异构信息空间中时间感知的查询时实体识别与数据融合

杨 丹[1] 陈 默[2] 王 刚[1] 孙良旭[1]
(辽宁科技大学软件学院 鞍山 114051)[1] (东北大学计算中心 沈阳 110004)[2]

摘 要 已有的传统的实体识别技术大多是以线下、非实时的方式,在静态数据集上进行,对于大数据集的执行通常需要大量的时间和系统资源。对于异构信息空间中具有时间信息、不断演化的异构实体来说,时间感知的查询时实体识别与数据融合越来越成为一种保证数据质量和满足用户需求的发展趋势。针对异构信息空间中使用时间上下文的关键字查询进行的实体搜索,提出一种时间感知的查询时实体识别与数据融合方法TQ-ER,以给用户提供准确的实体概貌(entity profile);提出一种迭代式时间感知的实体候选集生成算法。TQ-ER充分利用查询的时间上下文和实体的时间信息给正确的回答一个给定查询所需要的、最少的实体数据,以进行识别与数据融合。在真实数据集上的大量实验结果表明了TQ-ER的有效性和正确性。

关键词 时间感知,查询时实体识别,数据融合,异构信息空间
中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.03.045

Time-aware Query-time Entity Resolution and Data Fusion in Heterogeneous Information Spaces

YANG Dan[1] CHEN Mo[2] WANG Gang[1] SUN Liang-xu[1]
(School of Software, University of Science and Technology Liaoning, Anshan 114051, China)[1]
(Computing Center, Northeastern University, Shenyang 110004, China)[2]

Abstract Most of existing traditional entity resolution (ER) techniques mainly deal with static data sets by offline, non real-time methods. For large data sets, it usually requires a lot of time and system resources. In the face of evolved, heterogeneous entities with time information in heterogeneous information spaces, time-aware query-time ER and data fusion become a necessary trend to ensure data quality and user requirements. Aiming at entity search based on keyword query with temporal context in heterogeneous information spaces, this paper proposed a time-aware query-time ER approach TQ-ER to provide more accurate entity profiles to users. A time-aware iterative query expansion algorithm was proposed. TQ-ER leverags temporal context of query and temporal information of entities, which can identify the minimum entities to do ER and data fusion for a given query to be correctly answered. Extensive experimental results on real data sets show the effectiveness and correctness of TQ-ER.

Keywords Time-aware, Query-time entity resolution, Data fusion, Heterogeneous information spaces

传统的实体识别大多是通过线下的方式以批处理模式在静态的数据库上进行的,执行大数据量的数据集时通常需要大量的时间和系统资源。随着在线服务的增多,越来越多的机构和应用要求在进行实体搜索时执行实时的、在线的实体识别,即这些实体需要与已存在的实体数据库相匹配。例如,在线电子商务、电子图书馆、紧急事件响应系统、用户为救助突发的病人搜索医院等实时搜索(time-critical search)[1]的情况。对于用户的这类信息需求,需要实体识别满足在线、on-the-fly、实时性的要求,即查询时的实体识别。查询时实体识别的目标就是尽可能快速地在异构信息空间中识别查询中所包含的实体,响应时间要尽可能的短。例如,用户查找作者"Halevy"所著的所有论文时,系统没有必要识别异构信息空间中所有的作者实体或论文实体。异构信息空间中的时间信息无处不在,如文档的时间元信息(文档的创建时间、修改时间)、文档内容中与实体相关的时间上下文、时间记录(temporal record)的元组时间戳和属性时间戳、类型为时间型的实体属性的值(如出生日期、论文发表年代)等。用户经常会发出具有时间上下文的查询,如"Persons attended the SIGMOD 2015","Papers written by Halevy published between 2005 and 2010"。类似地,当用户想要查找Halevy在2012年发表的所有论文时,也没有必要识别该作者在所有年代发表的所有论文。

识别出的实体往往存在不一致的数据,需要对其进行数据融合以解决这些冲突和错误,从而提供给用户一个完整和

到稿日期:2015-10-11 返修日期:2015-12-31 本文受国家自然科学基金项目(61402213,61402093)资助。
杨 丹(1978-),女,博士,副教授,CCF会员,主要研究方向为数据集成、实体搜索,E-mail:asyangdan@163.com;陈 默(1983-),女,博士,讲师,CCF会员,主要研究方向为空间数据处理;王 刚(1978-),男,博士,副教授,主要研究方向为数据集成;孙良旭(1979-),男,硕士,副教授,主要研究方向为大数据。

