附件2：

****

**毕业设计（论文）开题报告**

**题 目：面向大规模实体识别的自动特征工程**

**专 业 数据科学与大数据技术**

**学 生 唐亚锋**

**学 号 1190201313**

**指导教师 王宏志**

**日 期 2022年10月21日**

**哈尔滨工业大学教务处制**

**说 明**

**一、开题报告主要内容**

1．课题来源及研究的目的和意义

**1.1研究的目的及意义**

随着互联网时代的发展，数据量得到爆炸性增长。为了提升数据质量，增强其可用性，数据治理愈加重要。在数据治理过程中，不同的数据库和结构化信息源之间往往需要进行交互和相互操作，这需要为用户提供一个完整的集成视图。提供这样的集成视图服务需要在多个应用程序和组织之间共享数据，并为一组自治和异构的数据源提供一个统一的访问入口，具有以上功能的系统被称作数据集成系统[1]。

数据集成系统需要处理来自一组数据源的数据，由于数据格式的记录方式的不同，同一现实世界的实体往往在不同数据源中有不同的表述方式，且由于语义的歧义性，同一表述也可能描述不同的真实世界实体。因此，实体识别任务旨在将有着共指关系的多个表述（co-referent manifestations）按照“粗粒度分块、细粒度匹配”的流程识别为同一个真实世界实体，是数据集成的关键操作，也是数据清洗[2]等下游任务的基础。

实体识别任务中，保证匹配的准确率非常重要。准确地进行实体识别往往比较困难，主要原因在于命名习惯不同、使用缩写、昵称等。如果把一个元组的所有属性连接在一起，把每一个元组看成是一个文本字符串，就可以利用文本字符串匹配的方法来解决实体识别问题。也可以将各个属性分开进行考虑，用领域知识等信息来进行数据匹配。如图 1 所示，如果两个元组中的邮箱（Work-mianbox）和电话号码（Tel）完全一样，我们就可以认为两个元组是匹配的。

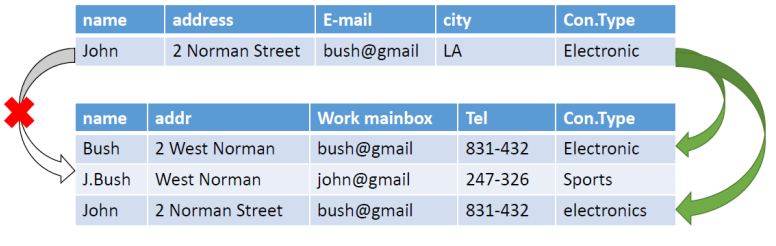


图1 实体识别示意

随着现代机器学习技术在实体识别上的应用，实体识别的准确性在公开数据集上已经得到大幅度提升。然而，在部署到实际的商业应用场景下时,该领域技术仍存在不足。一方面，基于传统机器学习方法的重复记录检测技术通常要求繁琐的特征工程，以及设计专门的技术对待检测的数据进行预处理，包括模式对齐、数据抽取与转换等。另一方面，基于深度学习的重复记录检测技术无需特征工程，但存在所需的标注数据量大、运行时间长、缺乏可解释性等问题。在面向大规模数据（行数在百万级以上）时，如何定制策略缩小搜索范围，降低深度学习模型的运行时间，同时利用深度学习模型输出的中心特征，简化机器学习的特征工程，提高整体实体识别模型的可解释性，是十分重要但尚待解决的问题。

在构建实体识别模型的过程中，往往伴随着复杂的特征工程。由于输入的关系数据属性多，包括数值型，字符串型等不同属性类型，存在着量纲不同，数据冗余，存在缺失值等问题。因此，为这些数据构建特征以提升模型效果是十分重要但尚未解决的问题。对于大规模数据的特征工程，如何构建好的特征空间合理表征数据的属性关系是一大难题。其次，由于数据量大，数据属性多，超参搜索空间庞大，单纯依靠模型训练进行参数学习的效率很低。并且由于数据标注量少，直接投入模型训练难以达到好的学习效果，需要人为进行大量的数据标注工作，耗费大量人力。本研究考虑设计特征空间，构建自动特征工程，并利用超参搜索调优等技术，减少模型运行时间和人力消耗。如图2所示，通过构建属性间的相似度函数库，进行自动参数选择并结合下游的实体识别模型反馈信息，进行实体识别的自动特征工程。

