基于上下文感知实体排序的缺失数据修复方法

陈肇强 李佳俊 蒋 川 刘海龙 陈 群 李战怀

(西北工业大学计算机学院 西安 710072)

摘 要 大数据环境下,数据缺失现象十分普遍,导致许多基于数据的决策出现偏差.传统的数据库缺失值修复方法主要是利用本地数据库来修复数值型数据,这些方法并不适用于利用互联网数据来修复数值型和非数值型数据.基于互联网的缺失值修复过程一般包括生成查询、检索文档集、抽取实体、实体排序4个步骤,其中候选实体的排序决定了最终用于修复数据库的信息.现有的利用互联网数据来修复缺失数据的研究主要集中在两个方面:一是提升查询和抽取的质量,然后对抽取的候选实体按频率进行排序;另一种是分析目标实体应该具有的特征,然后对候选实体计算特征值,最后用权值叠加进行排序.这两类方法都只是考虑了实体自身的因素,而忽略了实体之间的影响.文中针对候选实体的排序建立了图模型,基于该图模型提出了上下文相关的实体排序算法CER(Context-aware Entity Ranking),该算法能够把候选实体在网页中的上下文特征充分利用起来并用实体间的影响来推断新信息,从而得到更准确的排序结果.基于真实数据集的实验结果表明,相较于频率统计和权值叠加的实体排序算法,CER算法能利用互联网的海量数据对关系数据库中的缺失值进行更加有效的修复.

关键词 数据库缺失值修复;互联网;图;实体排序

中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2015.01755

A Context-Aware Entity Ranking Method for Web-Based Data Imputation

CHEN Zhao-Qiang LI Jia-Jun JIANG Chuan LIU Hai-Long CHEN Qun LI Zhan-Huai

(*School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an* 710072)

Abstract In Big Data era, data missing is very common in real life and it puzzles people since it makes decisions based on data unreliable. Most existing data imputation methods employ local database to repair missing numerical values, while these methods do not fit the case that repair missing numerical and non-numerical values using data from web. Web-based data imputation usually contains four steps, formulating queries, searching, entity extraction and entity ranking. During these steps, entity ranking plays a key role and makes the final decision on repairing. Recently works on web-based data imputation are major in two aspects, one makes efforts to improve query formulating and entity extracting, then uses frequency to rank, the other one makes efforts to analyze features that belong to target entities, then calculates and combines features' values to rank. Frequency-based or weighting-based entity ranking method considers factors related to entity itself while ignoring the influence between entities. In this paper, we propose a graph-based entity ranking method called CER(Context-aware Entity Ranking), it can take advantage of the context of candidate entities and make a comprehensive ranking utilizing the

收稿日期:2014-08-15;最终修改稿收到日期:2015-03-05.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316203)、国家自然科学基金(61332006,61472321)、西北工业大学基础研究基金(3102014JSJ0013,3102014JSJ0005)资助.**陈肇强**,男,1988年生,博士研究生,主要研究方向为数据质量管理.E-mail:chenzhaoqiang@mail.nwpu.edu.cn.**李佳俊**,男,1990年生,硕士研究生,主要研究方向为数据质量管理.**蒋 川**,女,1992年生,硕士研究生,主要研究方向为数据质量管理.**刘海龙**,男,1980年生,博士,讲师,主要研究方向为数据管理技术、网络软件技术.**陈 群**,男,1976年生,博士,教授,博士生导师,主要研究领域为数据管理、数据质量管理、信息检索.**李战怀**,男,1961年生,博士,教授,博士生导师,主要研究领域为数据库理论与技术.

graph model. Experiments based on real-world data collections demonstrate that CER performs a more effective data imputation utilizing massive web data than the existing entity ranking methods such as frequency-based and weighting-based.

Keywords data imputation; Web; graph; entity ranking

1 引 言

数据库系统是作为商业数据计算机化管理的早期方法而产生的,当用户向数据库系统提交一个查询时,希望数据库系统能给一个完整的答案.这就要求数据集中包含足够的数据来回答各种查询和支持各种计算[1].不幸的是,现实数据库的数据经常是不完整的,存在元组或者属性值的缺失[2].对于存在缺失值的数据,可以采用的处理方法有[3]丢弃具有缺失值的记录、进行缺失值的填补、采用模型对缺失值进行预测、直接使用.在数据质量管理过程中,我们更希望的是对缺失数据进行填补,因为缺失的数据不仅仅与用户的使用和数据分析有关,它还关系到数据的一致性、精确性、时效性和实体同一性的数据质量管理与修复.幸运的是,互联网的快速发展为我们带来了巨大的资料库,与此同时,也带来了巨大的挑战,即如何从海量数据中筛选出正确的实体来修复数据库中的缺失值.

利用互联网数据进行数据库缺失值修复的主要步骤有形成有效查询、利用搜索引擎检索相关网页并从中抽取实体、最后对候选实体进行排序并修复数据库.候选实体的排序是一个很具挑战的过程,其直接影响着缺失值修复效果,如:(1)如何把互联网本身所带来的噪声(不准确数据、错误数据、多样性的表达方式等)对候选实体的干扰降至最低;(2)如何准确评估候选实体与待修复元组的相关度;(3)如何判断候选实体的真实度与可信度;(4)如何充分利用网页片段所提供的信息来筛选候选实体.传统的排序算法中,简单的只利用频率[4],更精细的会考虑各种特征函数[5]来对候选实体进行评价,比如候选实体的位置、候选实体的频率、候选实体所在网页片段的网页排名、候选实体所在语句的结构信息等.它们的特点是只专注于单个网页片段,而没有考虑各个网页片段中不同实体间的关系和相互影响.我们认为,必须想办法把相关网页的信息进行综合考虑才能得到更合理的排序.

在表1的例子中,要寻找 Red Shift 的作者,两个候选实体 Alan Garner 和 A.E. Van Vogt 无法通过频率或者权值叠加的方法区别开来.为了解决这个问题,我们提出一种基于图的候选实体排序算法,用类似于主题相关的 PageRank[6]的思想,把候选实体的各个特征以及各网页片段中的实体间的关系都用图表达出来,然后进行迭代计算,最后得出候选答案的排序.在表1中,通过 fantasy novel 这个实体可以把网页片段1和网页片段2的信息综合起来,从而推断出 Alan Garner 是 Red Shift 的作者的结论.我们在两个数据集上进行了实验,结果表明,相比较于频率统计和权值叠加的实体排序算法,CER 算法能够提高缺失值填充的准确率.本文的主要贡献如下:

(1)针对利用互联网数据进行数据库缺失值修复问题进行了详细的分析,在此基础上,用图模型对实体间的关系进行精细建模.

(2)基于图模型提出了上下文相关的实体排序算法 CER(Context-aware Entity Ranking),该算法能够把候选实体的上下文特征充分利用起来,以此降低互联网数据的干扰并提升排序的可信度,进而提高了缺失值修复的准确率.

(3)CER 排序算法属于无监督的排序算法且具有较好的通用性,即不针对特定的数据集.基于真实数据集的实验证明了该算法的有效性.

