复杂数据上的实体识别技术研究

王宏志[1)] 樊文飞[1),2)]

[1)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

[2)](爱丁堡大学信息学院 英国爱丁堡 EH8 9AB)

摘 要 复杂数据当前有着广泛的应用.有效地使用复杂数据需要对其质量进行管理.实体识别是数据质量管理的基本操作,用于在数据集合中发现同一实体的不同描述,其在数据质量管理中可以用于错误检测、不一致数据发现等.由于包含复杂的结构信息,复杂数据上的实体识别与传统文本和关系数据上的实体识别不同,带来了新的技术上的挑战.该文介绍了复杂数据上实体识别的概念和应用,分别讨论了 XML 数据、图数据和复杂网络上实体识别技术的原理,最后展望了未来的研究方向.

关键词 数据质量;复杂数据;实体识别;XML 图;复杂网络

中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2011.01843

Object Identification on Complex Data: A Survey

WANG Hong-Zhi[1)] FAN Wen-Fei[1),2)]

[1)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

[2)](*School of Informatics, University of Edinburgh, Edinburgh, Scotland, UK* EH8 9AB)

Abstract It is increasingly common to find data with a complex structure in the real world. To effectively use complex data in practice, necessary techniques must be in place to improve the quality of the data. Entity resolution is a central issue in data quality management for complex objects. It is to find the data objects that refer to the same real-world entity, and to cluster such objects together. It has been proven extremely useful in data fusion, inconsistency detection and in data repairing. Nevertheless, the complex structures of data introduce new challenges and make object identification much harder than record matching on relational data. In response to the new challenges, there has been a lost of work on this topic. This paper aims to provide an overview of recent advances in the study of object identification, on complex objects including XML, graph data and complex networks. For XML data, we survey techniques of pairwise entity and group-wise entity resolution. For graph data, we focus on how to determine whether two graphs refer to the same real-world entity. We also present the metrics and methods for identifying vertexes that pertain to the same real-world entity in a complex network. Finally we discuss directions for future research.

Keywords data quality; complex data; object identification; XML graph; complex network

收稿日期:2011-08-11;最终修改稿收到日期:2011-09-15.本课题得到国家自然科学基金(61003046,61033015,61133002)、RSE-NSFC 交流项目(61111130189)、国家"九七三"重点基础研究发展规划项目基金(2012CB316200)以及教育部博士点基金(20102302120054)资助.

王宏志,男,1978 年生,博士,副教授,主要研究领域为数据管理,主要包括数据质量、XML 数据管理和图数据管理.**樊文飞**(通信作者),男,1963 年生,博士,教授,主要研究领域为数据库理论与系统,特别是数据质量、信息集成、分布式查询处理、查询语言、推荐系统、社会网络、Web 服务和 XML. E-mail: wenfei@inf.ed.ac.uk.

1 引 言

以数据为中心的系统当前已经得到广泛的应用.但是,这些系统通常假设所采集的数据是正确无误的.遗憾的是,这种假设在现实世界并不成立.现实信息系统中存储的数据多含有各种各样的错误,例如不正确、不一致、重复、不精确、不完整或者过时陈旧.最近统计表明,美国企业中 1%～30%的数据存在各类错误和误差[1].仅就医疗数据库而言,其中 13.6%～81%的关键数据不完整或陈旧[2].根据市场研究公司 Gartner 的调查,全球财富 1000 强公司中超过 25%的关键数据不正确或不准确[3].如果数据质量得不到保障,再先进的数据库系统和搜索引擎也无济于事,无法为用户查询提供准确的信息.

劣质数据经常产生严重的后果.根据美国医疗委员会的统计,由于数据错误引起的医疗事故仅在美国每年便导致高达 98000 名患者的死亡[4].据估算,数据错误每年造成美国工业界大约 6110 亿美元的经济损失,约占美国 GDP 的 6%[5].以电信为例,数据错误经常导致故障排除的延误、多余设备的租用和服务费收取的错误,并因此损害企业信誉并失去用户.美国零售业每年仅因标价错误即损失 25 亿美元[6].就银行而言,仅因数据质量管理失误而失察的信用卡欺诈在 2006 年即造成 48 亿美元的损失[7].劣质数据正日益成为困扰企业信息化的梦魇.统计数据显示,50%以上的数据仓库项目由于数据质量问题而不得不取消或延迟[8].在典型的信息系统项目中,时间和成本预算的 30%～80%实际用于清理数据而非系统开发[9].专家估算目前数据质量问题平均给每个企业增加的成本是企业收入的 10%～20%[10].

上述实例表明了数据质量管理技术的需求存在于信息社会的各个领域.实体识别在数据质量管理中起着重要作用,是数据质量管理的主流研究方向之一.在一个或多个数据库中,同一个现实世界对象可能具有多种描述方法.实体识别的目的是在一个或多个数据库中辨识描述同一个实体的不同表示方法,正确地识别出数据库中的所有不同实体.实体识别的结果是数据库中所有不同实体的集合以及每个实体的不同描述方法.实体识别的结果可以在数据质量管理的其它阶段得到广泛应用,如冗余数据去重、错误数据发现、不一致数据发现与冲突消解等.在不同的文献中,实体识别有着不同的名称,包括对象识别、冗余发现、实体消解等.

现实应用中许多数据具有复杂的结构,如 XML 数据、图数据和复杂网络等,我们称这类数据为复杂数据.复杂数据在多种应用中广泛存在,如企业信息系统中的 XML 数据、互联网中的数据、社会网络、化学与生物数据库中的数据等.同一实体具有不同复杂数据描述方式的问题在各种应用领域的信息系统中普遍存在.例如互联网中有 XML、关系数据库、可建模成图的 RDF、HTML 等多种复杂形式.为了有效对这些数据实施质量管理,需要对复杂数据进行快速有效的实体识别.

