概　　述

1.1　实体识别问题的提出

大数据时代，数据生成的速度和更新频率远超过去[1]，商业组织、公共部门和政府部门都在面临大量数据的冲击，高效地处理和分析这些数据有助于商业决策、公共政策制定、政府职能提升和国家安全维护。数据管理与数据挖掘是数据研究的核心领域。数据管理聚焦于高效地集成、存储和查询海量数据；数据挖掘则致力于从已有的数据中发掘潜在的信息和价值。

在大规模信息系统和大型的数据挖掘项目中，经常需要将来自多数据源的数据进行集成，提高数据质量，实现数据信息互补，为后续的数据分析与挖掘提供一个完整的、干净的、统一的数据集。集成后的数据集比之前分裂的多个数据集的价值更大，可以从中挖掘出更多的知识，为用户提供更多有价值的信息。在此过程中，一个非常重要的步骤是实体识别[2-13]，即将描述相同真实世界实体的不同数据对象识别出来，从而在数据融合时，能够将描述相同实体的数据对象合并成一个干净的、统一的、健全的数据记录，提高集成数据的质量。

实体识别的直接原因是数据冗余的存在。根据数据源是否单一，可以将数据冗余分为两类：单数据源数据冗余和跨数据源数据冗余。单数据源数据冗余通常由于在加入新的数据记录的时候没有执行严格的重复检测或者完全没有执行重复检测。比如，一个大型商场(在不同城市有分店)的客户信息记录，同一个客户可能进行了多次客户信息登记，而接待人员没有发现这些重复登记。造成这个状况的原因多种多样，如每次登记的姓名有差别、工作单位不同、家庭住址不同等。跨数据源的数据冗余则更加显而易见，当将多个数据集合成一个数据集时，来自不同数据集的数据记录很有可能描述相同的实体。比如，两家公司实行合并后，对它们的客户信息进行整合，需要将他们共同的客户信息找出来。跨数据源的实体识别中，模式匹配是前提。

实体识别中的数据对象(即数据记录)描述真实世界的实体，通常包括多个属性，如姓名、年龄和地址等。这里的数据对象是结构化的，符合一定的数据格式，比如客户信息的数据记录包括姓名属性、年龄属性、电话号码属性、地址属性和工作单位属性。实体识别中最常见的一类数据对象是描述人的数据对象，如商业数据库中的客户记录、公司数据库中的员工记录、航空公司数据库中的乘客记录、医院数据库中的病人记录和医疗保险记录、国家安全部门数据库中的嫌疑犯记录和政府数据库中的纳税人记录等[8]。除了人，还有其他的实体类型，如商业记录、出版记录、引文记录、产品记录等。例如，在商品比价应用中，由于不同电商网站的描述格式不同，识别出哪些商品记录描述着相同的商品有一定难度；还有引文记录中的会议(或出版社)全称和缩写的识别、作者单位全称与简称的识别等[8]。

1.2 实体识别研究的发展历史

实体识别起源于统计学家和公共健康研究领域，在单数据库内或多数据库中识别对应同一实体的重复记录。1946 年，Dunn 应用术语“记

录链接”（record linkage）[14]来描述现实世界中每一个个体的生命溯源，即从生到死整个生命周期中个体所经历的信息，如健康信息、社会保障、结婚、离婚等记录信息。20 世纪 50 年代末和 60 年代初，Howard Newcombe 等[15-16]提出应用计算机自动处理实体识别过程，并提出了基于概率的记录链接方法的成功理念。基于 Newcombe 的思想，在 1969 年，两个统计学家 Ivan Fellegi 和 Alan Sunter[17]为实体识别引入了正式的数学模型。

1999 年，由学者 Winkler[18]扩展并提高了最初的模型，最显著的工作是引入了字符串近似比较函数[19]来捕捉字符串的变化情况，以及应用期望(EM)算法[20]来改进概率记录链接中匹配参数的估计。同时，数据库研究团队从数据清洗需求出发，提出了重复记录识别技术[21]，用于改进数据库的质量[22]。但是，数据库研究者并没有采用由 Fellegi 和 Sunter 提出的基于概率的匹配方法，而是应用近似串比较函数计算属性相似度[23-24]，并通过属性比较来发现相似的记录[21,25]。

随着数据的丰富，计算机领域中有关实体识别的研究备受关注，尤其是在数据挖掘、机器学习和信息获取领域。此外，数据库和数据仓库研究团队[26]相应地也提出了一些新的实体识别技术[27]，如利用机器学习、自然语言处理和基于图的方法来改进数据质量。除此之外，近些年来还呈现出了面向时间记录的实体识别[28-29]，改善具有时间演化特性的同一实体的识别准确性；基于众包的实体识别[30-31]，通过混合人机来提升实体识别的准确性。同时，隐私保护下的实体识别[32,33]也成为了关注热点，以支持隐私数据的实体识别。