表1 候选实体排序的例子

查询	"Red Shift" "author"
网页片段1	"…Red Shift is a fantasy novel…"
网页片段2	"…Alan Garner is an English novelist best known for his children's fantasy novels…"
网页片段3	"…A.E. Van Vogt was… science fiction author…"
所有实体	Red Shift、Alan Garner、A.E. Van Vogt、fantasy novel、science fiction
候选实体	Alan Garner[频率:1]、A.E. Van Vogt[频率:1]
目标实体	Alan Garner

本文第2节回顾相关研究工作;第3节描绘基于互联网的数据库缺失值修复框架;第4节详细描述 CER 模型和算法;第5节通过在真实数据集上的实验证明 CER 的有效性;第6节对全文进行总结和展望.

2　相关工作

对于关系数据库的属性缺失问题，传统的修复方法有两个特点：(1) 它们的修复过程仅依赖于本地数据库的数据，比如均值填充、基于分类的方法、贝叶斯估计等现有的填充算法[7]；(2) 修复的对象主要是数值型数据，这种缺失值填补主要是为了防止数据分析时，由于相当部分的值空缺导致的分析偏差. 对于填补的单个数据，只具有统计意义，不具有个体意义[8]，如文献[9]对于传感器网络中的数值型数据提出了基于感知数据时间和空间相关性的缺失值估计算法；文献[3]的基于不完备数据聚类的缺失数据填补方法的实验数据集也是具有二态属性(值域为{0,1})和数值属性. 这些传统的修复方法并不适用于利用互联网数据来修复数值型和非数值型的缺失数据. 基于互联网数据的填充[4]，它把整个网络上的数据看作一个更大的数据库，并试图从这个更大的数据库中寻找约束进行缺失值的填充，如文献[10]利用互联网数据来对商品数据集中的缺失属性和缺失属性值进行修复. 基于互联网的数据填充一般包括生成查询、检索、抽取、排序.

生成查询，简单的方法是把缺失值所在列的其他所有已知属性作为查询的关键词；较为复杂的是分析已知属性的特性并通过训练的方法得到较优的查询[11]；更进一步的在自动问答系统研究领域中，对查询关键字进行同义词或时态扩展等.

形成查询后利用搜索引擎检索互联网上的相关文档，典型的搜索引擎有 Google、Bing、Yahoo、Indri 等. 它们能够返回与查询相关的网页片段，并按照一定的次序排名. 这些网页片段将成为利用网络数据进行缺失值填充的依据.

搜索引擎返回的数据特点不同，形式多样[12]，进行文本分析十分必要，相关技术有利用信息的特征来从异构的文本数据中抽取结构化的信息[13]，命名实体识别[14-15]，词性标注[16]. 典型的实体抽取技术有命名实体识别算法(NER)和基于模式(Pattern-based)的算法. 前者识别出网页片段中与查询的目标类型相匹配的命名实体作为候选实体进行排序，但由于 NER 受制于目标类型的可识别性及 NER 命名实体的训练库，如果候选实体类型无法识别或 NER 中没有某个候选实体的记录，则抽取失败；后者利用学习[17]或者人工指定的目标实体的模式进行实体抽取，但并不是所有目标实体都以一定模式呈现出来，对目标实体要求过于严格会使得很多候选实体无法被成功抽取.

现有的实体排序算法可以分为两类，基于频率排序和基于权值叠加排序. 对于基于频率排序的方法，出发点在于检索的网页片段一般都围绕着一定的主题进行阐述，尽管实体表达方式不一致，但相关统计数据表明，正确的实体很多情况下具有较高的频率[4]. 这种方法的关键之处在于要提高目标实体的频率和识别率，这就要求提升查询和抽取的质量，如文献[18]利用实体所在的片段来构建因子图，通过相似性的关系来提升抽取质量，然后利用候选实体的频率进行排序即可达到一定的正确率. 关于查询和抽取的其他相关工作前面已述，这里不再赘述. 对于基于权值叠加排序的方法，通常用到目标实体所具有的各种特征[5,19]，如频率、关键词之间的距离、关键词之间的语义相关度、网页返回的排名等. 文献[5]用加权的方法来综合这些不同的特征；文献[20]用支持向量机来叠加实体的特征，并引入了基于语法分析方法的语法关键路径的新特征；文献[19]则是将全部特征融入统一的概率框架来进行排序；文献[21]利用马尔可夫随机场来计算实体的特征值，然后叠加概率值来进行排序；文献[22]通过对检索的主题难易度预测来提升实体排序的有效性，这是对实体特征维度的一个扩展. 此外，还有一些特征维度用于特定的应用场景和数据集，如基于用户点击日志的实体排序[23]和基于用户观点的实体排序[24]. 以上的实体排序算法的共同之处在于：排序过程依据的是每个候选实体各自的评估值而忽略了实体之间的相互影响. 实体之间的影响的意义在于：候选实体特征的获取和计算存在不准确、不精确，而通过传导实体间的影响，以实体选举的方式可以提升实体排序的可靠性. 本文针对这点提出了基于图模型的上下文相关实体排序算法.

3　基于互联网的数据库缺失值修复框架

本节阐述了利用互联网数据来修复含有缺失值的关系数据库的框架，并通过分析修复过程得到结论：缺失值修复需要一个有效的排序算法.

3.1　问题描述

对于一个含有缺失值的数据库 D，即需要修复的目标数据库. 令 t_i 为 D 中某条含有部分缺失属性值的元组，假设该元组的属性值全集为 $U_{attr}=$

$\{a_0, a_1, a_2, \cdots, a_k\}$，缺失属性集为 $I_{attr}=\{a_j \mid a_j=\text{null}, 0\leq j\leq k\}$，$I_{attr}\subseteq U_{attr}$，则完整属性集为 $C_{attr}=U_{attr}-I_{attr}$. 我们的任务是根据 C_{attr} 和给定的文档集(在这里的场景中即是来自互联网的网页片段 Snippet 集)，找出合适的实体来修复缺失属性集 I_{attr}，使得修复后，对于 U_{attr} 中的每个属性值均不为空，即 $a_j\neq\text{null}$, $0\leq j\leq k$.

3.2 缺失值修复框架

基于互联网的缺失值修复工作一般包括以下 5 个步骤，其中前 4 个步骤属于主要步骤，第 5 个步骤在实际的修复过程中根据具体情况选择.

(1) 根据元组 t_i 的完整属性集 C_{attr} 构造合适的查询 $Q=\{q_0, q_1, q_2, \cdots, q_m\}$.

(2) 将查询 Q 提交到搜索引擎中，搜索引擎从互联网中检索相关文档，返回网页片段集 $S=\{s_0, s_1, s_2, \cdots, s_n\}$.

(3) 利用信息抽取技术从 S 中抽取候选的实体 $C_e=\{e_0, e_1, e_2, \cdots, e_l\}$.

(4) 运用排序算法对 C_e 进行排序，选择合适的实体，即目标实体对缺失的属性值进行修复.

(5) 对修复后的数据库进行有效性的验证，通过验证后修复完成，否则修复失败，重新修复.

具体的框架图如图 1 所示.