尽管当前的实体识别技术有很多,但主要集中在文本形式的词组或关系数据上[11],对复杂数据上实体识别的研究还刚刚兴起,研究人员从不同角度开展了复杂数据上实体识别的工作.本文拟对多种类型复杂数据上实体识别技术研究现状加以综述.

本文第 2 节对复杂数据上实体识别问题进行概述,讨论复杂数据上实体识别问题的定义、必要性以及困难和挑战;第 3 节综述 XML 数据上的实体识别方法;第 4 节综述图结构数据上的实体识别方法;第 5 节综述复杂网络上的实体识别方法;第 6 节总结全文并展望未来的研究方向.

2 复杂数据上的实体识别问题概述

本节首先定义复杂数据上实体识别问题,接下来介绍复杂数据上实体识别的必要性,最后讨论复杂数据上实体识别的困难与挑战.

2.1 复杂数据上的实体识别问题定义

复杂数据上的实体识别有不同的分类方法.根据识别结果的不同,复杂数据上的实体识别可以分为成对的识别和成组的识别.前者的目的是判定两个数据对象 o_1 和 o_2 是否是描述同一现实世界的实体;后者的目的是将数据对象集合 S 分为子集 S_1, S_2, …, S_k,满足 $S_1 \cap S_2 \cap \cdots \cap S_k = \varnothing$ 且 $S_1 \cup S_2 \cup \cdots \cup S_k = S$,使 $\forall i \in [1,k]$, $\forall o_1, o_2 \in S_i$, o_1 和 o_2 描述现实中同一实体.在成对识别中,人们经常使用基于相似性函数的实体识别方法,即定义相似性函数 sim 或距离函数 distance,对于两个对象 o_1 和 o_2 以及阈值 ε,当 $\mathrm{sim}(o_1,o_2) \geq \varepsilon$ 或 $\mathrm{distance}(o_1,o_2) \leq \varepsilon$ 时,o_1 和 o_2 被看做同一实体.

根据识别对象的不同,复杂数据上的实体识别可以分为 XML 数据的实体识别、图结构数据的实体识别和复杂网络中结点的实体识别,这 3 种数据

上的实体识别都有一定应用，下一节将具体讨论这些实体识别技术的应用.

2.2　复杂数据上实体识别的必要性

当前复杂数据的应用日益广泛，特别是 Web、社会网络、生物结构等数据量日益增大. 这些数据来自多个数据源，特别是如 Facebook，Twitter 一类 Web2.0 和社会网络数据由大量不同的用户维护，这些数据中可能存在大量不一致数据. 对这些数据进行质量管理需要有效的复杂数据实体识别技术. 复杂数据实体识别的直接应用有如下实例：

互联网中的网站是自治的数据源. Web2.0 网站中的信息由多个用户维护，因此不同网站甚至相同网站的不同部分对同一实体可能存在不同描述，因而互联网信息的检索结果可能包含同一实体的不同描述. 这些结果一方面使用户浏览过量类似信息，浪费用户浏览检索结果的时间；另一方面检索结果中的不一致信息和从中产生的错误统计分析结果会误导用户决策. 对检索结果进行实体识别，并把最终检索结果按照实体分类，使得同一类中的信息描述同一实体，可以提供给用户高质量的互联网信息检索结果，从而提高互联网信息的有效利用率. 而当前互联网中信息量很大且更新非常频繁，仅 google 索引的网页数量就达 1T 之多，有质量保证的互联网信息搜集与检索系统对海量、动态且支持复杂数据的实体识别技术提出了要求.

当前许多文档由 XML 形式的数据描述，包括电子商务、电子政务文档等. 而 XML 文档也是用来进行多种类型数据源信息集成的重要信息描述标准，来自不同数据源的 XML 数据具有不同模式，也可能存在同一实体的不同表示. 在未经过处理的数据上直接进行查询，会得到冗余或者不一致的结果. 在这样的数据上直接进行统计分析可能将同一对象统计多次，从而得到错误的统计分析结果，导致错误的决策. 如果对集成的 XML 数据进行实体识别，使得每一类数据对象描述同一实体，在经过实体识别的 XML 数据上会得到更高质量的查询结果和更加精确的统计分析结果.

当前的生物数据库，生物分子数据可以建模成图结构数据. 考虑到数据中的错误和不精确，对应来自多个生物分子数据库中的对应同一分子的分子结构描述可能有所不同. 如果不经过实体识别，在多生物分子数据库集成得到的数据集中查询会得到冗余的结果或者不一致的结果，进而对集成生物分子数据库的挖掘与分析会得到错误的结果. 而如果将图数据实体识别技术应用在集成的生物分子数据库上，则可以得到更加准确的挖掘分析结果.

此外，由于图结构还可以用来表示对象之间的关系，例如社会网络，图结构提供的信息可以帮助图结构结点的实体识别. 例如利用社会网络可以帮助人名的识别. 单独按照人名本身进行识别，对于一些重名的人则难以识别；即使增加邮件、工作单位等信息，如果人员的工作单位发生变动，也难以判定哪些信息描述同一个人，而考虑到一个人的社会关系短时间内变化不大，用社会网络信息可以对人的信息加以更有效的识别，如文献[12-13]中的实验结果所示，利用论文合作者信息识别不同论文中作者是否指同一人时，可以得到更加精确的识别结果.

2.3　困难和挑战

复杂数据实体识别和简单数据上实体识别相比，由于包含了复杂的结构信息，且应用更加复杂，这导致复杂结构数据上实体识别面临如下挑战：

(1) 实体识别需要快速有效地辨识描述同一实体的复杂数据对象(如 XML 文档和图). 复杂数据的特点是具有丰富的结构信息. 如何有效利用复杂数据的结构信息，设计有效的方法辨识描述同一实体的复杂数据对象是第 1 个挑战性问题.

(2) 很多实际应用中存在海量的复杂数据. 常见的复杂数据，如社会网络、网站和生物分子式，其量非常大. 如何设计高效的复杂数据对象分类算法，快速有效地在海量复杂数据上进行实体识别是第 2 个挑战性问题.