正确的具有时间信息的实体概貌(entity profile)。例如,来自不同数据源的同一作者在不同时期的地址属性值,一个是攻读博士学位时期,另一个是博士毕业后参加工作时期,两个属性值分别在不同时间是正确的,前者是过时的值,后者是当前的最新值。来自不同的数据源、描述同一个实体的信息彼此之间可能存在重叠的信息,也可能存在冲突、甚至是错误的值。

已有的查询时实体识别相关工作没有考虑时间上下文的查询,并且对于识别出来的具有时间信息的实体也没有考虑时间感知的数据融合。针对异构信息空间中具有时间上下文的关键字查询进行的实体搜索,本文提出时间感知的查询时实体识别与数据融合方法 TQ-ER,旨在提高实体识别的时间效率,在保证查询数据质量的同时,使得其满足用户查询的实时性要求。

1 相关工作

根据本文的研究内容,本节主要从两方面介绍相关工作,1)实时、查询时实体识别;2)数据融合的相关工作。

1.1 实时、查询时实体识别

已有的实体识别技术和相关工作大多都不是实时的,而是对线下的、静态的数据集的一种预处理、清洗工作。为数不多的研究实时、查询时的实体识别的相关工作主要有:文献[2]提出一种相似感知(similarity-aware)的索引技术,其主要思想是结合用于近似匹配的相似度计算和通常 Web 搜索引擎所使用的倒排索引技术。这种索引技术便于实时实体识别的实现,但是只适用于静态的数据库。文献[3]基于标准 Blocking 技术[4]扩展了上述索引,使之可以在动态的数据上工作。文献[5-6]提出了一种基于森林的 sorted neighborhood[7]索引,它使用具有不同排序键的多个索引树为实时实体识别提供便利。上述相关工作都是研究如何在尽可能短的时间内在索引的所有记录中匹配一个查询记录所表示的实体。

较早提出查询时实体识别的文献[8]为处理查询提出一个两阶段的集合实体识别策略,这两阶段为抽取阶段和识别阶段,文献[8]还给出了一个抽取与查询最相关的实体的算法。算法的思想是将用户提供的查询关键词逐步扩展后,找出与之关联的所有表象,并确定这些表象是否都指向用户所要检索的实体,使得可以在查询时同时识别实体,同时也保留集合识别的优势。但其在实体属性扩展时没有考虑属性值是时间类型的属性。文献[9]提出 On-the-fly 实体识别,用于解决数据不确定情况下的查询。文献[10]关注在线数据的查询,提出了一种查询驱动的实体识别方法 QDA,它只执行能够用来正确回答一个用户给定查询的必要的、最少的清洗步骤。文献[11]基于迭代缓存提出了一个 online 记录链接和融合框架 ORLF,把经常被查询的记录在线下清洗并缓存为未来记录链接的参考(认为一个有质量的查询 Q 的结果不能只根据查询 Q 的返回记录给出)。文献[12]提出查询时的实体识别,动态地去掉来自不同数据源的冗余实体信息。

1.2 数据融合

数据融合[13]是一个发生在实体识别之后的过程。给定一个实体的相关元组集合 I_t,将这些元组融合成一个元组 I 并解决这些元组冲突的过程称作冲突消解。解决数据内容冲突的方法大致分为两类:1)基于关系扩展的方法,如扩展关系操作及聚合函数;2)从多个冲突值中选择真值的方法。文献[14]提出了一种基于 Markov 逻辑网的两阶段数据冲突解决方法。文献[15]针对数据的多样性,将记录链接与数据融合相结合。文献[16]在考虑数据量的基础上,提出了在线数据融合,其主要思想是一旦系统认为其余的数据源不会改变已经探测的数据源对查询的计算结果,就停止探测额外的数据源。数据时效性(data currency)[17-18]是研究识别数据源中实体的当前值,使用当前值(最新值)来回答查询,是影响数据质量的重要因素之一。文献[19]在缺失数据源准确性信息的情况下,提出了一个基于数据时效性和数据一致性的冲突消解模型。

2 查询时实体识别与数据融合

解决基于关键字的实体搜索的查询时实体识别问题最简单的方法就是只识别查询关键字所指的实体,按照相似度函数找出可能的所有实体(集)并返回给用户。这种做法存在的问题是:1)只依靠字符串或属性值相似度进行一种单一的实体类型的实体识别往往是不准确的,甚至是不能识别的。例如,如果已知出现在“SIGMOD 2015”的论文和出现在“13th SIGMOD”的论文是同一篇论文,那么这两个字符串指的是同一个会议实体,这些信息有助于其他论文实体的识别(匹配),因此需要找出与该实体具有关联关系的实体以提供更多的识别证据。2)异构信息空间中的实体大多具有时间信息,在识别时需要考虑用户查询的时间上下文以过滤掉不必要进行识别的实体,从而提高识别效率。3)由于异构信息空间中多种时间版本的实体信息共存,识别后来自异构数据源的实体仍然存在各种冲突,需要进行时间感知的数据融合,将统一、正确的结果返回给用户。