根据识别对象的数据源的种类划分，已有的实体识别工作主要包括：在关系数据库、Deep Web 数据库上的实体(记录)识别[34]；Web 上的实体识别[35-36]；语义 Web(RDF 数据)上的实体识别[37-38]；数据仓库中的实体识别[39-40]]；非结构化文档中的实体识别[41]；复杂数据如 XML 数据、图数据、复杂网络上的实体识别[42]；社会网络中的实体识别[43-44]。

由于实体识别一直被各个领域从不同的方面研究，包括统计学领域、信息检索领域、人工智能领域、机器学习领域、数据库领域和工业界等，各种方法尤其是结构化数据上的识别方法相继被提出。在统计学和人工智能领域，已有的研究主要是把它看成一种分类问题，主要是基于统计和机器学习的方法(监督的方法和非监督的方法)。而在数据库领域，已有的方法通常是使用基于规则的方法。在不同的文献和研究领域中实体识别的英文名称有好多种，如 entity resolution、entity matching、fuzzy matching、fuzzy join、fuzzy duplicate elimination、approximate join、approximate string join、approximate matching、record linkage、merge/purge、identity uncertainty、duplicate identification、duplicate detection、record deduplication、coreference resolution、reference reconciliation、object identification 和 object matching 等，相应地中文名称也有很多，如重复探测、记录链接、对象区分、引用区分、引用协调、对象统一和实体统一等。

1.3 实体识别问题的描述

随着实体识别研究的深入，研究者对实体识别问题具有了统一认识，即实体识别是识别出对应同一真实世界实体的重复数据对象。通常，将一个真实世界的实体记作 χ，将一条描述实体的数据对象记作 r。一个数据对象通常包含多个属性，比如说，一个人可以通过姓名、生日、性别、婚姻状态、电话号码和地址等来描述。在“脏”数据集中，可能会存在多个数据对象描述同一实体；由于书写方式的多样性和拼写错误，描述相同实体的多个数据对象不一定字面上完全相同。如果两个数据对象描述相同的实体，那么这两个数据对象是重复的或冗余的或匹配的。实体识别问题的正式定义如下。

定义 1.1(实体识别) 给定一个脏数据集 $R=\{r\}$，实体识别就是利用一个数据对象-实体映射函数 $\varphi(r)=\chi$ 来确定描述同一实体的所有

数据对象，记作 $\Phi(R)=\{\{r|\varphi(r)=\chi\}|r\in R,\chi\in X\}$。其中，$X$ 是一个真实世界实体的集合，该集合并不需要是已知的。数据对象-实体映射函数 $\varphi(r)=\chi$ 将一个数据对象 r 映射到它所描述的实体 χ。然而，实体 χ 并不需要是已知的，实体识别只需要知道哪些数据对象对应相同的实体。比如，$\varphi(r_1)=\varphi(r_2)$，那么数据对象 r_1 和 r_2 描述相同的真实世界实体。

例如，如表 1-1 所示，数据集 $R=\{r1, r2, \cdots, r10\}$，$X=\{e1, e2\}$，则 $\Phi(R)=\{\{r1, r2, r3, r4, r5\}, \{r6, r7, r8, r9, r10\}\}$，$\varphi(r1)=\varphi(r2)=\varphi(r3)=\varphi(r4)=\varphi(r5)=e1$，$\varphi(r6)=\varphi(r7)=\varphi(r8)=\varphi(r9)=\varphi(r10)=e2$。

表 1-1 DBLP 中文章作者信息片段

eid	rid	Name	affiliation	Co-authors	year
*e*1	*r*1	Xin Dong	Univ of Washington	Halevy, Tatarinov	2004
*e*1	*r*2	Xin Dong	Univ of Washington	Halevy	2005
*e*1	*r*3	Xin Luna Dong	Univ of Washington	Halevy, Yu	2007
*e*1	*r*4	Xin Luna Dong	AT & T Labs-Research	Das Sarma, Halevy	2009
*e*1	*r*5	Xin Luna Dong	AT & T Labs-Research	Naumaunn	2010
*e*2	*r*6	Dong Xin	Univ of Illinois	Han, Wah	2004
*e*2	*r*7	Dong Xin	Univ of Illinois	Wah	2007
*e*2	*r*8	Dong Xin	Microsoft Research	Wu, Han	2008
*e*2	*r*9	Dong Xin	Microsoft Research	Chaudhuri, Ganti	2009
*e*2	*r*10	Dong Xin	Microsoft Research	Ganti	2010

需要指出的是，实体识别不同于自然语言处理中的命名实体识别(Named Entity Recognition，NER)和实体链接(Entity Linking)。命名实体识别是信息抽取的一个子任务，是指判断出文本中命名实体属于哪一类(提前定义好的)命名实体；实体链接是将文本中有歧义的实体指称项链接到给定的知识库中，从而实现实体歧义的消除。