(3) 一些应用中的复杂数据更新频繁，例如互联网信息和社会网络上的信息. 如何快速有效地处理更新频繁的动态复杂数据上的实体识别是第 3 个挑战性问题.

挑战不止上述几个方面，但即使对于上述问题，现有的技术也只是解决了一部分，特别是第 3 点，当前未有相应的方法提出.

3　XML 数据上的实体识别技术

由于其灵活性和可扩展性，XML 已经成为互联网、电子政务等多种类型数据表达与转换的标准，当前的信息系统中有大量数据以 XML 形式表示. 因而，有一些研究工作集中于 XML 数据上实体识别技术. 当前，XML 数据上实体识别的应用包括 Web 与 Peer-to-Peer 环境下描述同一实体数据的发现[14]、Web 上交换与发布的重复数据的发现[15-16]、

Web 上多数据源集成[17]等.

3.1 XML 数据的成对识别

XML 数据的成对识别用于发现描述同一实体的 XML 文档或者元素对,这类工作也称为 XML 文档匹配或元素匹配.成对识别的研究主要集中于定义 XML 数据的相似性或距离.

与结构化和非结构化数据相比,XML 数据最显著的特征是有丰富的结构信息,因此最常用匹配方法是使用结构信息描述 XML 文档相似性或距离.

由于 XML 文档可以用树结构来描述,因而树之间相似度描述方法可以用来描述 XML 文档之间的相似性.树编辑距离是最早衡量树之间相似度的方法,用于表示从一棵树变换到另外一棵树需要增加、删除或者修改标签的最少结点数量.Tai[18]提出了树编辑距离的概念并给出了第一个在 $O(n^6)$时间复杂度上计算树编辑距离的方法(其中 n 是树的结点数目).Milano 等人[19]提出基于覆盖的 XML 对象距离,两个 XML 树 U 和 V 之间的覆盖定义为可建立映射的最大结点个数.U 中的一个结点 u 映射到 V 中的一个结点 v 当且仅当它们从根到叶子的路径相同.

XMLDup 系统[20]用贝叶斯网络描述 XML 文档的相似性.该模型中的贝叶斯网络也可表示为树结构,以两个 XML 文档叶子上值之间相似性作为贝叶斯网络中叶子上的基本概率.对于两个元素而言,其相似性表示为一个概率,该概率由这两个元素的后代相似性对应的条件概率计算得到;两个 XML 文档之间的相似性定义为其根结点的相似性.文献[21]进一步讨论了用贝叶斯网络描述 XML 文档相似性过程中的优化策略,其策略是将 XML 文档的属性表示为一个向量,使用机器学习的方法确定文档的新结构,该结构更适合于基于贝叶斯网络的实体识别.

从另外一个角度看来,XML 文档可以提取成为一个集合,可以通过比较集合之间的相似性比较 XML 文档的相似性.文献[17]是较早研究 XML 数据实体识别的工作,提出了集成从 Web 中提取出树结构数据的方法.对于两个对象,通过将其转化为关键词向量并利用关键词向量的余弦相似性来计算其相似性.这种方法忽略了对象表示的层次结构,仅用相似性的权重表示向量中的对象相似性.文献[22]利用 XML 文档中根到叶子的路径集合表示 XML 文档,利用路径集合的相似性描述 XML 文档相似性.

文献[14,23]将结构和内容相似性结合在一起描述 XML 文档相似性.文献[14]中以结点相似性的平均值描述 XML 文档相似性,其中结点的相似性表示为结点名称相似性、路径相似性和结点所有后代的内容相似性的平均值.XDoI[23]中以两个 XML 树的内容相似性和结构相似性的积来描述 XML 树的相似性,其中内容相似性用匹配的公共叶结点个数来描述,结构相似性用匹配叶子的平均路径相似性之和描述.

还有一些工作集中于考虑 XML 数据的高效匹配.这类工作主要集中于基于树编辑距离匹配的加速算法,Zhang 和 Shasha[24]将树编辑距离的效率改进到了最坏情况 $O(n^4)$时间复杂度.Klein[25]提出了一种可以在 $O(n^3\log n)$时间内计算出树编辑距离的方法.Zhang 和 Klein 都是通过减小动态规划中子问题计算量加快树编辑距离的计算.Demaine 等人[26]在最近的一项工作中将树编辑距离算法时间复杂度提高到 $O(n^3)$.无论是理论分析还是实验结果,Zhang 和 Shasha 提出的算法在计算比较平衡的树的树编辑距离时是最好的,而 Demaine 的工作在计算任意树的树编辑距离的时候是最好的.

树编辑距离虽然能够很好地检测出相似的树对,但计算很复杂,其在大型 XML 数据库中的应用也会因此而受限.为了提高 XML 数据上近似连接的效率,很多工作都采纳了过滤-提取的方法,使用不同的方法快速估计出树编辑距离的上界和下界,从而减少计算树编辑距离的次数.比较知名的过滤-提取方法有 Deriving Upper and Lower Bounds[27], Binary Branch Distance[28],3 种直方图[29]和 Normal pq-Gram Distance[30].Deriving Upper and Lower Bounds 经常是很紧的,但对比较复杂的下界,计算它们需要 $O(n^2)$的时间.Binary Branch Distance 和直方图提供的下界可以被高效地计算出来,但是它们相对比较松,和准确的编辑距离可能相差较多.Normal pq-Gram 则可用于计算扇出树编辑距离的下界,这个下界也是一个可以快速计算出的近似下界.值得一提的是,通过组合这些下界进行多层的过滤,紧的下界也可以被高效地计算出来.