2.1 相关定义

定义 1(具有时间信息的实体) 异构信息空间中具有时间信息的实体表示为三元组:$e=\langle name, A, E_{type}\rangle @t|ts$,其中 $t|ts$ 表示该实体相关联的时间信息即时间点或时间间隔,t 表示某个时间点;ts 表示某个时间间隔,由开始时间和结束时间组成,记作$[ts_{begin}, ts_{end}]$;$name$ 是实体的名称,通常为实体主属性的属性值;E_{type} 表示实体 e 所隶属的实体类,A 表示实体属性-值集合,即 $A=\{(attribute_1, v_1), \cdots, (attribute_n, v_n)\}$。实体的时间信息记作 $e.t$ 或 $e.ts$;相应地,属性 $attr$ 的时间信息记作 $e.attr.t$ 或 $e.attr.ts$;实体 e 在某个时间点 t 或时间间隔 ts 的属性值记作 $(e.attr)_{t|ts}.value$。

例如,同一个作者实体在不同时间信息下的数据如下:〈Halevy,{(name, *Halevy*), (age, *50*), (affiliation, *Univ. of Washington*), (title, *professor*), … }〉@ 2005;〈Halevy, (name, *Halevy*), {(age, 50), (affiliation, *Google Inc.*), (title, *manager*), … }〉@[2010, 2015]。

定义 2(具有时间上下文的查询) 表示异构信息空间中具有时间信息需求的关键字查询。该类查询除了查询的关键

字部分，还包含时间部分，也称作具有时间限定语的查询。形式如下：查询 $Q=\{k_1,k_2,\cdots,k_n\}@\{t,ts\}$，由两部分组成，其中$\{k_1,k_2,\cdots,k_n\}$部分为查询 Q 的关键字（记作 Q_{text}），$@t|ts$ 部分是可选项，表示查询 Q 的时间上下文（记作 Q_{time}），t 表示某个（单个）时间点；ts 表示某个连续的时间间隔$[ts_{begin},ts_{end}]$。

例如，查询 {Xin Luna Dong, paper}@2014；{coauthor, Halevy }@[2001,2009]都是具有时间上下文的查询。

定义 3（实体类关系图 G_R）　由实体类、实体类之间的关系组成的图 $G_R=(V_S,E_S)$。其中，顶点是实体类；边表示实体类间的关系。

例如：作者实体类之间存在合作者关系：作者$\xleftarrow{\text{coauthor}}$作者；论文和会议实体类之间具有发表关系：论文$\xrightarrow{\text{publishedIn}}$会议。

定义 4（实体关联图 G_{Ass}）　由实体、实体之间具有时间信息的关联关系组成的图 $G_D=(V_D,E_D)$。其中，顶点是实体；边表示实体间的关联。

例如：Halevy $\xrightarrow{\text{supervisorOf@[2004,2009]}}$ Xin Dong 表示在 2004 年到 2009 年 Halevy 是 Xin Dong 的导师。

2.2　TQ-ER 框架概览

异构信息空间中时间感知的查询时实体识别与数据融合框架如图 1 所示，其中主要包括目标实体扩展、实体搜索、实体识别与数据融合。底层是异构信息空间中的实体类关联图和实体关联图。

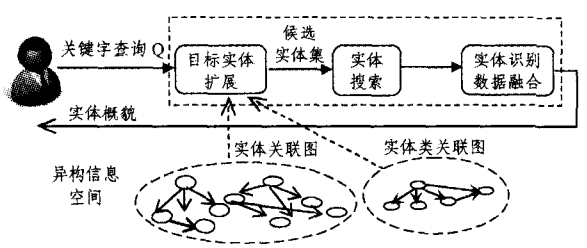


图 1　时间感知的查询时实体识别与数据融合

首先根据用户查询的关键字部分 Q_{text} 确定要查询的目标实体；然后根据实体类关联图、实体关联图和查询的时间上下文部分 Q_{time} 生成该实体的候选实体集，并在异构信息空间中进行实体搜索；最后进行实体识别，得到实体簇$\{C_1,C_2,\cdots,C_n\}$，其中每个实体簇 C_i 代表现实世界中的同一实体；进行时态的数据融合得到时间感知的实体概貌（entity profile）；最后将得到的查询时结果实体集 R 返回给用户。TQ-ER 的目标是根据用户的关键字查询的，在尽可能短的时间内提供给用户准确的、一致的、时间感知的实体信息，以满足用户的信息需求。