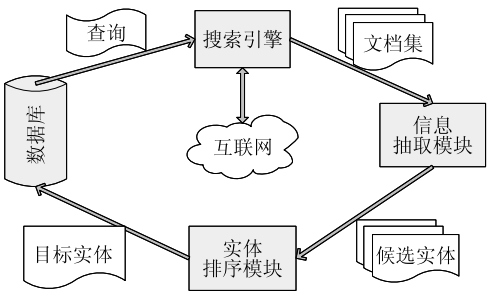


图 1 基于互联网的数据库缺失值修复框架

在这个修复过程中，每一个步骤都对最终的缺失值修复有着重要的影响. 对于第(1)步，如果生成的查询约束太大，那么返回的结果集数量相对较少，因目标实体不在结果集中而导致修复失败的风险也就相应增加；如果生成的查询比较宽泛，那么由于返回的结果集中包含大量无关的信息，进而导致寻找出用于修复的实体的难度增大. 对于第(2)步，搜索引擎是否能返回与查询相关的文档集直接影响到后续的实体抽取. 对于第(3)步，信息抽取技术决定了候选实体集，一个不好的候选实体集是不能产生出好的目标实体，这是显而易见的. 正是因为前面的 3 步中，每一步都会产生对缺失值修复不利的影响，这就给第(4)步的实体排序带来了挑战，即实体排序并不是一个平凡的过程，它是整个修复过程的最后一道关卡，决定着最终的目标实体.

因此，我们提出了一个实体排序算法 CER，它能够充分利用候选实体的上下文信息来降低(1)～(3)步的不利影响，并最终给出具有说服力和权威性的目标实体，极大地提升了缺失值修复的效果，下一节我们将详细说明 CER 模型和算法.

4 CER 模型和算法

本节首先针对候选实体的特征与互联网数据的特性构建了图模型. 图模型中的每个结点代表着一个实体，根据实体的作用分为关键字实体、中间实体和候选实体三类. 结点间的边代表着实体间的关系. 边的权重则表示实体间不同的影响，权重计算的因素包括实体间的距离、网页排名、关键字的影响因子. 然后在图模型的基础上提出了排序算法，该算法的核心思想类似于 PageRank，均用到了随机游走模型. 排序的开始阶段只有关键字实体具有初值并具有表征以较大概率从关键字开始的 D 值. 之后通过多轮随机游走过程，图中各个实体的值迭代收敛至稳定的值. 最后根据结点值给出排序结果，选择合适的实体进行缺失值修复.

4.1 模型描述

CER 算法是基于图的实体排序算法，这里的图是一个无向图 $G(V, E)$，V 是图的结点的集合，E 是图的边集，其中 E 是 $V\times V$ 的一个子集. 图中的每一个结点代表着一个实体，结点的值则代表着实体成为目标实体的程度，这里的值只具有相对意义，不具有具体的实际意义，即该值用于实体的排序. 图中的边是带权无向边，边权代表着两个实体间的关联度.

假设已经有根据查询 Q 从互联网中返回网页片段集 S，利用斯坦福大学的 NER[①](Named Entity Recognizer)技术从 S 中抽取所有的实体，构成一个集合 U_e，候选实体集 C_e 由实体类型与缺失属性值的类型相同的实体构成. 缺失属性值的类型的获得有两种方法：一是由在这个属性上值不为空的其他元组提供目标实体的样例，然后用 NER 判断候选实体的类型；二是人工来指定候选实体的类型. 关键字实体集 K_e 则是从查询中抽取出来的实体，一般地，K_e 为元组中的非空属性值集合的子集. 中间实体集

① http://nlp.stanford.edu/software/CRF-NER.shtml

E_e则是由实体类型与关键字实体集中实体类型相同的实体构成，且 $E_e \subseteq (U_e - K_e - C_e)$. 图模型中的结点由关键字实体集 K_e、中间实体集 E_e和候选实体集 C_e构成.

图模型中每个结点都对应有一个值，记第 j 个结点的值为 $VertexScore[j]$. CER 算法运行前，只有关键字实体结点的初值不为 0，其他结点初值为 0，即图中只有 $|K_e|$ 个结点具有初值，详细的初值计算见 4.2 小节. 关键字结点除了具有初值之外，每轮的随机游走过程还会加入一个 D 值的影响因子，表征从关键字开始的概率较大.

图模型中当且仅当：(1) 两个结点所代表的实体不能同时是关键字实体；(2) 至少有一个网页片段 s_j 包含有这两个结点所对应的实体；(3) 在 s_j 中，这两个实体之间不存在其他任何的 K_e、E_e、C_e 实体两个结点间存在边. 实体之间的边是无向边，因为在网页片段中，实体的前后位置并不能表征实体的指向关系. 每条边都对应有一个边权，代表了这条边所连接的两个实体之间的关联度，边权的计算为下面的式(1).

$$EdgeWeight_{vi \leftrightarrow vj} = \sum_{k=0}^{|S|} \left(\prod_{z=0}^{Z1} rel_{z,s_k}(vi, vj) \right) \quad (1)$$

式(1)中 $|S|$ 表示所有网页片段的数量，$Z1$ 表示所选取的衡量两个实体关联度的维数，$rel(vi,vj)$则是某个维度的关联度计算，具体的关联度特征的选取和计算参见 4.4 节.

至此，CER 算法的整个图模型已经构建起来，其示意图如图 2.

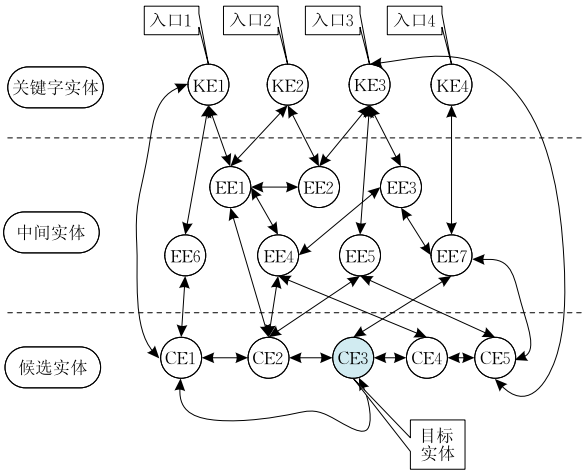


图 2　CER 算法的图模型

4.2　结点初值的计算

对于图模型中的结点，只有关键字实体集 K_e 具有初值，其他实体的初值均为 0. 当关键字实体及其同义词在网页片段集中出现的次数越少，则其初值越大. 具体的初值计算见式(2)：

$$VertexScore[j] = \begin{cases} \log_e\left[1 + |S| \Big/ \left(\sum_{i=0}^{|S|} contain(s_i, Vertex[j])\right)\right], \\ \qquad 当\ Vertex[j] \in K_e \\ 0, \qquad 当\ Vertex[j] \notin K_e \end{cases} \quad (2)$$

其中函数 $contain(s_i, Vertex[j])$的定义如式(3)：

$$contain(s_i, Vertex[j]) = \begin{cases} 1, & 当网页片段\ s_i 包含代表顶点 vertex[j] 的实体 \\ 0, & 其他 \end{cases} \quad (3)$$

4.3　入口 D 值的计算

图模型中入口 D 值表征了每轮随机游走过程从某个结点开始的概率. 这里，我们希望每次都从关键字实体开始传递结点的值，因此，只有关键字实体的结点具有不为 0 的 D 值，其他结点的 D 值为 0.

$$D[j] = \begin{cases} VertexScore[j] \Big/ \left(\sum_{i=0}^{|K_e|} VertexScore[i]\right), \\ \qquad 当\ Vertex[j] \in K_e \\ 0, \qquad 当\ Vertex[j] \notin K_e \end{cases} \quad (4)$$

4.4　边权特征的选取

为了准确刻画两个实体的关系，在计算式(1)中的 $rel_{z,s_k}(vi,vj)$时，本文采用了以下 3 个特征(以某个网页片段 s_k 为例)：

4.4.1　实体间的距离

实体间的距离这个特征表示了两个实体间影响的大小，距离越小，影响越大，反之，影响越小.

$$rel_0(vi,vj) = 1 - \min(dist(vi,vj))/length(s_k) \quad (5)$$

式(5)中，$dist(vi,vj)$指的是两个实体之间包含有的实体个数，如果在网页片段 s_k 中存在有多个 vi 和 vj 所代表的实体，则取最小的 $dist(vi,vj)$作为实体间的距离. $length(s_k)$指的是网页片段 s_k 总共的实体数. 在经典的计算实体间距离时，最常采用的是单词距离，而这里采用实体数来计算的优势在于：用现实的视角来对待实体，排除实体的不同单词个数所带来的影响，即与实体名字的长短无关. 例如，查询“Goodbye to the Buttermilk Sky”这本书的作者是谁时，返回的某个网页片段如图 3.