1.4 实体识别的处理流程

如图 1-1 所示，实体识别主要包括三个步骤：数据分块、数据对象相似度计算和数据对象对匹配决定。

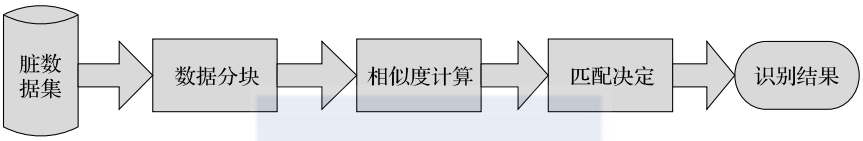

图 1-1 实体识别流程

首先，数据分块也称为数据索引，用于缩小搜索空间，减少无用的数据对象比较，提升识别速度。典型的数据分块技术有基于键值的分块方法、基于滑动窗口的分块方法、基于聚类的分块方法、基于值的分块方法等。数据分块是一个可选步骤。

其次，实体识别的一个重要环节是计算数据对象之间的相似度，如果一个数据对象对的相似度越大，该数据对象对匹配的可能性越大；相似度计算要用到相似度计算函数，具体见第 2 章。

最后，当获得了数据对象相似度之后，需要利用数据对象相似度来决定数据对象之间是否匹配(重复)，当前已有多种匹配决定的方法，典型的有基于阈值的决定方法、基于分类的决定方法和基于聚类的决定方法。

1.5 实体识别的挑战

在单个数据源或多个数据源中，将描述相同实体的不同数据对象识别出来存在一些挑战。接下来将介绍这些挑战。

1.5.1 相似度衡量问题

通常来说，待匹配的数据集之间不存在统一标识符(即 ID)，比如身份证号、社保号、商品唯一编号等。如果存在这样的统一标识符，那么实体识别将变成数据库连接操作。然而，现实世界的数据集中很少包含统一标识符，因此，为了进行实体识别，需要衡量数据对象的相似性。实体识别中的数据集通常都不是高质量的，它们可能包含错误属性、相同属性的不同表示形式、随着时间改变的属性(如年龄、联系电话)等。鉴于上述原因，实体识别中比较数据对象的属性时，不能采用精确的相似度比较函数(即返回布尔型结果)，而需要采用能衡量出属性值有多相似的相似度函数(返回介于 0 和 1 之间的数值，越接近于 1 代表越相似，反之则越不相似)。针对不同类型的属性，相似度衡量的方法应该不同。比如姓名经常存在不同的书写格式和不同的缩写形式，文章的标题通常是一个较长的字符串，年份、价格等数值型属性的相似度衡量不同于字符串等[45]。本书第 2 章将介绍不同的相似度函数，可用于不同的属性比较，从而解决不同的实体识别任务。

1.5.2 计算效率问题

给定两个数据集，其数据规模分别为 m 和 n，那么实体识别的比较次数为 $m \times n$。对于单数据源的情况，给定一个数据规模为 l 的数据集，那么需要进行 $l(l-1)/2$ 次的比较。通过上述分析可以发现，这种方式的实体识别随着数据规模的增长，其计算开销以平方级的方式增长。当数据量比较大的时候，这样大的开销是无法接受的。为此，应该快速、有效地去除掉不可能匹配的数据对象对，只保留那些有可能匹配的数据对象对。实体识别中分块技术的作用就是降低计算开销，通过分块技术只将可能匹配的数据对象分到相同的块中。本书第 3 章将介绍分块技术。

1.5.3 机器学习方法的应用问题

通过机器学习方法可以训练出实体识别模型的参数值和匹配规则，以避免人工确定字段权重和匹配阈值。基于机器学习的实体识别方法[49-54]主要可分为两类：基于分类器的实体识别方法和基于概率图模型的实体识别方法。基于分类器的实体识别方法将实体识别看作一个分类问题，即给定两个数据对象，判断两者是否匹配。一般情况下，这些待匹配的数据对象被看作独立且均匀分布的。常见的分类方法有决策树、贝叶斯分类器、支持向量机、主动学习、误差逆传播、遗传编程等。与基于分类器的实体识别方法不同，基于概率图模型的实体识别方法认为对象之间并非孤立，而是存在某种内在联系，利用这种内在联系可以避免对实体的孤立式匹配决策。这类方法将实体之间的内在联系表达为概率图模型，通过推理和学习来实现联合式实体识别，具体包括基于马尔可夫逻辑网络的实体识别和基于条件随机场的实体识别。为了达到较好的效果，需要选择和建立合适的机器学习模型。本书第 4 章将介绍基于机器学习的实体识别方法。

