变化，既不能增加也不能减少.

5.1.1　简单的对象合并方法

一个简单、直观的合并方法是为对象的每个属性保留所有非重复的值. 这种方法的优点是所有的对象信息都保留，不会违反对象合并原则. 它的缺点是，保留所有属性值需要较高的存储开销，并且为后续的对象比较带来了较高的计算复杂度；更坏的情况下相同属性的多个属性值之间存在冲突，如果保留冲突的数据，会影响后续的实体识别的准确性. 以图 1(b)中作者对象 a_3, a_6, a_8, a_9 为例，假定第一趟比较后 $a_3 \approx a_9, a_6 \approx a_8$，因此，$\langle a_3, a_9 \rangle$｛“Mike Leo Bush”，“Mike Bush”｝，$\langle a_6, a_8 \rangle$｛“M. Lee Bush”，“M. L. Bush”｝，花括号内表示作者姓名的值；第二趟比较后，$\langle a_3, a_9 \rangle \approx \langle a_6, a_8 \rangle$，因此，$\langle a_3, a_9, a_6, a_8 \rangle$｛“Mike Leo Bush”，“Mike Bush”，“M. Lee Bush”，“M. L. Bush”｝. 分析第二趟比较过程，对于某对对象的某个属性，总共要进行 $i \times j$ 次属性比较，i, j 表示两个对象的某个属性分别拥有值的个数. 这是一个二次方数量级的问题. 同时，注意到 $\langle a_3, a_9, a_6, a_8 \rangle$ 中作者姓名的两个值“Mike Leo Bush”和“M. Lee Bush”的中间名“Leo”和“Lee”之间存在冲突，这种冲突应该消除.

5.1.2　基于属性代表值的对象合并方法

针对简单的合并方法的缺点，本文提出一种基于属性代表值的合并方法. 该方法在对象合并后为每个属性选择最有代表性的值，这样的值包括了合并前该属性对应的值的所有信息；采用数据冲突解决的方法消除掉数据冲突. 继续使用上一小节的例子，识别顺序不变，采用本节的合并方法，第一趟比较后，有 $\langle a_3, a_9 \rangle$｛“Mike Leo Bush”｝，$\langle a_6, a_8 \rangle$｛“M. Lee Bush”｝，而不是 $\langle a_3, a_9 \rangle$｛“Mike Leo Bush”，“Mike Bush”｝，$\langle a_6, a_8 \rangle$｛“M. Lee Bush”，“M. L. Bush”｝，显然可以节省较大的存储开销；第二趟比较后，$\langle a_3, a_9 \rangle \approx \langle a_6, a_8 \rangle$，此时作者姓名属性的值出现冲突“Leo”和“Lee”，假定通过数据冲突解决技术判定，“Lee”是由于拼写错“Leo”才出现的，因此，$\langle a_3, a_9, a_6, a_8 \rangle$｛“Mike Leo Bush”｝. 数据冲突解决不是本文的研究主题，采用现有的数据冲突解决技术[19]. $\langle a_3, a_9, a_6, a_8 \rangle$｛“Mike Leo Bush”｝能够代表所有的输入值“Mike Leo Bush”，“Mike Bush”，“M. Lee Bush”，“M. L. Bush”. 注意到第二趟比较，$\langle a_3, a_9 \rangle$ 与 $\langle a_6, a_8 \rangle$ 的姓名比较时只进行了一次比较，而上小节进行了 $i \times j$ 次. 已有的对象合并方法没有考虑过数据冲突的影响. 对于有冲突的数据，如果不解决掉数据冲突，把所有值都保留下来，后续的匹配过程将产生错误的结果. 使用数据冲突解决技术，可以正确地解决大部分的数据冲突，保证了合并结果的正确性和后续匹配的正确性，这样最大限度地提高了实体识别的准确性. 综上分析，基于属性代表值的合并方法比简单的合并方法要更好.

5.2　局部图收缩

当匹配成功的对象合并后，需要对对象图中这两个对象及其周边执行局部图收缩（Local Graph Contraction，LGC）.

一个简单的局部图收缩的方法是将匹配的结点及连接它们的候选链接合并为一个新结点，并让它继承合并前结点的所有语义链接. 这个方法的问题是可能产生冗余的语义链接. 以图 7 为例，当 c_1, c_2

合并后，结点$\langle c_1,c_2\rangle$将继承 c_1和 c_2的原有语义链接 $l_1,l_2,l_3,l_4,l_5,l_6,l_7,l_8$，此时$\langle c_1,c_2\rangle$与$\langle ve_1,ve_2\rangle$出现两条相同的语义链接 l_7和 l_8，出现冗余. 当两结点间的多条语义链接被判定有冗余时，随机保留一条即可. 图 7 示例中，保留 l_7.

定义 17. 局部图收缩(LGC). $o_i,o_j\in O_t$，$o_i\approx o_j$，那么，将 o_i,o_j合并为结点$\langle o_i,o_j\rangle$，将合并前 o_i,o_j的语义链接与$\langle o_i,o_j\rangle$相连，并去除掉其中冗余的语义链接.