3　时间感知的实体识别

对于有时间上下文的查询时实体识别，以作者文献领域的实体搜索为例，观察到如下现象。

观察 1　由于录入错误或是数据源之间拷贝错误引起实体的时间信息不准确；而且有些实体的时间相关信息具有跨年的情况。例如，召开会议是在 2014 年，而论文检索时间是 2015 年，那么相邻年代的论文会出现时间上的不统一，因此需要时间轴上的扩展 X_T。

观察 2　由于异构信息空间中多时间版本信息共存的特点，时间信息可以作为重要的查询条件之一进行初步的过滤，从而缩小搜索空间以提高搜索效率。例如，查找某位作者近 5 年来发表的论文没有必要把他所有的论文都进行识别，因此提出时间轴的过滤 F_T。

观察 3　时间信息可以作为实体识别的有效证据，即判断两个属性相似度较高的实体在时间特征上是否具有相似性，这样可以有效提高实体识别的准确性。例如，两个作者发表论文的频率在时间轴上的相似性；两个作者发表论文的数量在时间轴上的相似性等。因此提出实体的时态相似性（temporal similarity）。

3.1　候选实体集生成迭代算法

给定异构信息空间中实体搜索的一个具有时间上下文的查询 $Q=\{k_1,k_2,\cdots,k_n\}@t|ts$，首先确定实体搜索的目标实体类 ec，根据查询的时间上下文进行时间轴上的扩展 $X_T(Q_{time})$；接着根据实体类关系图 G_R 进行实体类关系扩展 $X_R(e)$；然后根据搜索的目标实体 e 进行实体属性相似度扩展 $X_{Att}(e)$；再根据实体关联图 G_{Ass} 进行实体关联关系扩展 $X_{Ass}(e)$；最后通过进行时间轴上的过滤 $F_T(e)$来获得候选实体集 E'。时间感知的迭代式候选实体集生成算法 TIQE 如算法 1 所示。

算法 1　Time-aware Iterative Query Expansion(TIQE)

Input：查询 $Q=\{k_1,k_2,\cdots,k_n\}@[t \text{ or } ts]$；实体类关系图；实体关联图

Output：候选实体集 E

//Initialize 初始化

1. Query Mapping to obtain the target entity class ec
2. Temporal expands $X_T(Q_{time})$, put them into queue $Q_T=\{\}$

//Iterative 迭代

3. Entity Class Relation expands $X_R(ec)$, put them into queue $Q_R=\{\}$
4. Entity attribute expands $X_{Att}(e)$, put them into queue $Q_A=\{\}$
5. Entity Association expands X_{Ass}, put them into queue $Q_{ASS}=\{\}$
6. For each entity class $ec\in Q_R$, dequeue it, Association Expands
7. Repeat step 3−6 until Q_R is empty
8. Temporal filtering $F_T(Q_{time})$
9. Return E'

由算法 1 可知，迭代算法的终止条件是 Q_R 队列为空。关联关系扩展层数 L 表示在实体类关系图上两个节点间的路径长度。其中，属性扩展的条件为：属性相似度大于阈值 δ_{Low}。例如，查找“Xin Dong”时，可扩展为“Dong Xin”，“Xin Luna Dong”等。时间轴上扩展的原则为：根据查询的时间上下文 Q_{time}，扩展后的时间上下文 Q'_{time} 为：

$$Q'_{time}=\begin{cases}[t_1-1,t_2+1]\\ t-1,t,t+1\end{cases} \tag{1}$$

其中，1 为一个时间单位（如年、月、日），可根据不同领域的实体或数据源而有所不同。对于时间点的情况，扩展成 3 个时间点，即增加前一个时间点和后一个时间点；对于时间间隔的