1.5.4 关联对象的识别问题

现实世界中存在很多关联的数据，分为多类型关联数据和单类型关联数据。多类型关联数据包含多种类型的数据对象，彼此之间存在一定的关联关系，比如引文数据集中包括文章、作者、会议等，电影数据集中包括电影、导演、演员、出品公司等。由于不同数据对象之间存在依赖关系，当识别出一些数据对象后，与这些数据对象关联的其他数据对象的相似性会变大，从而实现了相似性传播[55,56]。可见，利用这种不同类型的数据对象之间的关联关系，可以实现更准确的实体识别。单类型的关联数据只包含一种相互关联的数据对象，比如社交网络和学术合作网络等。另外，实体识别中的特殊情况是不同实体同名(称为实体消歧或名字消歧)，比如 DBLP 数据库中有超过 50 个名叫 Wei Wang 的作

者[57]。在单类型的关联数据中，描述不同实体的数据对象与其他数据对象的关联强度不同，这些特性有助于解决实体消歧的问题。本书第5章将介绍基于关系的实体识别。

1.5.5 一些新的挑战

一些待识别的数据集的数据对象包含了时间戳或一些与时间相关的属性，这些属性描述了某个特定时间上实体的特征。比如，DBLP数据库有作者的相关信息：作者姓名、工作单位、合作者和年份，作者姓名、工作单位和合作者等属性都可能随着时间推移而发生演化，例如，姓名变化如Xin Luna Dong→Luna Dong；工作单位变更如University of Washingtion→Google，合作者变化如Halevy，Yu→Naumamn。上例中，同一属性在不同时间可能取不同的值，也就是说单纯的属性相似度无法准确反映数据对象的相似性。在识别这类数据时，需要考虑属性的演化特性[28,29]。如何利用时序信息和属性的演化信息来帮助实体识别是一个有意义的研究问题。

一些实体识别任务的难度非常大，单靠计算机的算法本身无法解决，比如涉及图片的实体识别。近年来，众包(Crowdsource)逐渐受到工业界和学术界的重视。众包就是借助互联网上大众的智力来解决一些计算机不能独立完成的任务。就实体识别而言，人通常可以比计算机更准确地判断两个数据对象是否描述相同的实体。如何利用众包来帮助计算机进行实体识别是一个新的研究问题[30,31]。

实体识别在金融、医疗、政府等领域具有广泛的应用。但是，当数据对象涉及个人隐私或敏感信息时，必须要考虑数据对象的隐私保护问题。例如，在分散的医疗体系中，某人的医疗信息可能分布在多个医院，找出同一个人在不同医院的诊断信息有利于更准确地分析病情，但由于涉及患者隐私，各医院并不希望暴露患者的医疗信息。在这种情况下，实体识别方法应当既找出某位患者在各医院的医疗信息，又保证各医院其他患者的医疗信息不被泄露[32,33]。如何在隐私保护的前提下进行

实体识别也是一个重要的研究问题。

本书第 6 章将介绍基于众包的实体识别、基于时间模型的实体识别和隐私保护下的实体识别。

1.5.6 实体识别评估

如何合理地评估识别结果的精确性和实体识别效率对于实体识别研究非常关键。同时，要评估实体识别方法，需要已知真实结果的数据集，如何获取有公信力的数据集也是实体识别评估的一个重要方面。本书第 7 章将介绍实体识别评估。

1.6 实体识别的应用

实体识别有着广泛的应用，如医疗卫生、人口普查、客户关系管理、网购比价、商业欺诈侦查、关联的开放数据和引文数据库等。实体识别对于大型的信息系统、政府部门、公共部门、商业组织和研究机构等都有着重要作用。

1.6.1 医疗卫生

医疗卫生是实体识别最早的应用领域之一，已经有几十年的应用历史，它也是推动实体识别研究的重要领域。一个人的一生当中会产生大量的医疗记录，这些医疗记录可能来自医生、医院、体检中心和医疗保险公司等。如果能将这些医疗记录匹配起来，就形成了一张医疗图谱。如果能将很多人的医疗图谱构建出来，对于医药研究的意义非凡。匹配的医疗数据可以用于新的医疗研究，减少新的数据获取的开销。比如，可以用于调查特定病人群体中的药物副作用反应。

英国的牛津记录链接研究开始于20世纪60年代，该项目致力于研发基于计算机的记录链接技术，并将该技术应用于大约350 000人的出生、死亡和医疗数据。这样就可以研究特定的疾病之间的关联关系，利用纵向的匹配数据可以分析不同职业的死亡率、移民情况和相关的社会经济学的因素。从20世纪90年代中期开始，澳大利亚进行了一个比较成功的医疗数据匹配项目。该项目将来自不同数据源的医疗记录以及一些非医疗记录匹配起来，为每个个体形成了一系列的匹配记录。从1995年到2003年，这个项目共输出700多项成果，其中包括一些卓越的成果：医疗政策的改革和诊所条例的更改。一些其他国家也实施了类似的医疗数据匹配项目。然而，医疗数据涉及病人的隐私信息，因此在进行实体识别时需要考虑隐私保护问题。