5.3 相似度传递

当已匹配的对象对完成联合式合并后，合并的对象结点的周边变得更紧密、语义关系更丰富，因此，该对象邻近的对象中可能会出现新的候选匹配对. 图 7(a)中，c_1和 c_2是一对候选匹配对，对它们进行联合式识别，$jm_{\text{GBi-JER}}(c_1,c_2)=\text{true}$，$c_1\approx c_2$，对象合并得到$\langle c_1,c_2\rangle$，进行局部图收缩操作后如图 7(b)中所示，$a_1$和 a_4，a_2和 a_5，a_3和 a_6之间各自的结构相似度增大，成为新的候选匹配对，将它们插入到候选队列 Q 的适合的位置. 这里的插入位置的选择是非常重要的，后文将对此专门阐述. 图 8 接图 7 示例，对 a_1和 a_4进行联合式匹配，匹配成功，后续过程与图 7 中(a)、(b)类似，不赘述. 总之，局部图收缩使得 GBi-JER 成为一个迭代的过程，引发周边对象的重新计算，称为相似度传递.

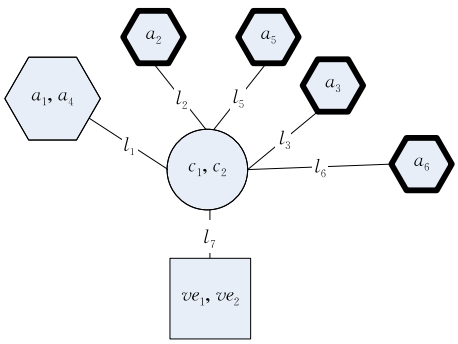

图 8 局部图收缩示例(续)

除了直接关联的对象，通过语义路径关联的对象间也可能发生相似度的传递，比如合作者关系，就是一种较强的对象关系，以图 1(b)为例，假定已生成$\langle a_1,a_4\rangle$，$\langle a_3,a_6\rangle$，当$\langle a_1,a_4\rangle$与 a_7匹配后，增加了$\langle a_3,a_6\rangle$与 a_8匹配的可能性，因此，生成一对候选匹配对$\langle a_3,a_6\rangle$与 a_8.

定义 18. 相似度传递. $o_i,o_j\in O_x$，$i\neq j$，$o_i\approx o_j$，如果存在$o_g,o_h\in O_y$，$1\leqslant x,y\leqslant T$，$schp(sp(o_i,o_g))=schp(sp(o_j,o_h))$，$con(schp(sp(o_i,o_g)))=con(schp(sp(o_j,o_h)))\geqslant\theta_\omega$，且 o_g,o_h满足 Canopy 分块条件，那么 o_g,o_h成为一对候选匹配对；特别地，如果 o_g,o_h是由于属性值部分缺失而导致不满足 Canopy 分块条件，则也将 o_g,o_h生成一对候选匹配对.

相似度传递产生的候选匹配对称为 II 类候选匹配对. 为了平衡开销和效率，本文限定定义 18 中相似度传递中语义路径 $sp(o_i,o_g)$长度最大为 2，即只考虑已匹配结点的直接相邻结点和次邻结点. θ_ω是相似度传递中语义路径的模式路径关联度的阈值，由领域专家给定的或通过实验测定. 定义 18 中特殊情况的设定，是为了保证对象在属性值部分缺失的情况下，也可以通过结构进行匹配. 本文引言中提出的第①类问题就属于此情况. a_1和 a_4不能生成 I 类候选匹配对，但通过相似度传递生成 II 类候选匹配对，并通过较高的结构相似度得以匹配成功，图 7(b)和图 8 展示了 a_1和 a_4匹配成功的过程.

新的候选匹配对在候选队列 Q 中的插入位置将决定其被识别的顺序，进而影响整个对象集的识别顺序. GBi-JER 希望先识别更可能匹配成功的候选匹配对，候选匹配对的匹配顺序将通过它们的属性相似度和相似度传递中语义路径的模式路径关联度估计，此处属性相似度直接使用 Canopy 分块时的代价较小的、近似的属性相似度，记作 $sim_{\text{canopy-abs}}$. 定义 19 给出一个优先级打分函数 $score$，GBi-JER 根据候选匹配对的 $score$ 来决定其在 Q 中的插入位置.

定义 19. 优先级打分函数. $o_i,o_j\in O_x$，$o_g,o_h\in O_y$是通过 $o_i\approx o_j$生成的候选匹配对，$schp(sp(o_i,o_g))=schp(sp(o_j,o_h))$，$1\leqslant x,y\leqslant T$，那么 o_g,o_h在 Q 中优先级由下面打分函数决定，

$$score(o_g,o_h)=(1-\eta)\times sim_{\text{canopy-abs}}(o_g,o_h)+\eta\times con(schp(sp(o_i,o_g))) \quad (7)$$

其中，$sim_{\text{canopy-abs}}$是 o_g,o_h的近似的属性相似度，$con(schp(sp(o_i,o_g)))$是相似度传递中语义路径的模式路径关联度，η是两者的权值分配系数，可由用户根据两者的重要性自主设定.