情况，分别将开始时间 t_1 和结束时间 t_2 扩展 1 个时间单位。

3.2 时态相似性

TQ-ER 在对具有时间信息的实体进行识别时利用了时态相似性。将实体表象 r_i 集合的时间向量 $t(r_i)$ 以某个时间单位沿时间轴划分为多个均等的槽(slot) $t_1, t_2, \cdots, t_n$，并选定某个随时间变化的实体属性 *attr* 在不同时间 t_i 的值存入向量。两个实体表象间的时态相似性记作 SIM_T，用它们的 2 个时间向量余弦相似度表示，如式(2)所示。图 2 示出了一个作者文献领域的例子，其来自两个数据源 D_1 和 D_2 的作者表象 a_1 每年(2010 年－2014 年)发表论文篇数的统计情况，同时进行时态相似性计算。

$$SIM_T=\frac{t(r_i)\cdot t(r_j)}{\|t(r_i)\|\cdot\|t(r_j)\|} \tag{2}$$

数据源	作者	#paper	year
D_1	a_1	1	2010
		2	2011
		3	2012
		4	2013
		1	2014
D_2	a_1	2	2010
		2	2011
		2	2012
		1	2013
		1	2014

时态相似性计算 ⟹ SIM_T=0.816

图 2 时态相似性的例子

3.3 时间感知的聚类算法

时间感知的聚类算法 TC 如算法 2 所示。由于是查询时实体识别，因此算法的主要目标是在尽量不损失准确率的前提下提高速度，并对具有时间信息的实体体现感知的特点。算法主要思想为：首先视每个实体为一个簇，再计算簇与簇的相似度，若大于阈值 τ 就将两个簇合并并更新簇的时间签名。所采用的实体表象相似度 *SIM* 的计算公式如式(3)所示，由 3 部分组成：

$$SIM(r_i,r_j)=\alpha SIM_A+\beta SIM_R+\gamma SIM_T \tag{3}$$

其中，SIM_A 为属性相似度，SIM_R 为关联关系相似度，SIM_T 为时态相似度；α,β,γ 为各部分的系数，并且 $\alpha+\beta+\gamma=1$。

算法 2 Time-aware Clustering(TC)

Input: entity references $R=\{r_1,r_2,\cdots,r_{|R|}\}$, sorted in increasing temporal order//时间升序排序

threshold of SIM: τ

Output: a set of clusters $C=\{C_1,C_2,\cdots,C_{|C|}\}$

Process:

//initialization:

1. $C_0\leftarrow\{C_1,C_2,\cdots,C_{|R|}\}$
2. L=0;//L is the iterative step number

//Iterative:

3. L←L+1
4. for all two clusters $C_i\in C_{L-1}, C_j\in C_{L-1}$ do
5. compute SIM(C_i,C_j)
6. if SIM(C_i,C_j)>τ then
7. new cluster C_{ij}←Merge (C_i,C_j)
8. Update cluster's time labels C_i. early, C_i. late and C_i. ts
9. $C_L\leftarrow C_{L-1}-\{C_i,C_j\}\cup\{C_{ij}\}$
10. else SIM(C_i,C_j)←0
11. end if
12. end if
13. end for
14. return C_L

4 时间感知的数据融合

对识别出的具有时间信息的实体集即生成的簇集合C 中的每个 C_i 中的实体进行时间感知的数据融合，包括冲突消解、时态数据融合，对不能为空的缺失属性值进行基于数据时效性的推理进而补充合理且可能的属性值，生成实体概貌。

4.1 相关定义

定义 5(实体概貌) 对于实体 *e*，TQ-ER 将融合后的具有时间信息的 *e* 的实体概貌记作 ε，与 ε 相关联的时间信息记作 $\varepsilon.t$ 或 $\varepsilon.ts$。

实体概貌是经过实体识别与数据融合后丰富的、干净的实体描述。

定义 6(实体的属性值时序关系) 实体 *e* 在属性 *attr* 上的属性值 (v_1-v_m) 在时间序列 $\langle t_1,t_2,\cdots,t_m\rangle$ (其中 $t_i\prec t_{i+1}$) 上的变化，记作 $e.Z_{attr}=[(v_1,t_1),\cdots,(v_m,t_m)]$，$e.Z_{attr}(t_i)\neq e.Z_{attr}(t_{i+1})$。实体的属性值时序关系 $t_i\prec attr_{t_j}$ 表示 t_j 时刻的实体属性 *attr* 的值比 t_i 时刻的值新。

定义 7(实体属性值时效约束) 实体的属性值在时间序列上的变化要满足的时效性约束条件，记作 *CC*(*attr*) satisfy if $t_i\prec t_j$, then$(e.attr)_{t_i}.value\prec(e.attr)_{t_j}.value$。