1.6.2 人口普查

人口普查是世界各国的一项基本公共事务。人口普查的数据涉及人口、文化、经济和环境等方面的信息。这些信息可以生成各种各样的统计报告，这些报告可以帮助政府和商业组织规划资金和资源的配置。实体识别技术是人口数据统计的一个重要工具。实体识别有助于复用已有数据集来编辑新的人口数据，从而减少管理大型人口数据集的开销。同时，在实体识别过程中，可以发现和纠正信息冲突和弥补信息缺失等，有助于提高数据质量和完整性。利用实体识别可以将不同年份获得的人口普查数据集匹配起来，从而生成纵向数据集，为公共及政府部门提供刻画人口的各种特征及其随时间的变化情况。在不同的国家，有不同的法律和法规规定什么样的数据可以进行实体识别。比如在澳大利亚，人口普查数据中的姓名和地址信息必须在收集后的一年内销毁。诸如此类的限制为构建纵向的数据带来了巨大的挑战，因为实体识别通常要依赖姓名、年龄、性别、宗教、职业、住址和工作单位等。美国人口普查局是最早采用实体识别技术的组织之一，同时该组织在较长的时期内都在实体识别研究中处于领先地位。美国人口普查局要处理的数据量在亿级，

因此该组织很早就提出针对大规模数据的分布式实体识别技术。

1.6.3 客户关系管理

大型商业组织通常会以不同的形式收集客户的相关信息。大型商业组织通常可能有不同的分支机构，如网上商店、在不同城市的实体店、售后服务机构、VIP 专项服务部门和广告推送部门等。每个分支机构都可能与客户产生联系，生成一些客户信息记录，并存储在各自的本地数据库中。这样一来，某位客户的信息可能会保存在该商业组织的不同分支机构的数据库中，导致不同分支机构的数据库中可能包含重复的客户记录。商业组织通常采用客户关系管理系统来管理海量客户的诸多信息。当客户关系管理系统收集客户的全部信息时，需要识别出描述相同客户的所有客户记录，以精准推送产品和服务广告，同时也避免因存在重复的客户记录而导致的资源浪费现象。

客户数据的实体识别面临以下挑战。

1）当人们换了住址后，他们的地址发生了变化；当人们结婚或离婚后(欧美国家)，他们的姓氏会发生变化。这些情况会导致商业数据库中出现重复地描述相同客户的记录。

2）大多数的客户并不关心商业数据库中是否存在多条描述他们的重复记录，他们只关心自己购买的产品或服务是否到货。即便收到多份来自相同商业组织的推送广告，他们也很少会主动报告。

3）当多个商业组织要进行合作时，比如要进行联合商业推广，需要将各种商业组织数据库中的客户记录集成在一起，构建一个统一的客户信息数据库，在此过程中需要利用实体识别技术将重复的客户记录找出来。一般来说，在不同商业组织的数据库中，客户记录的格式是不同的，可能存储了不同类型的信息，客户记录生成的时间也可能不同，这些情况都给客户记录的识别造成了极大的困难。

1.6.4 网购比价

随着互联网技术和电子商务的发展，越来越多的人习惯于从电商网站购买商品和服务。电子商务的繁荣催生了一个新的服务——网购比价。网购比价网站支持用户查询特定的商品或根据分类、价格或品牌来浏览商品。网购比价服务中面临的一个重大挑战是，如何确定来自不同电商网站的商品条目描述的是相同的商品。某些特定类型的商品有唯一标识符，比如图书有 ISBN 码、电子产品有 EPC 码。然而，大多数的商品(如衣服、家电、日用品等)并不存在唯一标识符，因此，这些商品在不同的电商网站的产品描述信息大不相同。如表 1-2 所示，四个来自不同电商网站的商品条目描述了相同的商品，然而，不仅它们的商品描述差异明显，而且它们的商品代码也十分不同。为了保证提供精确而全面的商品比价服务，比价网站必须准确找出所有描述相同商品的商品条目。相对于其他的实体识别应用，如人的姓名和地址等，商品条目的商品描述差异更大，因此需要采用完全不同的相似度计算函数。例如，表 1-2中四个商品条目中商品描述的字符串相似性是比较明显的。然而，这个相机的商品名称与它的上一代(Canon PowerShot G10)的商品名称只差一位数字，即 0 和 1。

表 1-2 来自不同电商网站描述相同商品的四个商品条目

商品描述	商品代码
Canon PowerShot G11 10 MP Compact Camera-6.10mm-30.50mm	Item # 927909
Canon-PowerShot G-11 10.0 Mega pixel	Item# CANPSG11
PowerShot G11 Point & Shoot Digital Camera	Canon 3632B001
Canon PowerShot G11 10 Megapixel Compact Camera	MFG #：3632B001