图 3　网页片段 s_k 样例

把这个网页片段简化为“Goodbye to the Buttermilk Sky [Deep South Books] (Julia Oliver) on Amazon.com”. 在计算实体 vi(Goodbye to the Buttermilk Sky)和实体 vj(Julia Oliver)之间的距离时,如果用单词距离来计算时,

$$rel_0=1-7/13=0.462.$$

如果用实体间距离的实体数且把网页片段也处理成由一个个实体组成,如图 3 中框所示,那么,

$$rel_0=1-1/4=0.75.$$

如果网页片段变成“Goodbye to the Buttermilk Sky [Deep South] (Julia Oliver), Amazon.com”,那么用单词距离计算的值为 0.455,而用实体距离计算的值为 0.75,保持不变. 用实体数来衡量实体间的距离不受实体的名字长度、连接词等无意义词汇的干扰,更为稳定合理.

注意到,根据 4.1 节中边存在的条件,如果两个实体之间有其他实体,那么边就不存在了,因此实体之间的距离只能是 0. 在这种情况下,直接按式(5)计算的话,这个因素就不起任何作用. 因此,这里的实体间距离计算我们修正为,式(5)中 $\min(dist(vi,vj))$ 恒等于 1,其他参数不变. 修正后的实体间距离这个因子的含义为,如果一个网页片段包含有大量的实体,而某两个实体紧挨着出现,那么它们的值较大;而如果网页片段包含的实体较少,那么实体紧挨的概率本身就较大,它们之间的值就小. 比如,极端情况下,从某个网页片段按序仅抽取出两个实体,除去非实体的字符串,那么它们必定是紧挨着的.

4.4.2 网页权值

网页权值表明了由网页片段 s_k 所决定的结点 vi 和 vj 的关联度具有多大的可信度. 网页排名越靠前,则由它定义的实体关系的可信度越高.

$$rel_1(vi,vj)=1-rank(s_k)/|S| \quad (6)$$

式(6)中,$rank(s_k)$ 表示网页片段 s_k 在返回的所有网页片段中的名次,这个值是由搜索引擎所决定的. $|S|$ 则表示所有网页片段的个数.

4.4.3 关键字的影响因子

关键字的影响因子表征了网页片段 s_k 所蕴含的实体关系的权威度,网页片段含有的关键字越多,它所提供的实体间的关系越重要,实体间的边权也就越大.

$$rel_2(vi,vj)=\left(\sum_{l=0}^{|K_e|}contain(s_k,K_e[l])\right)/|K_e| \quad (7)$$

式(7)中,$contain(s_k,K_e[l])$ 由式(3)所定义,表示关键字实体 $K_e[l]$ 是否在网页片段 s_k 中. $|K_e|$ 则是所有关键字实体的个数.

例如,书本“Cats at Work”的作者是“Rhonda Gray”,现在数据库中“Cats at Work”的作者缺失了,我们需要修复它. 输入的查询关键字为“Cats at Work+1991+Abbeville Press”,分别表示书名,出版年份,出版社. 现在从搜索引擎中只返回了两个名次一样的网页片段 s_1、s_2,其中 s_1 为“…In 1991, Cats at Work is written by Rhonda Gray … Abbeville Press…”,s_2 为“… Presidential Pets is written by Niall Kelly…Abbeville Press…”. 那么,在没有加入关键字影响因子的情况下构建的图如图 4,注意到,关键字“1991”与其他实体是没有边存在的,因为“1991”与其他实体间隔有一个关键字实体“Cats at Work”,根据 4.1 节的模型描述,“1991”与其他实体将没有边,而关键字实体间也是没有边存在的. 图中线段的粗细表示了边权的大小,在这种情况下,Rhonda Gray 和 Niall Kelly 的实体影响是一样的,那么最终 Rhonda Gray 和 Niall Kelly 的值将会是相等的,即无法有效筛选出 Rhonda Gray.

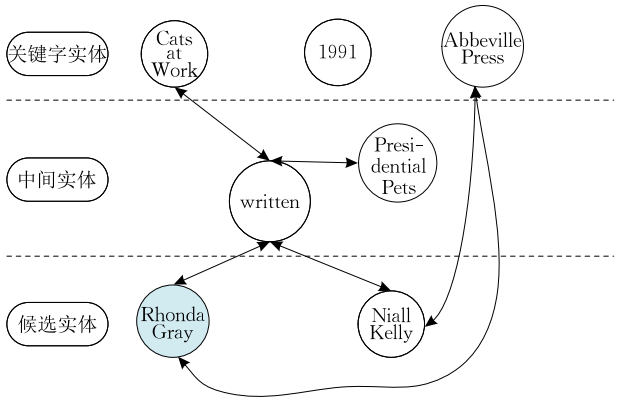


图 4 无关键字的影响因子构建的图

在加入了关键字的影响因子后,实体“written”与“Rhonda Gray”的边权将会增大,见图 5. 因为网页片段 s_1 中包含有关键字实体“1991”,这是网页片段 s_2 中所没有的. 因此,“Rhonda Gray”最终的值将会比“Niall Kelly”大.

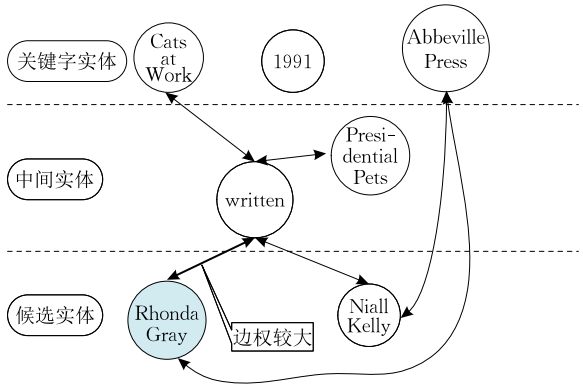


图 5 有关键字的影响因子构建的图

4.5 算法描述

CER 算法的运行过程类似于 PageRank，通过多轮迭代计算后获得每个结点的值，以此来获得候选实体的排名. CER 算法与 PageRank 的不同之处在于：(1) 我们指定了随机游走的入口，且从某个入口开始的概率不同，具体来说，如果某个关键字实体比较重要且有区分度，那么从它开始的概率会比较高；(2) 实体之间的边为双向，而这在网页的链接关系中并不常见；(3) 我们的边权的计算更为精细，使得随机游走过程更符合实体排序的需要. 每个结点的值的计算见式(8).

$$VertexScore[j]=(1-\alpha)\times D[j]+\alpha\times\sum_{i=0}^{In(j)}Value(i\rightarrow j) \tag{8}$$

式(8)中，α 为取值在 0 到 1 之间的阻尼因子，这里我们取 α 为 0.85. $D[j]$则代表了从结点 j 为入口的概率，在 CER 算法图模型中，只有关键字实体的 D 值不为 0，其余实体的 D 值均为 0. $In(j)$表示与结点 j 相连接且指向 j 的所有结点的个数. $Value(i\rightarrow j)$表示由某个指向结点 j 的结点 i 通过边$\langle i,j\rangle$传递给 j 的值，其计算见式(9).

$$Value(i\rightarrow j)=\frac{EdgeWeight(i\rightarrow j)}{\sum_{Out(i)}EdgeWeight}\times VertexScore(i) \tag{9}$$

式(9)中，$Out(i)$表示结点 i 的所有出度的结点个数.