XML 的成对识别与数据近似连接操作密切相关.很多前人的工作使用转化类的方法来高效实现 XML 数据的近似连接,即将树转化成其它易于比较的数据结构,并用这些数据之间的相似度来近似描述原 XML 之间的相似度.在方法 pq-Gram[31]中,每一棵树都被转化成了一个 pq-Gram 集合,每一个

pq-Gram都是一个有特定形状的子树.通过这种转化,树之间的相似性可以在$O(n\log n)$时间内被估计出来.最新的转化类方法是 hashing tree 方法.在 hashing tree 方法[32]中,每棵树被转化成了一个 pivot 的集合,每一个 pivot 记录了两个结点以及两个节点之间的最近公共祖先.然而,虽然计算效率较高,以集合为基础的转化方法的连接准确率明显低于树编辑距离.在文献[31]报告的实验结果中,其方法的准确率明显低于树编辑距离方法.为了在高效率和准确率之间寻求一个平衡,Li 等提出了基于字符串的转化类方法 g-string[33],树被转化成了一个字符串,其中每一个字符对应一个小的子树.基于 g-string 的近似连接的准确率明显高于集合类方法,而连接效率高于树编辑距离方法.

3.2　XML 数据的成组识别

XML 数据的成组识别当前工作较少,根据识别对象的不同,可以分为文档级的识别和元素级的识别.

文档级的识别用于对 XML 文档进行分类,使得每类中的文档描述同一实体.这类方法主要是关系数据上实体识别方法的扩展.其中 SXNM[15]将关系数据上邻居排序方法应用到 XML 数据上,其基本思想是将相似的文档加以分组,仅将同一分组中的文档进行成对的比较,从而避免 XML 文档上无用的比较.DogmatiX[16]中以对应扩展关系数据中基于元组匹配的识别方法[34]对 XML 数据进行实体识别,该方法包含 4 个步骤:首先将 XML 文档拆为候选元组,接下来根据候选元组类型选择计算相似性的方法,然后根据选择的相似性方法求得成对的相似文档,最后通过求得匹配成对相似文档的闭包得到实体识别结果.

元素级的 XML 数据识别用于对同一 XML 文档中的元素进行分类,使得每一类元素描述同一实体.文献[35]利用专家定义的元素之间关系来进行分类,即专家为每个元素 e 都定义一个判定集合 T,集合中的元素或属性用于判定 e 的分类;对于两个元素 e_1 和 e_2,如果其对应判定集合 T_1 和 T_2 中的每个对应元素或属性都已经被判定成为相似,则其二者被判定为描述同一实体.基于这种判定集合的定义方法,XML 文档中元素可以用图结构描述.如果这个图是 DAG,可以自底向上地进行分类;如果该图中存在圈,则多次遍历该图,不断对元素进行重分类直到分类不再发生变化.文献[36]提出领域无关的 XML 数据元素识别算法,基于同一 XML 文档中元素相似性进行识别,这种方法自顶向下地遍历 XML 树结构并标记每一层的重复元素,利用一个编辑距离函数和阈值来判定两个元素是否描述同一实体并通过传递闭包对元素进行聚类.为了最小化比较的次数,研究人员还提出了过滤策略以加速处理.XDoI[23]将 XML 文档划分成子树的集合,以子树为单位实现 XML 数据上元素的实体识别.

4　图数据上的实体识别技术

作为一种重要的数据类型,图模型可以用来描述网站结构、分子式等具有广泛应用的数据.图结构上实体识别技术的应用包括相似 Web 文档发现[22]、化学分子式和基因交互网络的识别[37]、Web 网站的识别[38-39]等.当前图数据实体识别技术主要集中在描述同一实体的图数据的判定上.该判定主要基于图数据的结构相似性.

一类方法基于图之间的结构映射关系进行实体判定,即若两个图之间的点满足某种映射关系,则这两个图判定为匹配,即描述同一实体.这类判定方法包括文献[38-41].其中文献[40]基于图同构判定图是否匹配,文献[38]将网站建模成为图,利用图模拟判定数据是否描述同一网站.文献[39]将图同态和同构拓展为 p 同态和 1-1p 同态来判定图是否匹配,其基本思想是将同构和同态中的边-边映射拓展为边-路径映射.考虑到图同构、p 同态和 1-1p 同态的判定都是 NP 完全问题,为了解决图同构和同态固有的难计算问题,文献[41]将同态映射修改为界限同态映射,即将一个图中的边映射为另一个图中长度有界限的路径,图之间基于该映射的匹配关系可以在$O(n^3)$时间内判定.相关概念定义如下,读者可根据其定义比较其相异之处,其中概念中的图是有标签的图,待判定的图是 $G_1=(V_1,E_1)$ 和 $G_2=(V_2,E_2)$,标签函数为 φ.

定义 1(图同构).　同构关系是一个双射函数 $\lambda:V_1\leftrightarrow V_2$,其中对于每个 $v\in V_1$,$\varphi(v)=\varphi(\lambda v)$,且 $(u,v)\in E_1$ 当且仅当 $(\lambda u,\lambda v)\in E_2$.如果在 V_1 和 V_2 上存在该同构关系,G_1 与 G_2 满足同构关系.

定义 2(图模拟).　模拟关系 $\leq$ 是一个定义在 V_1 和 V_2 上的二元关系,对于 $v_1\in V_1$,$v_2\in V_2$,如果:(1) $\varphi(v_1)=\varphi(v_2)$;(2) $\forall v'\in\{v\mid(v,v_1)\in E_1\}$,$\exists v''\in\{v\mid(v,v_2)\in E_2\}$ 满足 $v'\leq v''$,则 v_1 和 v_2 满足 $v_1\leq v_2$.如果 $\leq$ 是满射,则 G_2 是 G_1 的模拟.

定义 3(p 同态).　给定结点相似性矩阵 mat()

和相似性阈值 ε，如果存在从 $V_1 \sim V_2$ 的映射 λ 满足对于任意 $v \in V_1$，满足：(1) $\mathrm{mat}(v, \lambda v) \geq \varepsilon$；(2) $\forall (v, v') \in E_1$，G_2 中存在一条非空路径 $\lambda v / \cdots / \lambda v'$，则 G_1 和 G_2 满足 p 同态关系.