1.6.5 犯罪及欺诈侦查

实体识别技术是犯罪侦查信息系统的重要组成部分。利用复杂的信

息系统，警察部门可以准确地确定嫌疑犯的身份。犯罪侦查领域中的实体识别不同于其他领域的实体识别，它面临的一个重大挑战是，犯罪嫌疑人会蓄意修改个人信息，从而避免被识别出来。如当面对执法人员的询问时，罪犯通常会提供篡改过的或虚构的个人信息，如地址、出生日期、虚构的社保号或驾驶证号码等。犯罪嫌疑人的蓄意修改使得修改后的个人信息与真实的个人信息(极有可能是另外一个人)看起来十分相像，从而误导执法部门的侦查。实体识别技术可以帮助侦察人员确定虚假的个人信息是否对应一个真实的人。通过对比已有的罪犯数据库和待确定的个人信息记录，侦察人员可以确定当前嫌疑犯的真实身份。通常来说，在这种应用中，对实体识别的执行效率有较高的要求，因为嫌疑犯的身份需要尽快确定。

随着商业模式的多样化和互联网商业的发展，身份欺诈造成的财产损失越来越多。实体识别技术可以帮助身份验证，减少身份欺诈。身份欺诈的数量在各个国家都在不断增长，给金融组织带来了数以亿级的财产损失，同时带来了恶劣的社会影响。身份欺诈是指行骗者通过虚假的身份获取了服务和收益的权限。随着电子财务交易和在线公共和政府服务的广泛应用，这些服务和交易的参与者的身份验证变得非常重要。通过对美国的 3 亿银行账号进行统计发现，90%的欺诈账号是通过虚构的身份开通的，75%的银行财产损失是由虚构身份的欺诈造成的。实体识别是身份验证系统的关键组成部分。具体来说，就是将待验证的身份信息与各种包含已验证的、准确的个人记录进行比较，判断是否真实存在此人或是否是某个人。已验证的数据库包括选择投票注册数据库、驾照数据库、社保数据库、电话登记数据库等。通过与这些已验证的数据库进行匹配，就可以得到全面真实的个人信息，从而评估当前的身份信息是否为真。

1.6.6 关联的开放数据

随着互联网的发展，Web 上的数据越来越多，然而这些数据并没有

有效地集成，也无法准确地查询。语义网(Semantic Web)的研究者提出了关联的开放数据(Linked Open Data，LOD)的概念。LOD 的基本思路是：赋予每个数据对象一个 URI，用 HTTP 协议将数据对象关联起来(通常是 RDF 三元组的形式)，最后将这些数据公开发布在 Web 上。LOD 的目标是将来自各个数据源的 Web 数据集成起来，以便于用户快速查询和浏览。越来越多的数据源，如 DBpedia、Freebase 等加入到 LOD 项目中，使得用户可以方便地访问海量的信息。截至 2014 年 8 月，基于关联开放数据项目组织(LOD)发布的数据云图统计，已有约 570 个数据集，这些数据集之间通过 2909 个 RDF 链接。这些 RDF 链接中，有很大比例是链接两个匹配的数据对象，比如，Freebase 中描述美国银行的数据对象与维基百科中描述美国银行的数据对象之间存在一个链接。LOD 项目中的数据源涉及各个领域，包括公共服务领域、商业领域和科研领域等。这些数据源间经常会存在一定的交叠，这使得实体识别在 LOD 项目中不可或缺。最初的链接操作是人工进行的，今后将逐步地发展基于机器的链接操作。由于待链接的数据集常常是异构的，包含大文本属性，因此需要提出新的、特定的相似度函数。另外，由于数据量的巨大，实体识别算法的效率变得非常关键，需要提出快速的实体识别算法。

1.6.7 引文数据库

随着信息技术的发展，科研成果的发表逐渐电子化，大多数成果都提供了在线数据库的访问服务，比如 Springer、Elsevier、ACM Digital Library 和 IEEE Xplore 等。这些在线服务给科研人员带来了巨大的方便，使得科研人员在任何联网的地方都可以访问到海量的学术成果。这样的在线数据库被称为引文数据库。另外，一些在线机构(如 Thompson Web of Knowledge)提供文献引用和影响因子分析等服务。引文数据库不仅加快了新的科研成果的传播速度，而且它对于科研资助基金的分配有着重大影响——越来越多的科研资助基金机构(包括政府的和企业的等)基于权威的引文数据库来分析和评价科研人员、研究小组和研究机

构的学术成果和学术影响力。这样的学术评价方式将用于科研资助基金的分配以及科研人员的晋升。个人学术评价指标，如 h-index，根据科研人员的学术成果的引用情况来计算出一个数值，以此来评价科研人员的学术影响力。鉴于引文数据库对于学术成果传播和学术成果评价有着巨大的意义，引文数据库中的数据必须是高质量的，否则基于这些数据生成的学术评价指标将没有公信力。