当某 II 类候选匹配对已经在 Q 中时，将本次优先级与已有优先级进行比较，如果本次优先级更高则将对象对插入到新的位置，否则不作处理. 除了上述的基于打分值的优先级插入策略，还有插入到队首，插入到队尾，随机插入等策略. 插入到队首是指将新生成的 II 类候选匹配对插入到候选队列的头；插入到队尾，如果新生成的 II 类候选匹配对已存在于候选队列，则不做处理，否则将其插入到候选队列的尾；随机插入，如果新生成的 II 类候选匹配对已存在于候选队列，则不做处理，否则将其随机地插入

到候选队列中.

综合本小节内容可知，通过相似度传递可以解决引言中提出的第③类问题.

6　实验与分析

6.1　准备工作

实验环境. 处理器：Intel(R) Core(TM) i7-2600，主频 3.4 GHz，8 核；内存：8 GB；操作系统：Microsoft Windows 7 Ultimate，64 位.

数据集. 本文采用实体识别研究中常用的 Citeseer 数据进行主体实验，另外还采用了 IMDB 数据进行准确性实验以验证 GBi-JER 在不同领域的有效性. 本文从 Citeseer 数据库中采集了一个数据集，包括约 100 000 条引文条目，引用了约 9000 篇论文. Citeseer 中引文的数据模式是{title，{author，address，affiliation}*，date，venue}，后缀“*”表示一篇文章可能存在多个作者. 本文将每条引文条目分解为 4 类对象：文章{title，date，*written-by，*published-in}，作者{name，*affiliated-by}，单位{name，address}和会议{name，date}，其中*written-by，*published-in 和*affiliated-by 表示对象关系. 为了评价算法，本文作者对 Citeseer 数据集的论文、作者、单位和会议进行了人工标注得到真实的匹配结果. 本文从约 100 000 条引文条目中生成一个 50 000 条引文条目的子集，记作 Citeseer-1，用于除 6.2.3 节以外的所有实验. IMDB 数据集包括电影、导演、演员和出品公司 4 类数据对象，数据规模分别为电影(2000)、导演(312)、演员(4526)、出品公司(204)、电影与其他 3 类对象间存在语义关系，该数据集记作 IMDB，只参与 6.2.2 节实验.

评价指标. 本文采用 F 指数作为评价实体识别结果的精确性的指标. F 指数是准确率和召回率的调和平均数，公式如下：

$$F=\frac{2\times P\times R}{P+R} \tag{8}$$

其中，P 是准确率，R 是召回率.

模式图中单向关联权值测定. 从 Citeseer-1 中抽取 1 W 条引文条目记作 Citeseer-train，具体过程：随机抽取一条条目，参照人工标注结果，将包含该条目中涉及的文章、作者、会议和单位的条目抽取出来，重复上述过程直到提取数目达到 10 000. 将 Citeseer-train 中引文条目分解为 4 类对象，并参照人工标注结果构建对象图，根据式(2)，计算单向关联权值. 得到结果为 ω_{damp}(会议，文章)=1/56.8，ω_{damp}(文章，会议)=1，ω_{damp}(作者，文章)=1/5.1，ω_{damp}(文章，作者)=1/3.4，ω_{damp}(作者，单位)=1/1.2，ω_{damp}(单位，作者)=1/15.7. IMDB 数据集也采用类似的方法测定.

6.2　实验结果与分析

6.2.1　关键参数测定

本节针对 GBi-JER 方法中的关键参数进行测试，包括基于语义路径的相似度中的路径长度阈值 len，联合式匹配函数的权值分配系数 δ 和匹配阈值 $\theta_{m\text{-hyb}}$，相似度传递中的阈值 θ_{ω}，优先级打分函数的权重分配系数 η.

(1) 基于语义路径的相似度中的路径长度阈值 len

图 9 中，纵主轴是 F 指数，纵次轴是时间，F 均是 4 类对象识别的平均 F 指数(记作“F-均”). 观察发现，F-均随着路径长度的上界的增长而增长，当长度上界大于 8 后，F-均趋于平稳；算法的时间开销随长度的上界以幂率的方式增长，特别地，当长度大于 8 以后，时间开销增长巨大. 综合考虑 F-均和时间曲线，基于语义路径的相似度中的路径长度阈值 len 取 8.

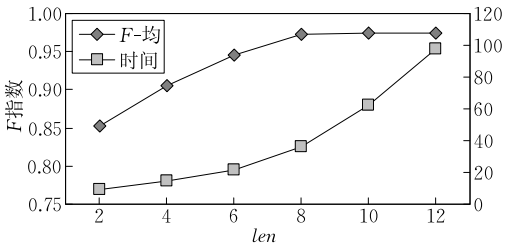

图 9　路径长度阈值 len 设置

(2) 联合式匹配函数的权值分配系数 δ

图 10 显示了权值分配系数 δ 对平均 F 指数的影响，当 δ 取 0.45，F-均取得最大值，因此，δ 取 0.45.