定义 4(1-1p 同态). 如果 G_1 到 G_2 存在的 p 同态映射 λ 是 1-1 映射，即对任意两个不同的 $v_1, v_2 \in V_1$，$\lambda v_1 \neq \lambda v_2$，$G_1$ 和 G_2 满足 1-1p 同态关系.

定义 5(界限同态). 如果存在二元关系 $S \subseteq V_1 \times V_2$ 满足

(1) $\forall u \in V_1$，存在 $v \in V_2$，满足 $(u, v) \in S$；

(2) $\forall (u, v) \in S$，(a) $\lambda u = \lambda v$；(b) 对于 $(u, u') \in E_1$，存在非空路径 $\rho = v / \cdots / v'$，满足 $(u', v') \in S$ 且 ρ 的长度不大于 k.

我们使用一个例子对比上述概念，在图 1 中，G_1 和 G_4 满足 simulation 关系，但和 G_2 不满足 simulation 关系；G_1 和 G_2 满足 p 同态关系但不满足 1-1p 同态关系，G_1 和 G_3 满足 1-1p 同态关系；对于 $k=2$，G_1 和 G_2 满足界限同态但由于 A 到 B 之间的路径为 3，G_1 和 G_5 不满足界限同态.

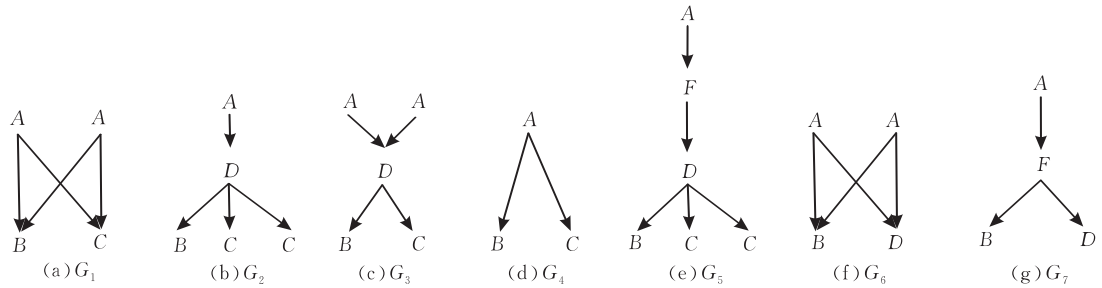

图 1

另一类方法基于图的相似性或距离判定图是否匹配，如果两个图的相似性大于某个阈值或距离小于某个阈值，则认为其描述同一实体. 上述映射关系可用来描述图相似性，例如基于图同构可以用图中不匹配点的数量定义图之间的距离[40]；基于 p 同态和 1-1p 同态可以依据可建立 p 同态或 1-1p 同态关系点的数量定义图之间的相似性[39]. 文献[42]利用图中边数减去最大公共子图中边数定义图之间的相似性，并提出基于子结构的索引支持近似查询处理. 文献[42]利用图之间编辑距离定义图相似性. 为了近似计算图相似性，该论文使用星形结构加速图相似性计算.

考虑图 1 中的图，G_1 和 G_6 之间基于图同构不匹配结点定义的距离为 1，因为去掉 C 点和 D 点则两图同构；基于 p 同态定义的 G_1 和 G_7 之间的相似性为 3/4，因为在建立的 p 同态关系中 G_1 中的 4 个结点有 3 个可以和 G_7 中的点建立 p 同态关系；G_1 和 G_6 的最大相似子图为包括两个 A 结点和 B 结点的子图，则 G_1 和 G_6 基于最大子图定义的距离定义为 $4-2=2$；G_2 和 G_7 的图编辑距离为 4，对应的 4 个操作是将 D 的标签修改为 F，删除一个 C 结点以及 D 结点到该结点中的边，将 C 结点的标签修改为 D.

文献[43-44]提出了一种核函数，该核函数基于图结构数据的特征定义两个图中结点的相似性，利用结点的相似性之和来定义图之间的相似性. 用来定义核函数的特征可以是邻居标签的直方图、结点的度、邻居的标签集合、结点所在的特殊子图等.

5 复杂网络上的实体识别技术

复杂网络可以用来描述互联网、社会网络或者更广泛地描述具有二元关系的对象集合. 复杂网络的结构信息有助于其中结点对应实体的识别. 复杂网络上实体识别技术的应用包括镜像网站的检测[45]、社会网络中人名的识别[12-13]、互联网信息检索等.

早期的工作面向具体应用采用简单的方法进行复杂网络中的实体识别. 文献[45]以检测镜像网站为背景，将网站建模为图中的结点，该识别过程分为两步. 第 1 步考虑数据库中每个主机中的 URL，将每个网站抽取为一个 URL 集合，并计算两个集合之间的相似性，输出可能的镜像，根据其相似性排序. 第 2 步中处理上一步得到的候选对集合，并检测其是否是镜像并判定其是否重叠. 文献[46]以社会网络为背景，提出一种交互式实体识别工具 D-Dupe 应用于管理图结构数据. D-Dupe 用户通过区分或合并社会网络的结点解决同名问题，利用字符串相似性描述结点相似性，通过用户反馈确定相似性函数的权重. 文献[47]提出的 GeoDDupe 将 D-Dupe 拓展为空间网络上结点的识别.

另一类工作研究复杂网络中的结点相似性描述方法.这类研究是复杂网络中结点信息识别的基础.其中 SimRank[48]是当前最著名的同一图中结点相似性递归定义方法.在该定义中,对于两个对象而言,如果与其相关的对象相似则两者相似,SimRank(a,b)定义为两个分别从 a 和 b 开始的随机游走遇到某个相同结点的概率.SimRank++[49]是 SimRank 的增强版本,以两个结点公共邻居的数量作为权重定义加权 SimRank.文献[50]利用蒙特卡洛法进行智能随机游走,其中 a 遇到 b 的概率以其邻居的 Jaccard 相似性定义.MatchSim[51]以 a 和 b 点邻居的最大匹配定义 a 和 b 点的相似性.文献[52]利用图的结构来定义两个图中对应结点之间的相似性,利用两个图的邻接矩阵递推地计算两个图中结点间的相似性矩阵.这五种方法中的结点相似性都使用递推方法定义,其相似性计算公式如表 1 所示,其中对于结点 u,$I(u)$表示 u 在图中邻居的集合,$I_i(u)$表示 u 在图中的第 i 个邻居.