引文数据是不断增长的，构建和维护引文数据库存在诸多挑战。一些大型的引文数据库包括了超过 25 000 000 条文献，维护起来十分困难。构建引文数据库的最大挑战是，数据库中很多作者的姓是一样的，名的缩写也是一样，甚至有很多这样的作者在同一个研究领域工作。即便提供了作者姓名的全称，也经常难以判断两篇文章是否是由同一作者发表。学术期刊和学术会议中的名字通常是缩写形式，而不是标准的全称，因此会导致同一个姓名的多种表达形式的出现。如表 1-3 所示，三条引文记录描述的是同一篇 VLDB 1994 学术会议上发表的文章，然而它们的字面表达有明显的差别。

表 1-3　描述同一篇学术论文的三条引文记录

引文记录
R. Agrawal, R. Srikant. Fast algorithms for mining association rules in large databases. In VLDB-94, 1994.
Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994.
Agrawal R., Srikant R. Fast algorithms for mining association rules in large databases. In VLDB Conference, 1994.

引文数据的一些特征使得它成为数据挖掘的热点研究对象。引文数据本身是公开发表的信息，因此在进行实体识别时不涉及隐私保护的问题。引文数据本身包括了多种类型的数据，如文章、作者、会议或期刊和作者工作单位等。利用这些多类型的数据对象，可以构建出一个异构的数据对象网络。在这个网络中，某一类型数据对象的匹配可以促进与之关联的其他数据对象的匹配。比如，两个作者记录的姓以及名的缩写相同，如果两者在同一大学工作，那么两者匹配的可能性比不在同一大

学工作的概率大。这样多类型同时识别称为联合式实体识别。

1.7 本章小结

本章从全局的角度介绍了实体识别技术的发展过程，给出了实体识别的问题描述以及实体识别的处理流程，详细地介绍了实体识别过程中涉及的挑战问题和实体识别技术在各个领域中的应用情况。

参考文献

[1] Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think[M]. Boston: Houghton Mifflin Harcourt, 2013.

[2] Ganti V, Sarma A D. Data cleaning: A practical perspective[J]. Synthesis Lectures on Data Management, 2013, 5(3): 1-85.

[3] Koudas N, Sarawagi S, Srivastava D. Record linkage: Similarity measures and algorithms[C]. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, 2006: 802-803.

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

[5] Köpcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems[C]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 484-493.

[6] 郭志懋，周傲英．数据质量和数据清洗研究综述[J]. 软件学报，2002，13(11)：2076-2082.

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

[8] Naumann F, Herschel M. An introduction to duplicate detection[J]. Synthesis

Lectures on Data Management, 2010, 2(1): 1-87.

[9] Thor A, Rahm E. MOMA - A mapping-based object matching system[C]. Proceedings of the Third Biennial Conference on Innovative Data Systems Research, 2007: 247-258.

[10] Singla P, Domingos P. Entity resolution with markov logic [C]. Proceedings of the Sixth International Conference on Data Mining, 2006: 572-582.

[11] Wellner A M B. Conditional models of identity uncertainty with application to noun coreference[C]. Proceedings of the 2004 Annual Conference on Neural Information Processing Systems , 2004: 905-912.

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

[13] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning[C]. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002: 269-278.

[14] Dunn H. Record linkage[J]. American Journal of Public Health, 1946, 36 (12): 1412.

[15] Newcombe H, Kennedy J. Record linkage: making maximum use of the discriminating power of identifying information[J]. Communications of the ACM 1962,5(11), 563-566.

[16] Newcombe H, et al. Automatic linkage of vital records[J]. Science, 1959, 130 (3381): 954-959.

[17] Fellegi I P, Sunter A B. A theory for record linkage[J]. Journal of the American Statistical Association, 1969, 64(328): 1183-1210.

[18] Winkler W E, Thibaudeau Y. An application of the Fellegi-Sunter model of record linkage to the 1990 U. S. decennial census[R]. Tech. Rep. RR1991/09, US Bureau of the Census, Washington, DC, 1991.

[19] Porter E H, Winkler W E. Approximate string comparison and its effect on an advanced record linkage system[R]. Tech. Rep. RR97/02, US Bureau of the Census, 1997.

[20] Winkler W E. Using the EM algorithm for weight computation in the Fellegi-Sunter model of recordlinkage[R]. Tech. Rep. RR2000/05, US Bureau of the

Census, Washington, DC, 2000.

[21] Hernandez M A, Stolfo S J. The merge/purge problem for large databases[C]. ACM SIGMOD, 1995:127-138.

[22] Hernandez M A, Stolfo S J. Real-world data is dirty: Data cleansing and the merge/purgeproblem[J]. Data Mining and Knowledge Discovery, 1998, 2(1): 9-37.

[23] Rahm E, Do H H. Data cleaning: Problems and current approaches[J]. IEEE Data Engineering Bulletin, 2000, 23(4): 3-13.

[24] Monge A E. Matching algorithms within a duplicate detection system[J]. IEEE Data Engineering Bulletin, 2000, 23(4): 14-20.