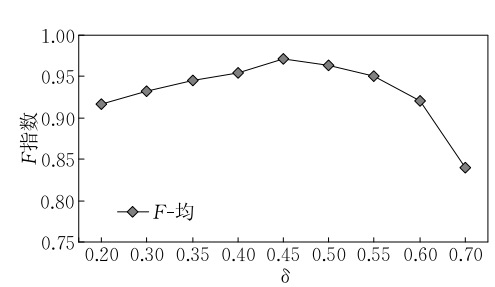

图 10　联合式匹配函数的权值分配系数 δ 设置

(3) 联合式匹配函数的匹配阈值 $\theta_{m\text{-hyb}}$

图 11 显示了联合式匹配函数的匹配阈值 $\theta_{m\text{-hyb}}$ 对平均 F 指数的影响，本文将 $\theta_{m\text{-hyb}}$ 设置为 0.85.

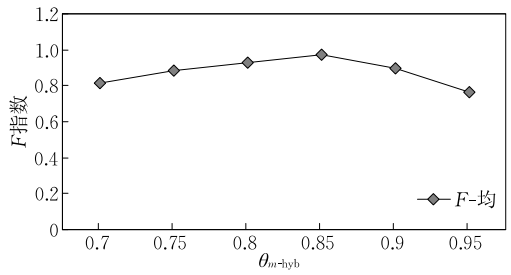

图 11 联合式匹配函数的匹配阈值 $\theta_{m\text{-}hyb}$ 设置

(4) 相似度传递中的阈值 θ_ω

图 12 中,纵主轴是时间,纵次轴是 F 指数. 相似度传递的阈值 θ_ω 对联合式实体识别的时间开销和平均 F 指数的影响如图 12 中曲线所示,当 θ_ω 取 0.025 时,F-均取得最大值,而此时的时间开销较小,因此,将 θ_ω 设置为 0.025.

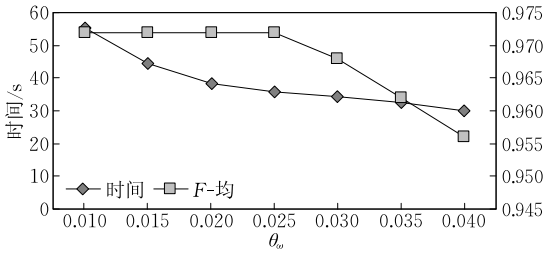

图 12 相似度传递中的阈值 θ_ω 设置

(5) 优先级打分函数的权重分配系数 η

图 13 显示了优先级打分函数的权重分配系数 η 对联合式实体识别的时间开销的影响,当 η 为 0.65 时,时间开销最小,因此,将 η 设置为 0.65.

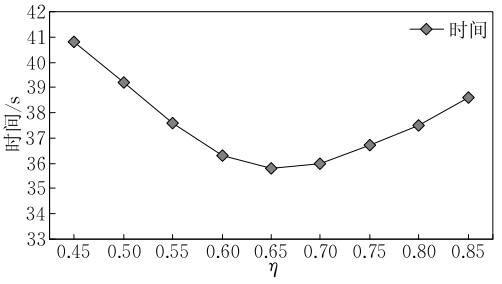

图 13 优先级打分函数的权重分配系数 η

6.2.2 GBi-JER 与已有方法的对比

传统的实体识别方法仅比较对象的属性相似度,记作传统方法. Kalashnikov 等人[6]提出了一种基于对象图中对象关联强度的单类型实体识别方法,记作 RelDC. Dong 等人[1]的方法是联合式实体识别方法,基于对象依赖关系构建对象对的依赖图,计算结构相似度时只考虑直接相邻的对象,一定程度上实现了相似度传递,记作 DepGraph.

图 14 中,在 Citeseer-1 数据集上,GBi-JER 对 4 类对象的实体识别的准确性都要优于其他 3 个方法;图 15 中,在 IMDB 数据集上,GBi-JER 对 4 类对象的识别的准确性也优于其他 3 个方法. 因为 GBi-JER 能充分地发掘对象语义关系(直接的和间接的),更准确地计算结构相似度,并随着局部图收缩和相似度传递的不断迭代,对象图逐渐收敛,可进一步促进后续实体识别. RelDC 和 DepGraph 都优于传统方法,因为二者都利用了对象的结构相似度;DepGraph 利用相邻的对象间相似度传递,不同对象的识别促进了其他对象的识别,DepGraph 的准确性高于 RelDC.