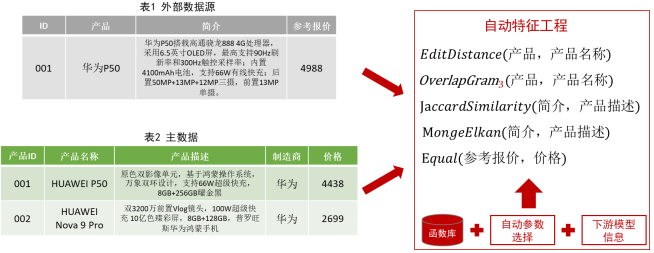


图2 实体识别的自动特征工程

1. 国内外在该方向的研究现状及分析

国内外学者对于实体识别任务及其上的特征工程已经进行了广泛的研究。本文将从主流实体识别方法及实体识别的特征工程两个方面对该方向进行介绍。

**2.1 主流实体识别方法**

实体识别的主流方法包括基于规则的传统方法、基于众包的人在回路方法和基于机器学习的智能方法等。

（1） 传统实体识别：基于规则的方法

基于规则的方法主要思想为设计相似度函数和阈值，是实体识别领域最经典的方法。文献[3]表明，早在 20 世纪 70 年代，就有学者使用属性相似度均值的方法实现实体识别。文献[4]提出的 Swoosh 实体识别算法是十分具有代表性的基于规则的方法，当匹配操作和合并操作满足 ICAR 特性：幂等性、交换性、结合性、可被代表性，就能使用 FSwoosh 或 R-Swoosh 算法降低匹配操作的次数、提高实体识别的效率。文献[5]在依赖推导的基础上提出 3 个算法：SiFi-Greedy，SiFi-Gradient，SiFi-Hill，为匹配规则和候选规则选择相似性函数和阈值。文献[17]以数据质量管理中的去重操作为实体识别的应用

场景，比较和分析了交集类、加权和类、语言模型类、编辑类、组合类等不同相似度函数。

（2） 人在回路的实体识别：基于众包的方法

众包是人在回路的实体识别的早期尝试。基于众包的实体识别方法将实体对候选集合发布到众包平台上，由大量人员对实体对匹配与否给出判断来得到结果，传统众包方法代价高且存在错误回答的风险，因此研究人员提出了许多优化方法。文献[7]首先将未标记的实体对训练集分成两部分，一部分利用众包进行标记，另一部分利用记录之间的传递关系进行标记，以最小化众包对数量、降低成本。文献[8]同样着力于利用传递性来减少众包操作中询问的数量，证明了文献[7]中提出的寻找最优询问策略是 NP 难问题，并提出两种方案：随机顺序询问和基于优先级的询问。文献[9]提出了一个基于偏序的众包框架 POWER，通过构造一个基于偏序的询问图进行提问和推断，该框架降低了众包成本，且可以容忍众包和偏序引入的错误。

近年来，人在回路的实体识别的研究逐渐脱离众包的限制，开始显现出系统化和智能化的趋势。文献[10]提出了一种 Magellan 的实体识别系统，该系统提供了完整的操作指南以及多步骤的工具供选择，已被应用于许多工业场景中。文献[11]则提出了基于主动学习的实体识别框架，允许用户对组件进行选择，对于某些数据集甚至可以超过监督学习的性能。

（3） 智能实体识别：基于机器学习和深度学习的方法

实体识别可以视作二分类问题，智能实体识别方法把传统的分类、特征提取等思想应用在实体对上以得到二分决策器的方法，通常有更高的精确率。早期的智能实体识别方法包括决策树[12]、朴素贝叶斯[13]、支持向量机[14]等，其中最具代表性的是[12]提出的运算符树（operator tree）结构，把匹配规则模型转化为树：根节点是并集操作符，中间节点是相似性函数，叶子节点是数据，最后通过训练集来产生结果。这类树模型的思路逐渐演变为随机森林[15]等方法，这类方法迄今在很多计算资源有限、数据量巨大、可解释性要求较高的工业界场景下还有很多应用。文献[16]着力于解决模型训练需要大量标签数据的问题，提出了无监督学习方法 ZeroER，该方法基于高斯混合模型生成训练集，提出特定正则化技术改善过拟合，最后利用传递性提高准确率，可以达到与监督学习相当的性能。文献[17]探索了自动机器学习（AutoML）技术在实体识别领域的应用，得到了接近于深度学习方法的性能。

深度学习的方法使得实体识别的效果大大提升，引领了该领域的进展。文献[18]提出了一种准确、高效且易于使用的基于元组分布式表示的新型实体识别系统——DeepER，元组表示学习采用了双向递归神经网络，并通过端到端的全局方法进行调整，显示出优越的性能。文献[19]从实验的角度对 word2vec、GloVe、fastText、RNN、Attention等深度学习技术在不同实体识别问题上的表现进行了对比、评估。