[25] Monge A E, Elkan C P. The field-matching problem: Algorithm and applications[J]. ACM SIGKDD, 1996:267-270.

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

[27] Winkler W E. Overview of record linkage and current research directions[R]. Tech. Rep. RR2006/02, US Bureau of the Census, Washington, DC, 2006.

[28] Pei Li, Xin Luna Dong. Linking Temporal Records[C]. PVLDB, 2011, 4(11): 956-967.

[29] Y H Chiang, A Doan, J F Naughton. Modeling entity evolution for temporal record matching[C]. SIGMOD, 2014.

[30] 冯剑红，李国良，冯建华．众包技术研究综述[J]. 计算机学报，2015，38(9)：1713-1726.

[31] Jiannan Wang, Tim Kraska, et al. CrowdER: Crowdsourcing Entity Resolution[J]. PVLDB2012, 5(11): 1483-1494.

[32] V S Verykios, A Karakasidis, V Mitrogiannis. Privacy preserving record linkage approaches[J]. International Journal of Data Mining, Modelling and Management, 2009, 1(2): 206-221.

[33] D Vatsalan, P Christen, V S Verykios. A taxonomy of privacy-preserving record linkage techniques[J]. Inf. Systems, 2013, 38(6): 946-969.

[34] 寇月，申德荣，李冬等．一种基于语义及统计分析的 Deep Web 实体识别机制[J]. 软件学报，2008，19(2)：194-208.

[35] Ergin E, Min-Yen K, Lee D W, et al. Web based linkage[C]. In Proc. of WIDM, 2007:21-128.

[36] Kopcke A, Thor E, Rahm. Learning-based approaches for matching web data entities[J]. IEEE Internet Computing, 2010, 14(4): 23-31.

[37] W Hu, J F Chen, Y Z Qu. A Self-training approach for resolving object coreference on the semantic Web[C]. In Proc. of WWW ,2011:87-96.

[38] F Sais, N Pernelle, M Rousset. L2R: a logical method for reference reconciliation[C]. In Proc. of AAAI, 2007:329-334.

[39] Rohit A, Surajit C, V. Ganti. Eliminating fuzzy duplicates in data warehouses [C]. In Proc. of VLDB, 2002:586-597.

[40] Surajit C,Kris Ganjam, V Ganti, et al. Robust and efficient fuzzy match for online data cleaning[C]. In Proc. of SIGMOD, 2003:313-324.

[41] Surajit C,V Ganti, Dong Xin. Mining document collections to facilitate accurate approximate entity matching[J]. PVLDB, 2009,2(1): 395-406.

[42] Melanie Weis, Felix Naumann, et al. Industry-scale duplicate detection[C]. In Proc. of VLDB, 2008:1253-1264.

[43] M Bilgic, L Licamele, L Getoor, et al. D-Dupe: an interactive tool for entitiy resolution in social networks[C]. In Symposium on Graph Drawing, 2005:505-507.

[44] Malin. Unsupervised name disambiguation via social network similarity[C]. In Proc. of SIAM international conference on data mining(SDM),2005.

[45] Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records [C]. Proceedings of the Kdd Workshop on Data Cleaning and Object Consolidation, 2003: 73-78.

[46] Christen P. A survey of indexing techniques for scalable record linkage and deduplication[C]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(9): 1537-1555.

[47] Hernández M A, Stolfo S J. The merge/purge problem for large databases[C]. ACM Sigmod Record, 1995, 24(2): 127-138.

[48] Hernández M A, Stolfo S J. Real-world data is dirty: Data cleansing and the merge/purge problem[J]. Data Mining and Knowledge Discovery, 1998, 2(1): 9-37.

[49] S Tejada, C Knoblock, S Minton. Learning object identification rules for information integration[J]. Information Systems, 2001, 26(8): 607-633.

[50] Elfeky M G, Verykios V S, Elmagarmid A K. Tailor: A record linkage toolbox [C]. In Proceedings of 18th International Conference on Data Engineering, 2002:17-28.

[51] W E Winkler. Methods for record linkage and bayesian networks[R]. Technical report, Series RRS2002/05, U. S. Bureau of the Census, 2002.

[52] William E Winkler. The state of record linkage and current research problems [R]. Technical report, Statistical Research Division, U. S. Census Bureau, 1999.

[53] M Bilenko, R Mooney. Adaptive duplicate detection using learnable string similarity measures[C]. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003: 39-48.

[54] S Sarawagi, A Bhamidipaty. Interactive Deduplication Using Active Learning [C]. Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002: 269-278.

[55] Bhattacharya I, Getoor L. Deduplication and Group Detection using Links[C]. LinkKDD, 2004.

[56] Dong X, Halevy A, Madhavan J. Reference Reconciliation in Complex Information Spaces[C]. SIGMOD, 2005.

[57] Kalashnikov D, Mehrotra S. Domain-independent data cleaning via analysis of entity-relationship graph[C]. TODS, 2006.