算法 1.　CER 算法.

输入：关键字实体集 K_e，中间实体集 E_e，候选实体集 C_e，网页片段集 S

输出：目标实体集 T_e

1. FOR(每个结点 i)，初始化结点的值 $VertexScore[i]$；

2. FOR(每个结点 i)，初始化结点的入口 D 值 $D[i]$；

3. FOR(每条有向边$\langle i,j\rangle$)，初始化边权 $EdgeWeight[i][j]$；

4. 初始化迭代次数 $Iteration=0$；

5. 设定结点值的变化量 λ 为 0，复制结点的值至数组 $lastRoundVertexScore$；

6. 根据式(8)计算每个结点的值，更新数组 $VertexScore$；

7. 计算结点值的变化量，FOR(每个结点 i)，$\lambda+=|VertexScore[i]-lastRoundVertexScore[i]|$；

8. 迭代次数 $Iteration++$；

9. 如果 λ 的值大于指定的终止阈值并且迭代次数 $Iteration$ 小于设定的最大迭代次数，返回第 5 步；

10. 根据结点值的大小获得候选实体集 C_e 的一个降序排序结果；

11. 根据实际情况需要选取前 N 个候选实体为最后的目标实体集 T_e.

4.6 算法的时间效率分析

CER 算法的运行包括两个部分，图模型的构建和利用随机游走进行排序，不妨记构建图模型的时间复杂度为 T_1，排序过程的时间复杂度为 T_2，总的时间复杂度为 $T=T_1+T_2$. 令输入的网页片段集的大小为 $|S|$，关键字个数为 $|K_e|$，平均每个网页片段中含有的实体数为 $\overline{|E_s|}$，则在构建图模型中，结点初值的计算需要时间为 $O(|S|\times|K_e|)$；关键字的 D 值计算为很小的常数时间，因为关键字的个数通常很少，可忽略这部分时间开销；边权的计算时间为 $O(|S|\times\overline{|E_s|})$，我们可以将计算结点初值和计算边权融合在一起，故最终 $T_1=O(|S|\times(\overline{|E_s|}+|K_e|))$. 记图中的结点数为 $|V|$，随机游走过程收敛需要迭代的次数为 C，则排序过程主要为矩阵运算的时间开销 $T_2=O(C\times|V|\times|V|)$. 综上，算法的总时间开销为

$$T=O(|S|\times(\overline{|E_s|}+|K_e|))+O(C\times|V|\times|V|) \tag{10}$$

5 实验及分析

这一节描述了实验的数据集、实验的设置、对比实验和实验结果，并对实验结果进行了分析讨论.

5.1 实验数据集

本文实验使用了两个数据集 Book-Crossing[①](后面简称为“书本数据集”)和 Fortune 1000 Contact Information[②](后面简称为“公司数据集”). 其中，(1) 书本数据集是 Cai-Nicolas Ziegler 在 Book-Crossing 社区爬取的，它包含了 278 858 名匿名用户对 271 379 本书的 1 149 780 个评分. 本文采用书本数据集中的 BX-Books 表进行实验，去除表中的图像信息等无法供给实验用的字段后，BX-Books 的字段有 ISBN 号、书名、作者、出版年份和出版商；(2) 公司数据集收集了《财富》杂志评选的 1000 强企业的相关信息，该数据集由 Andy Pavlo 提供. 公司数据集包含的属性字段有公司名、地址、所在城市、所在州名、邮编、电话、网址、通用邮箱、CEO 及 CEO 邮箱.

5.2 实验设置

为了充分验证 CER 算法进行关系数据库缺失值修复的有效性，本文实验采取了以下 3 个方法：

① http://grouplens.org/datasets/book-crossing/

② http://www.cs.cmu.edu/~pavlo/datasets/fortune1000/

(1)随机从数据集中抽取元组进行实验;(2)随机选取元组中的某个属性字段,将该字段的属性值设置为空值,然后进行修复工作,最后评测修复的属性值与原属性值是否一致;(3)与两个算法进行比较,一个算法提供基准值,该算法以实体在网页片段中的频率作为实体排名因素;另一个算法则采用了权值叠加的方法来进行候选实体的筛选.

数据库的缺失值修复过程涉及查询生成、文档集检索、实体抽取和实体排序.本文实验着重对比和分析实体排序过程,其余的技术采用最常见的方法来实现,这么做并不会引起不必要的实验噪声,因为所有的实体排序算法均在同样的前提基础上运行.

实验中,我们生成查询的方式是选取除缺失属性值之外的其余所有属性值来构建查询.例如,书本数据集中的缺失元组 Tuple1(ISBN, Title, Author, Publication Year, Publisher)=(0001047663,"Matter Of Honour", null, 2000, Trafalgar Square Books),则构建的查询为〈"0001047663" "Matter Of Honour" "Author" "2000" "Trafalgar Square Books"〉.在这种查询生成方式下,两个数据集的查询召回率如表2所示.本文使用的文档集是从互联网上抓取的网页片段,文档集的检索采用商业搜索引擎必应 Bing①.它是由微软公司推出的全新搜索引擎服务.这里设置每个查询使用的网页片段数目最大值为100个,这个数值参考了文献[4]的实验结论.更具体的网页片段数则与不同的元组有关.两个数据集的网页片段个数情况如表2所示.从网页片段集中抽取实体采用的是 NER 实体抽取技术.本文把符合以下两点特征的网页片段称为目标网页片段:(1)网页片段中包含有目标实体;(2)NER 技术可以抽取出该网页片段的目标实体.表2列出了每个元组用到的目标网页片段数的平均值.显然,能够用的目标网页片段数越多,则越有利于目标实体的寻找,表2的数据解释了表3中公司数据集的排序效果好于书本数据集的原因.实体排序算法是本文实验的分析重点,总共采用了3种不同的算法:基于频率的排序算法、基于权值叠加的排序算法和本文提出的 CER 排序算法.更详细的实验信息参见下一小节的实验对比.