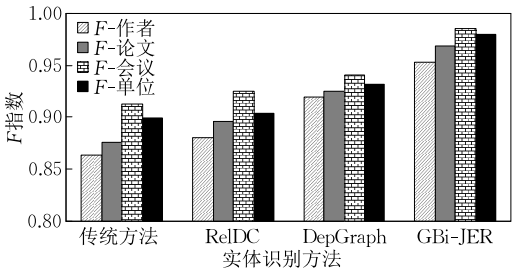

图 14 GBi-JER 与已有方法准确性对比(1)

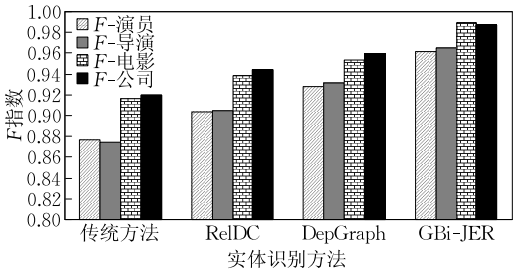

图 15 GBi-JER 与已有方法准确性对比(2)

6.2.3 GBi-JER 性能测试

为了测试 GBi-JER 的性能,将其分别在规模为 10 000, 20 000, 30 000, …, 100 000 的数据集上进行实验,相邻数据集规模相差 10 000. 观察图 16,GBi-JER 随着数据集的增长,其时间开销的增长接近于线性,这得益于相似度的传递,生成候选匹配对,引导实体识别以高效的顺序进行,避免平方级的增长.

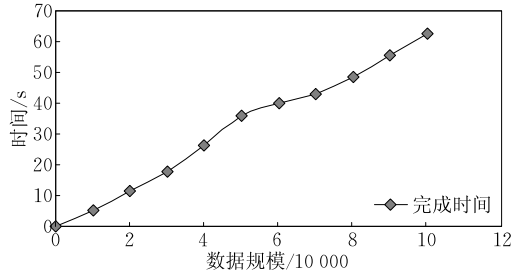

图 16 GBi-JER 性能测试

6.2.4　GBi-JER 的关键组成部分测试

本节测试 GBi-JER 的关键组成部分. 为了更加清晰地说明后文实验中各种方法的组成部分的异同，给出表 1 和表 2. 表 1 中“+”表示存在该组成部分，“−”表示不存在该组成部分，英文缩写代表具体的组成部分. 英文缩写的意义见表 2.

表 1　实体识别方法组成

实体识别方法	ABS	SBS	对象合并策略	局部图收缩	相似度传递	插入策略
传统方法	+	−	S-mg	−	−	−
GJ-1	+	SPBS	S-mg	−	−	−
GJ-2	+	SPBS	ARB-mg	−	−	−
GJ-3	+	SPBS	ARB-mg	+	−	−
GBi-JER	+	SPBS	ARB-mg	+	+	PB-in
GJ-RW	+	RWBS	ARB-mg	+	+	PB-in
GJ-SRW	+	SRWBS	ARB-mg	+	+	PB-in
GJ-Smerge	+	SPBS	S-mg	+	+	PB-in
GJ-InQH	+	SPBS	ARB-mg	+	+	QH-in
GJ-InR	+	SPBS	ARB-mg	+	+	R-in
GJ-InQT	+	SPBS	ARB-mg	+	+	QT-in

表 2　表 1 中缩写注解

缩写	全称
SPBS	基于语义路径的相似度算法
RWBS	基于随机游走的相似度算法
SRWBS	基于双向随机游走的相似度算法
S-mg	简单的对象合并方法
ARB-mg	基于属性代表值的对象合并方法
PB-in	基于优先级的插入策略
QH-in	插入到队首
R-in	随机插入
QT-in	插入到队尾

(1) 关键组成部分贡献测试

为了测试 GBi-JER 各组成部分的作用，进行如下比较：GJ-1 比传统方法增加了基于语义路径的相似度，GJ-2 将 GJ-1 的对象合并策略替换为基于属性代表值的合并方法，GJ-3 比 GJ-2 增加了局部图收缩，GBi-JER 比 GJ-3 增加了相似度传递.

观察图 17，GJ-1 比传统方法的准确性高，说明结构相似度(基于语义路径的相似度)对实体识别准确性有促进作用，结构相似度有助于解决引言中的第①、②类问题；GJ-2 与 GJ-1 的准确性大体相当，说明基于属性代表值的合并方法和简单的合并方法对实体识别准确性作用大体相同；GJ-3 比 GJ-2 的准确性高出很多，可见局部图收缩对联合式实体识别的作用是非常关键的，对象匹配成功后执行图收缩，提高周边的对象之间的结构相似度，促进了后续的实体识别；GBi-JER 比 GJ-3 的准确性高，说明相似度传递提高了联合式实体识别的准确性，相似度传递有助于解决引言中的第①、③类问题.

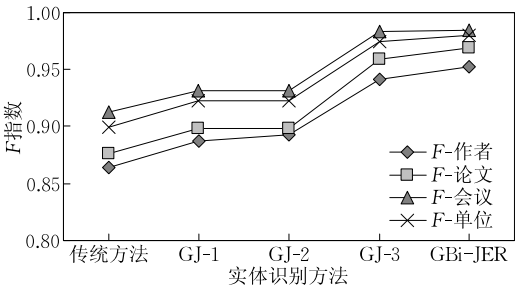

图 17　关键组成部分测试

接下来对 GBi-JER 的关键组成部分单独测试.

(2) 基于语义路径的相似度

将 GBi-JER 的联合式匹配函数的结构相似度用不同方法实现，其他部分不变，基于单向随机游走算法的记作 GJ-RW，基于双向随机游走算法的记作 GJ-SRW.