例如，某个人的状态属性 *state* 值只能从工作状态变化到退休状态；从退休状态变化到死亡状态，但是不能从死亡状态变化到工作状态或退休状态。

4.2 数据融合与冲突消解规则

在 TQ-ER 中根据时间序列、时效约束定义时间感知的数据融合、冲突消解规则如下。

Rule 1 时态合并(Temporal Merge)规则，即对于实体的两个时间间隔 ts_i 和 ts_j，如果存在$(e.attr)_{ts_i}.value=(e.attr)_{ts_j}.value$，并且 $e.ts_i\supseteq e.ts_j$，那么 $\varepsilon.ts=e.ts_i$。如果 $e.ts_i\cap e.ts_j\neq\varnothing$，那么 $\varepsilon.ts=[ts_{ibegin},ts_{jend}]$。对于两个时间点 t_i 和 t_j，如果存在 $(e1.attr)_{t_i}.value=(e2.attr)_{t_j}.value$，并且 t_i 和 t_j 在时间序列上相邻，那么$(\varepsilon.attr).ts=[t_i,t_j]$。

Rule 2 对于属性值不为空的属性，当前属性值时间最近(recency)规则，即 $(\varepsilon.attr)_{current}.vaule=(e.attr)_{max(TQ)}.vaule$。

Rule 3 对于属性值不为空的属性，向前就近原则，即如果$(e.attr)_t.vaule=$null，$t_1\prec t\prec t_2$，并且$(e.attr)_{(t_1,t_2)}.vaule=$null，那么$(\varepsilon.attr)_t.vaule=(e.attr)_{t_1}.vaule$。

1) http://www.cs.umass.edu/~mccallum/data/cora-refs.tar.gz

Rule 4　时效约束规则，即如果 $t_1 \prec t_2$，那么 $(e.attr)_{t_1}.value \leqslant (e.attr)_{t_2}.value$。

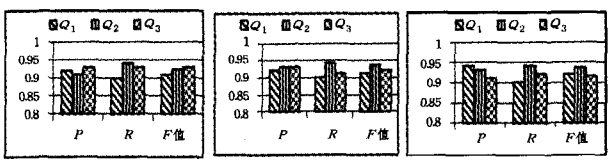
Rule 5　最保守冲突消解原则，如果 $ts_{ibegin} \prec ts_{jbegin}$，$ts_i \cap ts_j \neq \emptyset$ 并且 $(e.attr)_{ts_i}.vaule \neq (e.attr)_{ts_j}.vaule$，那么 $(e.attr)_{[ts_{ibegin}, ts_{jbegin}]}.vaule = (e.attr)_{ts_i}.vaule$，$(e.attr)_{[ts_{iend}, ts_{jend}]}.vaule = (e.attr)_{ts_j}.vaule$。

5　实验评价

选择学术文献领域的两个数据集 DBLP 和 Cora[1]，并对其进行了手工扩展（增加并补充了一些实体的属性及属性值），包括作者、论文和会议实体。在两个数据集上分别随机生成下列 3 种类型的查询集：不具有时间上下文的查询 Q_1，即不具有 Q_{time} 部分的查询；具有显式时间上下文的查询 Q_2；具有隐式时间上下文的查询 Q_3。实验所使用的机器配置为：3.5GHz 因特尔酷睿 i3 处理器、4GB 内存和 500GB 硬盘。

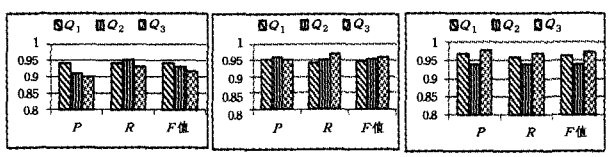
(1)实体识别性能评价

采用被广泛使用的评价标准即准确率、召回率和 F 值进行实体识别性能评价。在两个数据集上的查询时实体识别实验结果如图 3、图 4 所示。从图 3 和图 4 中可以看出，在两个数据集上，对于不同的实体类实体、不同的查询类型（Q_1 — Q_3），F 值都在 0.95 以上。



(a)作者　　(b)论文　　(c)会议

图 3　不同实体类实体的识别结果(DBLP)



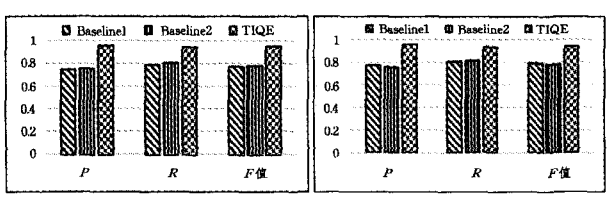
(a)作者　　(b)论文　　(c)会议

图 4　不同实体类实体的识别结果(Cora)