表 1　图中结点相似性对比

方法	相似性计算公式
SimRank[48]	$sim_{k+1}(u,v)=\frac{C}{\|I(u)\|\|I(v)\|}\sum_{i=1}^{\|I(u)\|}\sum_{j=1}^{\|I(v)\|}sim_k(I_i(u),I_j(v)),C\in(0,1)$
SimRank++[49]	$sim_{k+1}(u,v)=\frac{C}{\|I(u)\|\|I(v)\|}\sum_{l=1}^{\|I(u)\cap I(v)\|}\frac{1}{2^l}\sum_{i=1}^{\|I(u)\|}\sum_{j=1}^{\|I(v)\|}sim_k(I_i(u),I_j(v)),C\in(0,1)$
文献[50]	$sim_{k+1}(u,v)=c\cdot\left[\frac{\|I(u)\cap I(v)\|}{\|I(u)\cup I(v)\|}\cdot 1+\frac{\|I(u)\backslash I(v)\|}{\|I(u)\cup I(v)\|}\cdot\frac{1}{\|I(u)\backslash I(v)\|\|I(v)\|}\sum_{\substack{u'\in I(u)\backslash I(v)\\ v'\in I(v)}}sim_k(u',v')+\frac{\|I(v)\backslash I(u)\|}{\|I(u)\cup I(v)\|}\cdot\frac{1}{\|I(v)\backslash I(u)\|\|I(u)\|}\sum_{\substack{u'\in I(v)\backslash I(u)\\ v'\in I(u)}}sim_k(u',v')\right]$
MatchSim[51]	$sim_{k+1}(u,v)=\frac{\hat{W}_k(u,v)}{\max(\|I(u)\|,\|I(v)\|)}$,其中 $\hat{W}_k(u,v)$是依据 $sim_k(u,v)$得到的图中 $I(u)$和 $I(v)$的最大匹配的权重
文献[52]	$\boldsymbol{S}_{k+1}=\boldsymbol{B}\boldsymbol{S}_k\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{B}^{\mathrm{T}}\boldsymbol{S}_k\boldsymbol{A}$

基于结点的相似性,一些工作面向具体应用提出基于结点结构相似性的快速实体识别方法.Zhao 等人[53]用入邻居和出邻居的 SimRank 来描述共同引用(入邻居)和论文合作关系(出邻居).文献[12]提出的 GHOST 系统通过共同作者关系构成的社会网络解决文献中作者的同名问题.该方法分为 5 个步骤,首先将合作作者关系建模成为图结构,然后依次经过校验路径选择、相似性计算、名字聚集和用户反馈步骤,实现同名的识别.该系统用到的相似性是两点之间路径长度的倒数和.文献[13]提出实体识别框架 EIF 利用图结构进行实体识别,首先根据结点信息相似性将图中结点分成重叠的类,再根据结点的共同邻居信息确定分类.

文献[54]使用结点之间的关系加速实体识别.在初始化阶段,根据实体之间的关系建立图结构,并使用优先队列保存候选对象集合;在迭代过程中,每一轮从优先队列中取出候选对象对并判定其是否相似,并根据结果更新该优先队列.该工作还针对方法的可扩展性进行了优化.

6　总结与研究展望

由于具有复杂结构数据的广泛存在,其实体识别有着重要应用价值.复杂数据上实体识别的研究方兴未艾,本文介绍了复杂数据上实体识别的问题,并讨论了该问题的应用以及挑战性问题.我们依据面向数据类型对不同复杂数据上的实体识别方法进行了分类介绍.

基于本文的讨论,我们认为复杂数据上的实体识别方面还有如下有待研究的问题,希望对本领域的其他研究者有所启发.

(1)海量复杂数据上的实体识别技术.当前复杂数据上的实体识别技术,特别是图数据上的实体识别技术,主要面向实体识别的有效性提出,着眼于如何有效判定两个数据对象是否描述同一实体,但是面向实体识别效率的工作较少,特别是面向大数据集合上实体识别的工作还比较初级,对大规模图集合上数据实体识别的工作尚未开展研究.但现实应用中有许多海量复杂数据集合,因而需要开展有效率保证的海量复杂数据上实体识别理论与算法的研究.

(2)更新频繁复杂数据上的实体识别技术.当前复杂数据上的实体识别技术假设数据是不发生变化的静态数据,因而可以在其上建立不支持更新的索引,如 signature 等.但是现实中的复杂数据可能更新很频繁,例如互联网上具有结构的信息、社会网

络等都存在着频繁的更新.如果在每次数据发生更新时对数据集合重新进行实体识别,则需要大量的识别时间,难以满足效率要求,因而需要设计更新频繁复杂数据上的增量实体识别技术,为更新的数据确定其所描述的实体.

(3)多类型复杂数据上的实体识别技术.现有的复杂数据上实体识别技术仅考虑同类型数据上的实体识别,而且其主要方法都是基于其结构与内容相似性加以识别.但现实应用中需要对包含多类型复杂数据的数据集加以识别,例如互联网上的信息,包含结构化数据(如隐藏数据库中的数据)、树结构数据(如 HTML 和 XML)、图结构数据(如 RDF)和非结构化数据(如普通文本),对互联网上信息进行有效的查询、集成和分析需要多类型复杂数据的实体识别技术.