表2 Web 网页片段相关信息统计(设定网页片段个数为100个)

数据集	选取的元组数目	能够搜索到答案的元组数	召回率	平均每个查询返回的网页片段数目	平均目标网页片段数目
书本数据集	507	400	0.79	68.13	9.91
公司数据集	307	285	0.93	61.50	14.46

表3 排序结果(设定网页片段个数为100个)

数据集	算法	Top1 元组数	Top3 元组数	Top5 元组数	Top10 元组数	*MRR* 值
书本数据集	基于频率	155	258	308	345	0.5518
	基于权值叠加	162	269	322	358	0.5695
	本文	**248**	**320**	**350**	**377**	**0.7312**
公司数据集	基于频率	216	246	255	268	0.8238
	基于权值叠加	224	248	262	269	0.8436
	本文	**245**	**271**	**271**	**271**	**0.9062**

5.3 实验对比

本文用于评价实验结果的因素有3个:

(1) Top$N$. 代表排序结果中,目标实体出现在前 N 个候选实体.

(2) Top$N$ 的准确率. Top$N$ 的元组数除以参与测试的所有元组的数目.由于本文的关注重点在于排序算法,为了去除查询和抽取带来的不必要影响,计算准确率的分母采用的是能够搜索到答案的元组数.

(3) *MRR*. 所有的目标实体的排名位置的倒数和除以元组数. *MRR* 的值在0到1之间,值越接近1表示排序的效果越好.

$$MRR=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank(ans_i)} \tag{11}$$

式(11)中 $rank(ans_i)$ 表示第 i 个元组的目标实体在最终的排序结果中的位置.

(1)与基于频率排序算法的比较

本文的算法在候选实体抽取阶段使用了 NER 技术,为了验证我们排序算法的有效性,我们测试了对 NER 抽取的候选实体只使用频率进行排序的方法.从表3可以看到,本文的算法与单纯使用频率进行排序相比更加有效,书本数据集排在第一名的提高了23.25%,*MRR* 提高了32.51%.而在公司数据集中目标实体排第一名的提高了10.18%,*MRR* 提高了10%.并且,若采用排名第一的候选答案进行数据库缺失值填充,本文的算法对于能够抽取出候

① https://datamarket.azure.com/dataset/bing/search

选答案的元组，在书本数据库中能达到 62% 的准确率，而公司数据集能达到 85.96% 的准确率. 这是因为在 Web 中有更多与公司相关的记录. 本文的算法能够比基于频率进行排序的方法好，这是很容易理解的，因为基于频率的算法只利用了目标实体在网页片段中出现次数更多的这一个特点. 然而，这个特点并不具有普遍适用性，存在很多的实体，它们频率并不是最高，但却是目标实体，对于这种情况，基于频率的算法就失效了. 所以，必须挖掘出与目标实体相关的更多特性，比如与关键词实体的距离，与其他各个实体间的关系等，本文的算法正是做到了这一点. 比如表 1 中的例子，抽取出来的候选实体 Alan Garner 和 A.E. Van Vogt 在给定的网页片段集的背景下，其频率均为 1，那么就无法通过统计频率的方法来筛选出目标实体.

(2) 与权值叠加排序算法进行比较

权值叠加排序算法也考虑候选实体的各种特征，比如实体间的距离，网页片段的排名等，并对这些特征值进行加权和，从而得出候选实体的排序. 从实验结果表 3 可以看出，简单的权值叠加虽然有效，但效果并不显著. 与基于频率进行排序的算法相比，书本数据集的 *MRR* 值仅提高了 3.21%，公司数据集的 *MRR* 只提高了 2.40%. 权值叠加算法虽然能够通过挖掘目标实体的特点来提高频率低的目标实体的最终排名，但是它并没有利用各个网页片段中实体间的关系，如果一个网页片段中不包含候选实体，那么它对基于权值叠加的算法基本没有用处. 然而事实上，当人为地利用这些网页片段进行判断的时候，每一个网页片段都是有用的，因为人能够把各个网页片段所提供的信息综合起来进行推断. 本文的算法正是把各个网页片段通过关键词实体，中间实体和候选实体用图的方法综合起来，并且用边权表征实体间的联系，最后通过迭代计算，得到与关键词实体最相关的候选实体. 例如表 1 中的例子，基于权值叠加的算法无法利用所提供的网页片段来排序，因为不存在关键字与候选实体同时出现的网页片段，那么权值叠加就无法进行. 但是，本文的算法能够利用中间实体“fantasy novel”把网页片段 1 和网页片段 2 结合起来进行推断并排序. 本文的算法充分利用网页片段所提供的候选答案的上下文环境，所以才得到了更高的准确率，相比于权值叠加排序算法，书本数据集中目标实体排在第一名的提高了 21.5%，*MRR* 值提高了 28.40%；公司数据集中目标实体排在第一名的提高了 7.37%，*MRR* 值提高了 7.42%.

5.4　分　析

本文用于修复关系数据库缺失值的数据源是 Web 中的网页片段集，前面分析了在固定(设定数量 100 个)网页片段集下不同排序算法针对不同目标数据库的修复效果，下面将会测试在不同数量的网页片段集下排序算法的表现.

通过分析查询返回的网页片段我们发现，实际的网页片段非常复杂和混乱. 主要体现在 4 个方面：(1) 网页片段并不是完整的句子，同一个查询返回的网页片段描述的内容不同；(2) 同一个意思又有不同的表述方式，比如“THIS SIDE OF PARADISE is written by F. Scott Fitzgerald”，“F. Scott Fitzgerald finished his another book—THIS SIDE OF PARADISE in 1988.”；(3) 同一个名称可以有不同的写法，比如作者的名字可以写成“F. Scott Fitzgerald”“F. Scott”“Scott”；(4) 信息的公开程度，流行程度不同，导致网上具有的信息并不相同，比如一本流行的书，会有更多的关于它的讨论的信息，而对于一本冷门的书则只有很少的网页会提及. 图 6 展示了在不同的网页片段数的情况下，网页片段集包含目标实体的元组个数的变化情况. 随着网页片段数的增加，能够召回目标实体的元组数也呈上升趋势，但网页片段数到达一定的值后，这样的元组数达到一个稳定的值. 对于书本数据集，在网页片段数达到 70 个以后，能召回目标实体的元组数平均增长幅度只有 0.93%；对于公司数据集，在网页片段数达到 50 个以后，能召回目标实体的元组数增长的个数只有 1 到 2 个. 这说明公司数据集所对应的网页片段集能提供较为精准和相关的信息，因为大部分(276/307)的元组在设定网页片段数为 50 个的情况下就能够找到目标实体. 这也解释了表 3 中，不同的排序算法下，公司数据集的排序效果均好于书本数据集的原因.