观察图 18，从 GJ-RW，GJ-SRW 到 GBi-JER，实验结果 F 指数依次增长. GBi-JER 比 GJ-RW，GJ-SRW 的精确性更高，说明基于语义路径的相似度算法比单向和双向的随机游走算法能更准确地衡量对象图中对象的相似度.

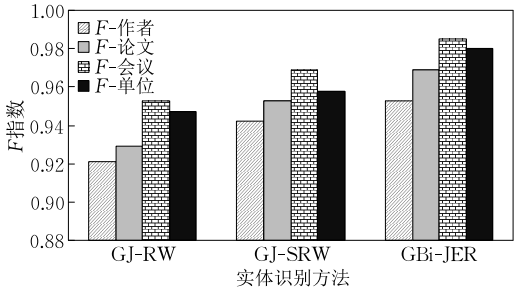

图 18　基于语义路径的相似度算法与其他方法对比

(3) 对象合并策略

将 GBi-JER 的对象合并策略采用简单的对象合记作 GJ-Smerge.

观察图 19，GBi-JER 与 GJ-Smerge 的准确性大体相当，只有在作者的实体识别中 GBi-JER 的准确性稍高，作者的名字变化的情况较多，比如缩写方式不同、部分缺失等，基于属性代表值的合并方法选择

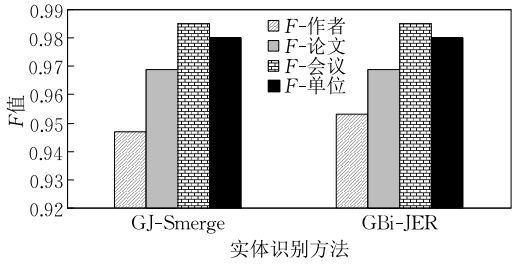

图 19　对象合并策略对比

合并后最具代表性的值且解决了数据冲突,因此,获得了更高的准确性.

GJ-Smerge 的完成时间为 41.9 s,而 GBi-JER 的完成时间为 35.8 s,GJ-Smerge 比 GBi-JER 花费的时间多出 17%.综合准确性和时间开销来看,基于属性代表值的合并方法要比简单的合并方法要更优.

(4)相似度传递测试

GJ-3 比 GBi-JER 减少了相似度传递.为了评价相似度传递对识别速度的影响,本小节将测试 GBi-JER 和 GJ-3 随着时间变化,4 类对象识别的 F 指数的平均值的变化情况.

观察图 20,0 到 35.8 s 区间内,GBi-JER 的 F-均增长速度远大于 GJ-3,并在 35.8 s 处达到最大值 0.972,而此时 GJ-3 的 F-均为 0.4 左右;在图 14 时间轴最大值 120 s,GJ-3 的 F-均达到 0.8 左右,与 GBi-JER 的 F-均仍有较大差距;GJ-3 要达到其 F-均最大值,仍然需要较长时间开销.可见相似度传递提高识别速度的作用非常大.

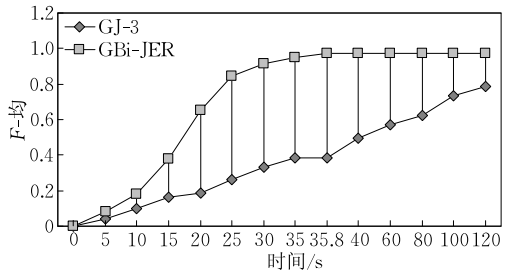

图 20 相似度传递对识别速度的影响

(5)II 类候选匹配对插入策略

虽然相似度传递时 II 类候选匹配对如何插入候选队列不会影响 GBi-JER 的准确性,但会影响 GBi-JER 的性能.本小节针对不同的插入策略进行性能测试,将 GBi-JER 的插入策略分别采用插入到队首、随机插入和插入到队尾,分别记作 GJ-InQH、GJ-InR 和 GJ-InQT.

观察图 21,不同的插入策略的时间开销是不相同的,其中 GBi-JER 的时间开销最小,说明本文提出的基于优先级的插入策略是一个好的插入策略,能够更快地识别出匹配的对象.基于优先级的插入策略根据数候选匹配对匹配的可能性来决定匹配顺序,比其他 3 种策略更优;插入到队首的策略将 II 类候选匹配对一律插入到队首,可以快速地识别出一部分 II 类候选匹配对,但是有一些 II 类候选匹配对由于其周边的结构还不够紧密,语义信息不够丰富,因此并不能被识别出,需要经过反复地匹配,经历周边结构的变化,直到相似度达到阈值才能匹配成功,造成了较高的代价;随机插入策略比插入到队首的时间开销更高,因为该策略没有及时地识别 II 类候选匹配对;插入到队尾的时间开销最高,因为该策略将新生成 II 类候选匹配对插入到候选队列的尾,无法发挥出 GBi-JER 的优势.