(4)复杂数据上实体识别结果的评价.当前尽管有一些复杂数据上实体识别方法提出,但是一个显著的问题是缺少有效的评价方法.当前主要使用准确率和召回率评价实体识别方法.准确率可以相对容易地根据每组实体识别得到的结果进行人工检测,但是召回率的计算需要整个数据集合上实体识别的准确结果,对于海量复杂数据集,人工获取准确结果非常困难.因此设计复杂数据上有效的结果评价方法以及公共测试集合是复杂数据实体识别研究中一个亟待解决的问题.

参 考 文 献

[1] Redman T C. The impact of poor data quality on the typical enterprise. Communications of ACM, 1998, 41(2): 79-82

[2] Miller D W, Yeast J D, Evans R L. Missing prenatal records at a birth center: A communication problem quantified. AMIA Annual Symposium, 2005, 2005: 535-539

[3] Nikki S. Gartner warns firms of "dirty data". Information Management Journal, 2007, 41(3). http://www.allbusiness.com/company-activities-management/operations-quality-control/8901885-1.html

[4] Kohn L T, Corrigan J M, Donaldson M S. To err is human, building a safer health system. Washington, D.C., USA: National Academies Press, 2000

[5] Eckerson W. Data quality and the bottom line: Achieving business success through a commitment to high quality data. The Data Warehousing Institute: Technical Report, 2002. http://download.101com.com/pub/tdwi/Files/DQReport.pdf

[6] English L. Plain English on data quality. Information Management Magazine, 2003. http://www.information-management.com/issues/20030401/6535-1.html

[7] Business Technology. SAS Improves Fraud Protection for HSBC, 2007. http://www.sas.com/offices/NA/canada/en/news/preleases/103107/hsbcfraudprotection.html

[8] Gartner. Gartner says more than 50 percent of data warehouse projects will have limited acceptance or will be failures through 2007. http://www.gartner.com/press_releases/asset_121817_11.html

[9] Shilakes C, Tylman J. Enterprise information portals. Merrill Lynch, 1998

[10] Redman T C. Data: An unfolding quality disaster. Information Management Magazine, 2004. http://www.information-management.com/issues/20040801/1007211-1.html

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

[12] Fan X, Wang J, Pu X, Zhou L, Lv B. On graph-based name disambiguation. Journal of Data and Information Quality, 2011, 2(2): 10

[13] Li L, Wang H, Gao H, Li J. EIF: A framework of effective entity identification//Proceedings of the 11th International Conference of Web-Age Information Management. Jiuzhaigou, China, 2010: 717-728

[14] Kade A M, Heuser C A. Matching XML documents in highly dynamic applications//Proceedings of the 2008 ACM Symposium on Document Engineering. Sao Paulo, Brazil, 2008: 191-198

[15] Puhlmann S, Weis M, Naumann F. XML duplicate detection using sorted neighborhoods//Proceedings of the 10th International Conference on Extending Database Technology. Munich, Germany, 2006: 773-791

[16] Weis M, Naumann F. DogmatiX tracks down duplicates in XML//Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland, USA, 2005: 431-442

[17] Carvalho J C P, Silva A S. Finding similar identities among objects from multiple web sources//Proceedings of the 5th ACM CIKM International Workshop on Web Information and Data Management. New Orleans, Louisiana, USA, 2003: 90-93

[18] Tai K C. The tree-to-tree correction problem. Journal of ACM, 1979, 26(3): 422-433

[19] Milano D, Scannapieco M, Catarci T. Structure aware XML object identification//Proceedings of the 1st International VLDB Workshop on Clean Databases. Seoul, Korea, 2006: 1-8

[20] Leitão L, Calado P, Weis M. Structure-based inference of xml similarity for fuzzy duplicate detection//Proceedings of the 16th ACM Conference on Information and Knowledge Management. Lisbon, Portugal, 2007: 293-302

[21] Leitão L, Calado P. Duplicate detection through structure optimization//Proceedings of the 21th ACM Conference on Information and Knowledge Management. Gloscow, Scotland, 2011: 215-327

[22] Joshi S, Agrawal N, Krishnapuram R, Negi S. A bag of paths model for measuring structural similarity in Web documents//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2003: 577-582

[23] Viyanon W, Madria S K. A system for detecting xml similarity in content and structure using relational database//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 1197-1206

[24] Zhang K, Shasha D. Simple fast algorithms for the editing distance between trees and related problems. SIAM Journal of Computing, 1989, 18(6): 1245-1262

[25] Klein P N. Computing the edit-distance between unrooted ordered trees//Proceedings of the 6th Annual European Symposium on Algorithms. Venice, Italy, 1998: 91-102

[26] Demaine E D, Mozes S, Rossman B, Weimann O. An optimal decomposition algorithm for tree edit distance//Proceedings of the 34th International Colloquium on Automata, Languages and Programming. Wroclaw, Poland, 2007: 146-157

[27] Guha S, Jagadish HV, Koudas N, Srivastava D, Yu T. Approximate xml joins//Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data. Madison, Wisconsin, 2002: 287-298

[28] Yang R, Kalnis P, Tung A K H. Similarity evaluation on tree-structured data//Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland, 2005: 754-765

[29] Kailing K, Kriegel H P, Schonauer S, Seidl T. Efficient similarity search for hierarchical data in large databases//Proceedings of the 9th International Conference on Extending Database Technology. Heraklion, Crete, Greece, 2004: 676-693

[30] Augsten N, Bohlen M H, Gamper J. The *pq*-gram distance between ordered labeled trees. ACM Transaction on Database Systems, 2010, 35(1): 4:1-4:36

[31] Augsten N, Bohlen M H, Gamper J. Approximate matching of hierarchical data using *pq*-grams//Proceedings of the 31st International Conference on Very Large Data Bases. Trondheim, 2005: 301-312

[32] Tatikonda S, Parthasarathy S. Hashing tree-structured data: Methods and applications//Proceedings of the 26th International Conference on Data Engineering. Long Beach, California, USA, 2010: 429-440

[33] Li F, Wang H, Zhang, Hao L, Li J, Gao H. Approximate joins for xml using *g*-string//Proceedings of the 7th International XML Database Symposium Database and XML Technologies. Singapore, 2010: 3-17