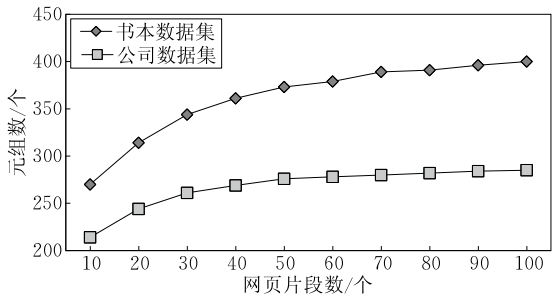
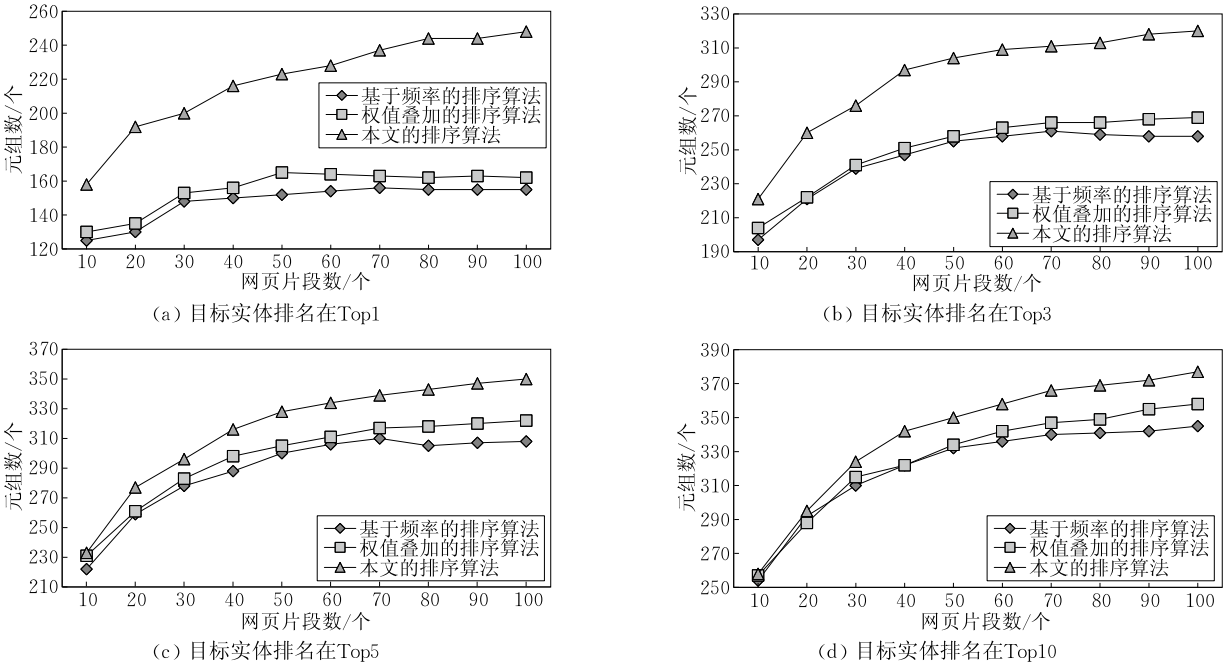


图 6　网页片段包含有目标实体的元组数随网页片段个数的变化

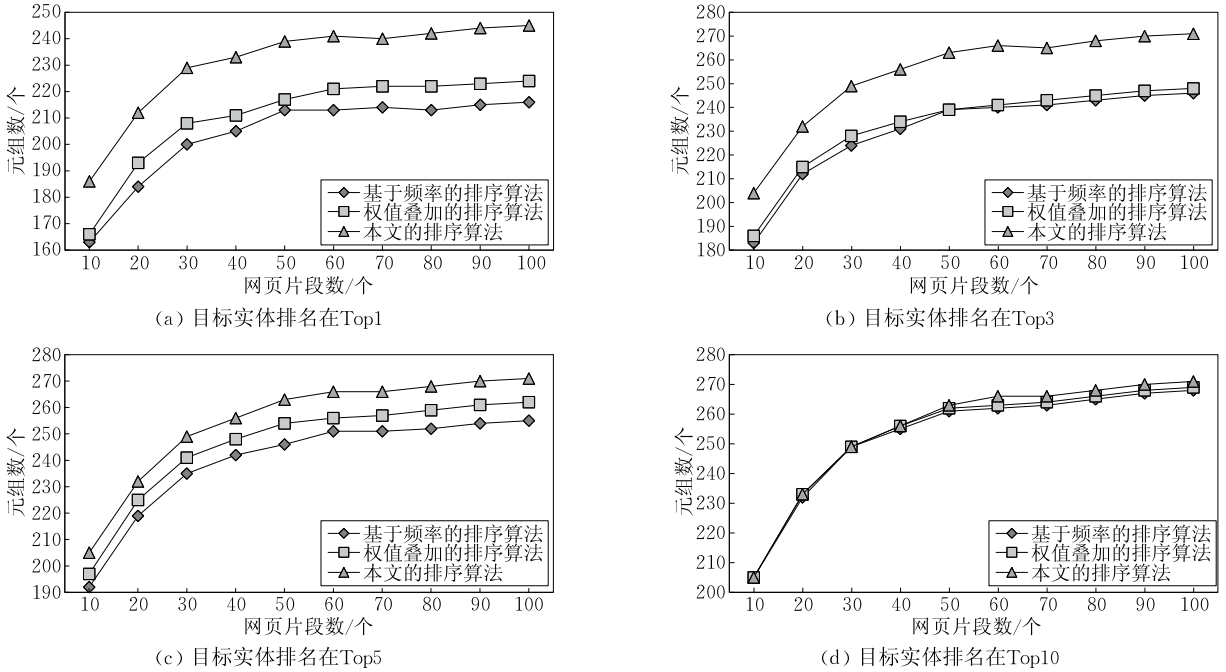
本文利用网页片段提供的信息构建实体的关系图,会受到网页片段集的影响.在网页片段比较少的情况下,我们能够利用的信息相应较少,构建的候选实体的上下文关系具有一定的局限性,这是不可避免的.但是,即使在这样的情况下,本文的排序算法效果还是优于基于频率和基于权值叠加的排序算法,这是因为不管信息源的多寡,本文的算法可以综合利用网页片段所提供的信息.当然,网页片段提供的关于候选实体的信息越全面越多,我们构建的图越接近实体本身的关系图,我们的算法也就越有效.图 7 和图 8 分别展示了不同的排序算法在书本数据集和公司数据集上目标实体排名在 Top$N$ 的元组数目随着网页片段数量的变化而变化的情况.从这两个图可以看出,本文的算法能够很好地把目标实体排在第一,这对于数据库缺失值修复是很重要的,因为通常人们采用排名最靠前的候选实体来修复数据库.对于需要多个目标实体来修复数据库的情况,比如,一本书有多个作者,Top3 和 Top5 的排名则



(a) 目标实体排名在Top1 (b) 目标实体排名在Top3

(c) 目标实体排名在Top5 (d) 目标实体排名在Top10

图 7 书本数据集目标实体在 TopN 的元组个数随网页片段个数的变化



(a) 目标实体排名在Top1 (b) 目标实体排名在Top3

(c) 目标实体排名在Top5 (d) 目标实体排名在Top10

图 8 公司数据集目标实体在 Top$N$ 的元组个数随网页片段个数的变化

具有指导作用. Top10 的排名则指示了排序算法能够把目标实体返回供用户选择的能力. 3 个排序算法的效果随着网页片段数的增加而得到提升，本文算法的提升效果优于其余两种排序算法. 注意到，3 个算法在公司数据集中目标实体排在 Top10 的元组数的变化曲线几乎重叠，这是因为前面已经提到过的原因，即公司数据集所对应的网页片段集本身能提供较为精确的信息，但在 Top1、Top3 和 Top5 上，本文的排序算法还是具有优势，这说明了本文算法的整体的排序效果较好.

6　总结和展望

本文针对大数据环境下，利用互联网数据修复关系数据库中缺失值的问题，提出了一个上下文相关的实体排序算法 CER. 该算法首先通过完整属性值构建合理的查询并抽取 Web 实体，然后根据实体的关联特征构建图模型，最后用指定入口的随机游走思想迭代计算，为候选实体进行排序. 实验证明，相较于基于频率统计和基于权值叠加的实体排序方法，本文算法能提高缺失值修复的准确率.