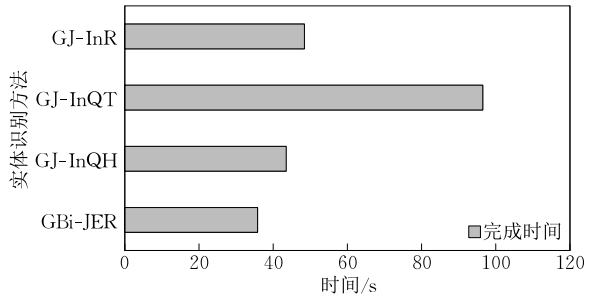

图 21 插入策略对比

7 相关工作

实体识别研究已经具有几十年的历史,吸引了数据库,数据挖掘,人工智能以及统计学等领域的诸多研究者,它有其他的名称:实体解析、实体匹配、记录匹配、记录去重、实体辨别、合并与清除等[1,3-4,20-22].传统实体识别方法基于对象属性相似度(ABS),给定两个对象及其关系模式,传统实体识别方法计算两个对象的对应属性的相似度并将其加权求和得到综合的相似度,然后将该相似度与给定阈值进行比较来决定两个对象是否匹配.传统方法解决单类实体识别问题[1,3-4].然而,现实世界里的数据一般都是存在关联的而非独立存在,比如引文数据库包括文章、作者、单位和会议,电影数据库中包括电影、演员、导演、电影工作室,电子商务数据库包括客户、商品、厂商、消费记录等,称为关联的数据.为了更准确、快速地识别关联的数据集中的实体,需要充分利用对象关系,同时进行多类实体的联合识别.

目前,已存在一些基于对象关系的实体识别方法.Dong 等人[8]通过对象关系,如作者、文章、email 等,将可能匹配的对象对作为一个结点,构建出对象对的依赖图来进行联合式实体识别.当一对对象匹配成功,通过依赖关系,将这个结果传递给它们的直接相邻的其他对象对,它们的匹配结果,增加了与它们直接相邻的对象对之间的结构相似度,形成一个迭代过程.依赖图中的边引导联合式实体识别的进行.该方法针对个人信息管理提出.Bhattacharya 等人[5-7]利用引文数据中的共同作者关系,进行基于关

系的聚类，来实现联合式实体识别，该方法只识别作者且局限于引文领域. 以上两类方法都用到了对象间直接关联关系. Kalashnikov 等人[9-11]默认只有一类实体未被识别出，利用对象关系构建对象图，没有考虑关系模式，通过分析图上的对象关系来促进实体识别，解决了同名不同实体的问题，该方法不是联合式实体识别方法. 在机器学习领域，有一些基于关系的实体识别方法[20-22]，此类方法通过学习训练数据，得到全局概率模型，由这个模型来指导实体识别. 该类方法只适合于比较少的关系和异构性不太强的数据，对于复杂结构的数据不适用，无法保证识别质量和速度. 还有一些研究者致力于提高联合式实体识别的伸缩性[23-25]，这类方法采用分布式技术或分块技术来处理大数据量的实体识别. 而本文针对识别速度的优化是通过更少的比较次数来识别出更多的匹配对，两者有明显区别.

本文提出的 GBi-JER 方法，充分发掘对象图，通过迭代地利用逐渐收敛的对象关系（包括直接的和间接的关系），提高实体识别的准确性. GBi-JER 可以更快速、准确地进行联合式实体识别，Dong，Bhattacharya，Kalashnikov 等的方法是 GBi-JER 的子集.

8　结束语

实体识别是数据清洗的重要方面，数据挖掘和数据集成都离不开它. 大数据时代，数据呈现多样性和关联性，如何对多类型的、关联的对象集进行高效、准确的联合式实体识别成为一个重要问题. 针对已有联合式实体识别方法和基于对象关系的实体识别方法的不足，本文提出一种面向关联数据的联合式实体识别方法——GBi-JER，该方法领域无关，适合于任何关联的数据. GBi-JER 充分发掘逐渐收敛的对象图，迭代地对多类型的、关联的数据进行联合实体识别，利用不同对象间的关系来促进彼此的匹配，实现了准确、高效的联合式实体识别. 下一步工作将研究增量的联合式实体识别方法.

参　考　文　献

[1] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(1): 1-16

[2] Sun Y, Han J. Mining heterogeneous information networks: A structural analysis approach. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20-28

[3] Thor A, Rahm E. MOMA—A mapping-based object matching system//Proceedings of the 3rd Biennial Conference on Innovative Data Systems Research. Asilomar, USA, 2007: 247-258

[4] Benjelloun O, Garcia-Molina H, et al. Swoosh: A generic approach to entity resolution. The International Journal on Very Large Data Bases, 2009, 18(1): 255-276

[5] Bhattacharya I, Getoor L. Iterative record linkage for cleaning and integration//Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Paris, France, 2004: 11-18

[6] Bhattacharya I, Getoor L. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 5

[7] Bhattacharya I, Getoor L. Entity resolution in graphs// Cook D, Holder L eds. Mining Graph Data. Hoboken: Wiley-Interscience, 2006: 311-342

[8] Dong X, Halevy A, Madhavan J. Reference reconciliation in complex information spaces//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore, USA, 2005: 85-96