[34] Ananthakrishna R, Chaudhuri S, Ganti V. Eliminating fuzzy duplicates in data warehouses//Proceedings of the 28th International Conference on Very Large Data Bases. Hong Kong, China, 2002: 586-597

[35] Weis M, Naumann F. Detecting duplicates in complex XML data//Proceedings of the 22nd International Conference on Data Engineering. Atlanta, GA, USA, 2006: 109

[36] Weis M, Naumann F. Detecting duplicate objects in XML documents//Proceedings of the International Workshop on Information Quality in Information Systems. Paris, France, 2004: 10-19

[37] Zeng Z, Tung A K H, Wang J, Feng J, Zhou L. Comparing stars: on approximating graph edit distance. PVLDB, 2009, 2(1): 25-36

[38] Cho J, Shivakumar N, Garcia-Molina H. Finding replicated Web collections//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas, Texas, USA, 2000: 355-366

[39] Fan W, Li J, Ma S, Wang H, Wu Ying-Hui. Graph homomorphism revisited for graph matching. PVLDB, 2010, 3(1): 1161-1172

[40] Tian Y, Patel JM. TALE: A tool for approximate large graph matching//Proceedings of the 24th International Conference on Data Engineering. Cancún, México, 2008: 963-972

[41] Fan W, Li J, Ma S, Tang N, Wu Y. Graph pattern matching: From intractable to polynomial time. PVLDB, 2010, 3(1): 264-275

[42] Yan X, Yu P S, Han J. Substructure similarity search in graph databases//Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland, USA, 2005: 766-777

[43] Wang X, Smalter A M, Huan J, Lushington G H. G-hash: Towards fast kernel-based similarity search in large graph databases//Proceedings of the 12th International Conference on Extending Database Technology. Saint Petersburg, Russia, 2009: 472-480

[44] Wang X, Huan J, Smalter A M, Lushington G H. Application of kernel functions for accurate similarity search in large chemical databases. BMC Bioinformatics, 2010, 11(S-3): 8

[45] Bharat K, Broder A Z. Mirror, mirror on the Web: A study of host pairs with replicated content. Computer Networks, 1999, 31(11-16): 1579-1590

[46] Bilgic M, Licamele L, Getoor L, Shneiderman B. D-Dupe: An interactive tool for entity resolution in social networks//Proceedings of the 13th International Symposium on Graph Drawing. Limerick, Ireland, 2006: 43-50

[47] Kang H, Sehgal V, Getoor L. GeoDDupe: A novel interface for interactive entity resolution in geospatial data//Proceedings of the 11th International Conference on Information Visualisation. Zürich, Switzerland, 2007: 489-496

[48] Jeh G, Widom J. SimRank: A measure of structural-context similarity//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, 2002: 538-543

[49] Antonellis I, Garcia-Molina H, Chang C C. Simrank++: Query rewriting through link analysis of the click graph. PVLDB, 2008, 1(1): 408-421

[50] Fogaras D, Rácz B. Scaling link-based similarity search// Proceedings of the 14th International Conference on World Wide Web. Chiba, Japan, 2005: 641-650

[51] Lin Z, Lyu M R, King I. MatchSim: A novel neighbor-based similarity measure with maximum neighborhood matching// Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 1613-1616

[52] Blondel V D, Gajardo A, Heymans M, Senellart P, Dooren P V. A measure of similarity between graph vertices. The Computing Research Repository, 2004, cs. DC(0407001): 1-19

[53] Zhao P, Han J, Sun Y. P-Rank: A comprehensive structural similarity measure over information networks//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 553-562

[54] Herschel Melanie, Naumann Felix. Scaling up duplicate detection in graph data//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, California, USA, 2008: 1325-1326

WANG Hong-Zhi, born in 1978, Ph. D., associate professor. His research area is data management, including data quality, XML data management and graph management.

FAN Wen-Fei, born in 1963, Ph. D., professor. His research area is database theory and systems, in particular data quality, data integration, distributed query processing, query languages, recommender systems, social networks, Web services and XML.

Background

This paper aims to provide an overview of recent advances in the study of object identification. Object identification refers to the problem of entity resolution on data with a complex structure. It is to identify objects that pertain to the same real-world entity. While there has been a lot of work on entity resolution, previous work has mostly focused on identifying tuples in relational data that refer to the same real-world entity, also known as record matching. In contrast, object identification has to deal with objects with a tree, DAG or cyclic graph structure. It is far more intriguing than record matching. Indeed, the problem of object identification is challenging because real-life data typically contain errors, have different representations in various data sources, and above all, it is hard to extract the essential properties of an object from its massive and often irrelevant structure. When it comes to complex objects, the study of object identification has raised as many questions as it has answered, from theory to practical techniques. Indeed, many important issues remain to be studied, including effective techniques for object identification on massive complex data, dynamic complex data, and data in multiple complex forms, as well as metrics and methods for evaluating the accuracy and performance of object identification techniques. As it is common to find real-life data with such a complex structure, the need for object identification is evident in data quality management, data fusion, data integration, inconsistency detection and data repairing, among other things. Therefore, to effectively manage complex data, effective object identification techniques have to be developed.

This paper surveys the latest work in this important and practical line of research, namely, object identification techniques for complex objects including XML data, graph data and complex networks. For XML data, we present pairwise entity and group-wise entity resolution methods. For graph data, we focus on methods for determining whether two graphs refer to the same real-world entity. We also survey metrics and methods for identifying vertices that pertain to the same real-world entity in a complex network. Finally, we identify certain open research issues for future work.

This work is supported in part by the National Basic Research Program (973 Program) of China under Grant No. 2012CB316200; National Science Foundation of China under Grant Nos. 61003046, 61133002, 61033015, and 61111130189, Doctoral Fund of Ministry of Education of China under Grant No. 20102302120054. It is also supported in part by the RSE-NSFC Joint Project Scheme.