(2)不同候选实体集生成方法的比较

比较在如下 3 种候选实体生成方法情况下实体识别的准确性，结果如图 5 所示。从图 5 中可以看出，在两个数据集上，TIQE 的性能都明显好于两种 Baseline 方法。

- Baseline1：只进行属性扩展；
- Baseline2：只进行属性扩展和关系扩展；
- TIQE：TQ-ER 所采用的时间感知的迭代式候选实体生成方法。



(a)DBLP　　(b)Cora

图 5　不同候选实体生成方法的比较

(3)TQ-ER 的可伸缩性

在 DBLP 数据集上比较 TQ-ER 在不同数据量情况下的执行时间，实验结果如图 6 所示。从图 6 中可以看出，随着表象数量的增加，执行时间呈 sub-liner 增长，说明 TQ-ER 具有较好的可伸缩性。

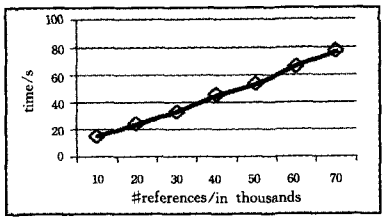


图 6　不同表象数量下的执行时间

结束语　本文提出了一种异构信息空间中时间感知的查询时实体识别与数据融合方法，其能在满足用户信息需要的实时性要求的情况下保证数据质量。在确定保证正确回答查询所需要的实体范围上，利用查询的时间上下文和实体的时间信息，提出时间感知的迭代式候选实体集生成算法来提高识别效率。大量实验结果表明了本文所提方法的有效性和正确性，其有效地提高了异构信息空间中实体搜索的效率。

参 考 文 献

[1] MISHRA N, WHITE R W, IEONG S, et al. Time-critical search [C]// SIGIR. 2014.

[2] CHRISTEN P, GAYLER R, HAWKING D. Similarity-Aware indexing for real-time entity resolution [C] // Proc of CIKM. 2009.

[3] RAMADAN B, CHRISTEN P, LIANG H, et al. Dynamic similarity-aware inverted indexing for real-time entity resolution[C]// PAKDD Workshops. 2013.

[4] CHRISTEN P. Data Matching-Concepts and techniques for record linkage, entity resolution, and duplicate detection [M]. Springer-Verlag Berlin and Heidelberg GmbH & Co. k, 2012.

[5] RAMADAN B, CHRISTEN P. Dynamic sorted neighborhood indexing for real-time entity resolution[C]// ADC. 2014.

[6] RAMADAN B, CHRISTEN P. Forest-based dynamic sorted neighborhood indexing for real-time entity resolution [C] // CIKM. 2014.

[7] HERNANDEZ M A, STOLFO S J. The merge/purge problem for large databases[C]// SIGMOD. 1995.

[8] BHATTACHARYA I, GETOOR L, LICAMELE L. query-time entity resolution[C]// KDD. 2006: 529-534.

[9] IOANNOU E, NEJDL W, NIEDEREE C, et al. On-the-fly entity-aware query processing in the presence of linkage [C] // VLDB. 2010: 429-438.

[10] ALTWAIJRY H, KALASHNIKOV D V, MEHROTRA S. Query-driven approach to entity resolution[C]// VLDB. 2013.

[11] REZIG E K, DRAGUT E C, QUZZANI M, et al. Query-time record linkage and fusion over web databases[C]// International Conference on Data Engineering. Seoul, South Korea,

（下转第 236 页）

logy,2012,41(2):163-175. (in Chinese)
朱郁筱,吕林媛. 推荐系统评价指标综述[J]. 电子科技大学学报,2012,41(2):163-175.

[8] TEJEDA-LORENTE Á, PORCEL C, PEIS E, et al. A quality based recommender system to disseminate information in a university digital library[J]. Information Sciences, 2014, 261(5): 52-69.

[9] ANBARASU V, LINDA X, MAHALAKSHMI S. An Efficient Recommender System based on Collaborative Filtering[J]. Asian Journal of Applied Sciences, 2015, 3(1).

[10] MIU P. Research on Weibo user interest model based on information push technology [D]. Wuhan: Wuhan University of Technology, 2012. (in Chinese)
谬平. 基于微博用户兴趣模型的信息推送技术的研究[D]. 武汉:武汉理工大学,2012.

[11] MONTUSCHI P, LAMBERTI F, GATTESCHI V, et al. A Semantic Recommender System for Adaptive Learning[J]. It Professional, 2015, 17(5): 50-58.