[9] Kalashnikov D V, Mehrotra S, Chen Z. Exploiting relationships for domain-independent data cleaning//Proceedings of the 2005 SIAM International Conference on Data Mining. Newport Beach, USA, 2005: 262-273

[10] Chen Z, Kalashnikov D V, Mehrotra S. Adaptive graphical approach to entity resolution//Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. Vancouver, Canada, 2007: 204-213

[11] Nuray-Turan R, Kalashnikov D V, Mehrotra S. Adaptive connection strength models for relationship-based entity resolution. Journal of Data and Information Quality, 2013, 4(2): 8

[12] Sun Y, Han J, et al. PathSim: Meta path-based top-k similarity search in heterogeneous information networks// Proceedings of the 37th International Conference on Very Large Data Bases. Seattle, USA, 2011: 992-1003

[13] McCallum A, Nigam K, Ungar L H. Efficient clustering of high-dimensional data sets with application to reference matching//Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 2000: 169-178

[14] Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching//Proceedings of the 18th IEEE International Conference on Data Engineering. San Jose, USA, 2002: 117-128

[15] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records//Proceedings of the 9th ACM SIGKDD Workshop on Data Cleaning and Object Consolidation. Washington, USA, 2003: 73-78

[16] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031

[17] Sun Chen-Chen, Shen De-Rong, et al. WSR: A semantic relatedness measure based on wikipedia structure. Chinese Journal of Computers, 2012, 35(11): 2361-2370(in Chinese)
(孙琛琛, 申德荣等. WSR: 一种基于维基百科结构信息的语义关联度计算算法. 计算机学报, 2012, 35(11): 2361-2370)

[18] Lovász L. Random walks on graphs: A survey//Miklós D, Sós V T, Szönyi T eds. Combinatorics: Paul Erdos is Eighty. Budapest: Janos Bolyai Mathematical Society, 1993: 1-46

[19] Dong X L, Naumann F. Data fusion: Resolving data conflicts for integration//Proceedings of the 35th International Conference on Very Large Data Bases. Lyon, France, 2009: 1654-1655

[20] Singla P, Domingos P. Entity resolution with Markov logic//Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, China, 2006: 572-582

[21] McCallum A, Wellner, B. Conditional models of identity uncertainty with application to noun coreference//Proceedings of the Advances in Neural Information Processing Systems 18. Vancouver, Canada, 2005: 905-912

[22] Culotta A, McCallum A. Joint deduplication of multiple record types in relational data//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany, 2005: 257-258

[23] Rastogi V, Dalvi N, Garofalakis M. Large-scale collective entity matching//Proceedings of the 37th International Conference on Very Large Data Bases. Seattle, USA, 2011: 208-218

[24] Herschel M, Naumann F, et al. Scalable iterative graph duplicate detection. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(11): 2094-2108

[25] Böhm C, de Melo G, et al. Linda: Distributed web-of-data-scale entity matching//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 2104-2108

SUN Chen-Chen, born in 1987, Ph. D. candidate. His research interests include entity resolution and information network mining.

SHEN De-Rong, born in 1964, Ph. D., professor, Ph. D. supervisor. Her research interests include distributed data management and data integration.

KOU Yue, born in 1980, Ph. D., associate professor. Her research interests include entity search and data mining.

NIE Tie-Zheng, born in 1980, Ph. D., associate professor. His research interests include data quality and data integration.

YU Ge, born in 1962, Ph. D., professor, Ph. D. supervisor. His research interests include database and big data management.

Background

Entity Resolution (ER) is a crucial aspect of data cleaning and is important to data mining and data integration. Traditional ER approaches rely upon entities' Attribute-Based Similarity (ABS), focusing on a single class of independent entities. However, the world is interconnected that most data is related, such as citation dataset consisting of papers, authors, venues and semantic relations between them. ABS ER approaches can hardly exploit the semantic relations of related dataset. To exploit the semantic relations and jointly resolve multiple entities promotes ER accuracy. This type of ER approaches are named joint ER. There are some joint ER approaches in literature. Bhattacharya et al analysis co-occurrence relationships in citation dataset to help resolve authors. They proposed a relational clustering method for ER. Dong et al build a dependency graph of candidate entity pairs. They exploit the dependency graph to resolve multiple classes of entities. Both of them only consider direct relationships between entities so that they are limited.

We propose a joint ER model based on set partition. Then we implement the joint ER model with an entity relationships graph and propose a graph-based iterative joint ER approach—GBi-JER. The core idea is to exploit entity relationships (both direct and indirect ones) to jointly resolve multiple classes of entities in a dynamic entity graph. GBi-JER hires a hybrid similarity to match two data objects, consisting of an attribute based similarity and a semantic path based one. GBi-JER merges two matched data objects and contracts the neighborhood of them. After the subgraph contraction, the neighborhood becomes denser, which triggers similarity propagation and makes GBi-JER an iterative process. With the joint ER going on, the entity graph gets denser and denser, which positively influences later resolution. GBi-JER outperform existing joint ER approaches.

This work is supported by the National Basic Research Program (973 Program) of China under Grant (2012CB316201) and the National Natural Science Foundation of China under Grant (61472070).