在图模型中，实体间的影响直接关系到最终实体的排名，如何更高效合理地细化实体间的影响有待进一步研究.

参　考　文　献

[1] Li Jian-Zhong, Liu Xian-Min. An important aspect of big data: Data usability. Journal of Computer Research and Development, 2013, 50(6): 1147-1162(in Chinese)
(李建中，刘显敏. 大数据的一个重要方面：数据可用性. 计算机研究与发展，2013，50(6)：1147-1162)

[2] Fan W, Geerts F. Foundations of data quality management. Synthesis Lectures on Data Management, 2012, 4(5): 1-217

[3] Wu Sen, Feng Xiao-Dong, Shan Zhi-Guang. Missing data imputation approach based on incomplete data clustering. Chinese Journal of Computers, 2012, 35(8): 1726-1738(in Chinese)
(武森，冯小东，单志广. 基于不完备数据聚类的缺失数据填补方法. 计算机学报，2012，35(8)：1726-1738)

[4] Dumais S, Banko M, Brill E, et al. Web question answering: Is more always better?//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002: 291-298

[5] Li X, Meng W, Yu C. T-verifier: Verifying truthfulness of fact statements//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering (ICDE). Hannover, Germany, 2011: 63-74

[6] Page L, Brin S, Motwani R, et al. The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, California, USA: Technical Report: 422, 1999

[7] Grzymala-Busse J W, Hu M. A comparison of several approaches to missing attribute values in data mining//Ziarko W, Yao Yiyu eds. Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science 2005. Berlin Heidelberg: Springer, 2001: 378-385

[8] Han Jin-Yu, Xu Li-Zhen, Dong Yi-Sheng. An overview of data quality research. Computer Science, 2008, 35(2): 1-5 (in Chinese)
(韩京宇，徐立臻，董逸生. 数据质量研究综述. 计算机科学，2008，35(2)：1-5)

[9] Pan Li-Qiang, Li Jian-Zhong, Luo Ji-Zhou. A temporal and spatial correlation based missing values imputation algorithm in wireless sensor networks. Chinese Journal of Computers, 2010, 33(1): 1-11(in Chinese)
(潘立强，李建中，骆吉洲. 传感器网络中一种基于时-空相关性的缺失值估计算法. 计算机学报，2010，33(1)：1-11)

[10] Wang Li, Zhang Rong, Sha Chao-Feng, et al. A product normalization method for e-commerce. Chinese Journal of Computers, 2014, 37(2): 312-325(in Chinese)
(王立，张蓉，沙朝锋等. 电子商务商品归一化方法研究. 计算机学报，2014，37(2)：312-325)

[11] Li Z, Sharaf M A, Sitbon L, et al. WebPut: Efficient web-based data imputation//Wang X S, Cruz I, Delis A, Huang Guangyan eds. Web Information Systems Engineering-WISE 2012. Lecture Notes in Computer Science 7651. Berlin Heidelberg: Springer, 2012: 243-256

[12] Lin J. The web as a resource for question answering: Perspectives and challenges//Proceedings of the 3rd International Conference on Language Resource and Evaluation (LREC 2002). Las Palmas, Spain. 2002

[13] Zhao S, Grishman R. Extracting relations with integrated information using kernel methods//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Michigan, USA, 2005: 419-426

[14] Mikheev A, Moens M, Grover C. Named entity recognition without gazetteers//Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. Bergen, Norway, 1999: 1-8

[15] Collins M, Singer Y. Unsupervised models for named entity classification//Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Maryland, USA, 1999: 100-110

[16] Ratnaparkhi A. A maximum entropy model for part-of-speech tagging//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Pennsylvania, USA, 1996, 1: 133-142

[17] Riloff E. Automatically generating extraction patterns from untagged text//Proceedings of the National Conference on Artificial Intelligence. Portland, USA, 1996: 1044-1049

[18] Sun H, Duan N, Duan Y, et al. Answer extraction from passage graph for question answering//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Washington, USA, 2013: 2169-2175

[19] Ko J, Nyberg E, Si L. A probabilistic graphical model for joint answer ranking in question answering//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, Netherlands, 2007: 343-350

[20] Li Bo, Gao Wen-Jun, Qiu Xi-Peng. Constructing an answer ranking model using semantic analysis and statistical method for question answering. Journal of Chinese Information Processing, 2009, 23(2): 23-27(in Chinese)
(李波，高文君，邱锡鹏. 基于语法分析和统计方法的答案排序模型. 中文信息学报，2009，23(2)：23-27)

[21] Raviv H, Carmel D, Kurland O. A ranking framework for entity oriented search using Markov random fields//Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search. Portland, USA, 2012: 1

[22] Vercoustre A M, Pehcevski J, Naumovski V. Topic difficulty prediction in entity ranking//Geva S, Kamps J, Trotman A eds. Advances in Focused Retrieval. Berlin Heidelberg: Springer, 2009: 280-291

[23] Mottin D, Palpanas T, Velegrakis Y. Entity ranking using click-log information. Intelligent Data Analysis, 2013, 17(5): 837-856

[24] Ganesan K, Zhai C X. Opinion-based entity ranking. Information Retrieval, 2012, 15(2): 116-150

CHEN Zhao-Qiang, born in 1988, Ph. D. candidate. His research interest is data quality management.

LI Jia-Jun, born in 1990, M. S. candidate. His research interest is data quality management.

JIANG Chuan, born in 1992, M. S. candidate. Her research interest is data quality management.

LIU Hai-Long, born in 1980, Ph. D., lecturer. His research interests include data management and software technology.

CHEN Qun, born in 1976, Ph. D., professor, Ph. D. supervisor. His research interests include data management, data quality management and information retrieval.

LI Zhan-Huai, born in 1961, Ph. D., professor, Ph. D. supervisor. His research interest is database theory and technology.

Background

Data integrity is an essential aspect of data quality. Unfortunately, data missing is very common in real life and it makes data-based decisions unreliable. And it's important to repair the missing data in activity of data management.

There already exist some traditional data imputation methods. The common characteristic of these methods is that they utilize data come from database and they are restricted to the limited available data. In recent years, some researchers proposed utilizing web data to fill the missing data. Most of these researches focused on formulating effective queries to retrieve more related documents from web. And some researches mentioned entity ranking issue, for example, frequency- based ranking method, weight-based ranking method, pattern-based ranking method and so on. The entity ranking issue is also part of Question Answering domain and part of Natural Language Processing domain. According analysis of these ranking methods, we found three shortages of them: First, some of them are too simple and result in bad ranking results(w. r. t. Frequency-based); Second, Some of them are too sophisticated to use in reality(w. r. t. Pattern-based); Third, Some of them need training data while getting proper training data is not an easy thing (w. r. t. Weight-based).

This paper proposed an entity ranking method called CER(Context-aware Entity Ranking). It does not need training data and it can take advantage of the context of candidate entities to make an comprehensive ranking utilizing graph model. This method is well suited to the web-based data imputation problem.

This work is supported by the National Basic Research Program(973 program) of China under Grant No. 2012CB316203, the National Natural Science Foundation of China under Grant Nos. 61332006, 61472321 and the NWPU Basic Research Foundation (3102014JSJ0013, 3102014JSJ0005).