[12] MEHDI M, BOUGUILA N, BENTAHAR J. Probabilistic approach for QoS-aware recommender system for trustworthy web service selection[J]. Applied Intelligence, 2014, 41(2): 503-524.

[13] YU Z W, LI L, WONG H S, et al. Probabilistic cluster structure ensemble[J]. Information Sciences, 2014, 267(5): 16-34.

[14] SIHME A, LEHOCINEM B, MINIAIH A. Batch Adsorption of Phenol From Industrial Waste Water Using Cereal By-Products As A New Adsrbent[J]. Energy Procedia, 2012, 18(4): 1135-1144.

[15] CAO Y M. Research and implementation of personalized news recommendation algorithm based on collaborative filtering [D]. Beijing: Beijing University of Posts and Telecommunications 2013. (in Chinese)
曹一鸣. 基于协同过滤的个性化新闻推荐算法的研究与实现[D]. 北京:北京邮电大学 2013.

[16] ALPHY A, PRABAKRAN S. A Dynamic Recommender System for Improved Web Usage Mining and CRM Using Swarm Intelligence[J]. Scientific World Journal, 2015, 2015: 1-16.

[17] XIA P Y. Collaborative filtering algorithm in personalized recommendation technology [D]. Qingdao: Ocean University of China, 2011. (in Chinese)
夏培勇. 个性化推荐技术中的协同过滤算法研究[D]. 青岛:中国海洋大学,2011.

[18] RAHMAN R M, SIDDIQUEE M R, HAIDER N. Movie Recommendation System Based on Fuzzy Inference System and Adaptive Neuro Fuzzy Inference System[J]. International Journal of Fuzzy System Applications, 2015, 4(4): 31-69.

[19] CHEN L, CHEN G, WANG F. Recommender systems based on user reviews: the state of the art[J]. User Modeling and User-Adapted Interaction, 2015, 25(2): 99-154.

[20] WANG L C, MENG X W, ZHANG Y J. Chinese Journal of context aware recommendation system[J]. Software, 2012(1): 1-20. (in Chinese)
王立才,孟祥武,张玉洁. 上下文感知推荐系统[J]. 软件学报,2012(1):1-20.

[21] HUANG Q. Personalized recommendation algorithm of network book resources [D]. Chengdu: Southwest Jiaotong University, 2014. (in Chinese)
黄琼,网络图书资源个性化推荐算法研究[D]. 成都:西南交通大学,2014.

[22] DING B Z. Collaborative filtering algorithm based on citation information[D]. Jilin: Jilin University, 2014. (in Chinese)
丁彬钊. 基于引文信息的协同过滤算法研究[D]. 吉林:吉林大学,2014.

[23] CHEN N Y. Design and implementation of recommendation system based on personalized recommendation engine[D]. Guangzhou: South China University of Technology, 2012. (in Chinese)
陈诺言. 基于个性化推荐引擎组合的推荐系统的设计与实现[D]. 广州:华南理工大学,2012.

(上接第 219 页)

[12] HERZIG D M, MIKA P, BLANCOR, et al. Federated entity search using on-the-fly consolidation[M]// The Semantic Web—ISWC 2013. 2013: 167-183.

[13] DONG X, NAUMANN F. Data fusion-resolving data confiicts for integration[C]// VLDB. 2009.

[14] ZHANG Y X, LI Q Z, PENG Z H. 2-Stage Data Conflict Resolution Based on Markov Logic Networks[J]. Chinese Journal of Computers, 2012, 35(1): 101-111. (in Chinese)
张永新,李庆忠,彭朝晖. 基于 Markov 逻辑网的两阶段数据冲突解决方法[J]. 计算机学报,2012,35(1):101-111.

[15] GUO S, DONG X, SRIVASTAVA D, et al. Record linkage with uniqueness constraints and erroneous values[J]. PVLDB, 2010, 3(1): 417-428.

[16] LIU X, DONG X L, OOI B C, et al. Online data fusion[J]. PVLDB, 2011, 4(12): 932-943.

[17] FAN W F, GEERTS F, TANG N, et al. Inferring data currency and consistency for conflict resolution[C]// ICDE. 2013: 470-481.

[18] LI M H, LI J Z, GAO H. Evaluation of Data Currency[J]. Chinese Journal of Computers, 2012, 35(1): 2348-2360. (in Chinese)
李默涵,李建中,高宏. 数据时效性判定问题的求解算法[J]. 计算机学报,2012,35(11):2348-2360.

[19] FAN W F, GEERTS F, et al. Determining the currency of data [C]// PODS. 2011.