自然语言处理领域出现 Transformer 模型[20]后，基于该类模型的实体识别方法效果超过了原有的深度模型。文献[21]分析了四种最新的 Transformer 结构：BERT，XLNet4，RoBERTa 和 DistilBERT在实体识别任务中的表现，Transformer 类方法效果优于经典的深度学习方法。文献[22]提出了基于预训练语言模型 BERT 的实体识别系统 Ditto，首先将实体记录对建模成序列对进行预训练，并加入了领域知识注入、概括过长序列、数据增强的模块。文献[23]提出了双目标 BERT 训练方法——JointBERT，使得模型在预测匹配或不匹配之外预测训练对中每个实体的实体标识符。

**2.2实体识别的特征工程**

由于手动构建特征工程耗费大量人力，且需要大量的领域知识，并且无法广泛的适配实际场景，我们希望从数据集中自动创建候选特征，且从中选择若干最佳特征进行训练，即自动特征工程。自动特征工程旨在不使用人力或者借助少量人力，自动化地构建模型的输入特征集合，使得后续的模型能够更高效准确地识别出重复记录，进行实体识别任务。

对于表格数据进行实体识别任务时，从大量的数值型，离散型等属性上进行特征提取和特征选择，构建合理的特征空间，是保证下游模型性能的重要一环。文献[24]使用分组采样策略和编辑距离作为相似度函数构建特征，Magellan[10]将属性分成多种类型，包括数值型，布尔型，字符串型等，并为每种类型的属性提供了多种相似度函数。对于给定的实体对，实体属性是确定的，通过计算实体对间涉及每种属性的相似度，作为每一维特征的取值。AutoML-EM[25]则构建了一个所有属性通用的相似度函数库，构建实体对的特征时，直接计算库中所有函数的相似度数值。上述两种方案都是通过计算文本字符串相似度得到特征取值，DeepMatcher[26]则借鉴了NLP社区处理文本信息的技术，采用word-embedding获取文本数据的语义信息用于深度模型训练。

由于实体识别任务面向的数据集存在大量实体，而标注数据很少，直接将数据输入模型效果一般。针对此问题，文献[27]提出了通过输入少量数据进行自训练，然后利用Boosting进行集成学习提升整体效果。AutoML-EM[25]利用主动学习技术计算标签预测置信度，迭代式的人工标注低置信度数据。这种方式可以利用具有很少标注量的初始数据进行模型训练，并且利用较高的置信度保证，减少了需要人为标注数据的数量。文献[28]的Autofuzzyjoin进一步减少了人力参与的程度，利用迭代式集合运算，要求输入的数据仅需两个待实体识别的表L与表R，以及期望的准确度阈值即可，而与初始数据是否存在标签数据无关。

**2.3国内外文献综述的简析**

从上述对于国内外研究现状的介绍中，我们可以发现，实体识别作为一个问题背景复杂的研究领域，学者们早期主要关注如何设计复杂的规则得到合适的实体识别判别逻辑，尽管这一类方法具有泛化能力弱、对专家知识要求过高等问题，其作为特征工程的尝试还是驱动了智能实体识别方法的发展。随着自动机器学习的发展，学者们开始采用相似度函数构造特征空间的方法，进行自动化的特征工程。通过迭代式的自动化训练，减少了人力需求。但一方面，基于传统的机器学习方法的实体识别技术通常要求繁琐的特征工程，以及设计专门的技术来处理异构数据，例如，模式对齐、数据抽取与转换等。另一方面，基于深度学习的实体识别技术无需特征工程，但存在所需的标注数据大、运行时间长、可解释性差等问题。因此，尽管现有智能实体识别方法在公开数据集上取得了优异的效果，进行了特征工程，但是仍需要领域专业知识，缺乏自动化能力适应不同领域的数据集，在可扩展性和模型效果要求较高的实际应用场景下，难以达到用户需求。

1. 主要研究内容

根据第 2 节中对该课题的研究现状的分析，我们发现面向大规模实体识别的自动特征工程研究刚刚兴起，诸多难点仍未能得到有效的解决，因此我们确定了本课题的研究内容，针对大规模实体识别的自动特征工程开展研究。主要的研究内容如下：

研究内容1：实体识别特征工程空间构建；

研究内容2：自动化特征工程的超参搜索；

研究内容3：时空有效的自动特征工程加速；

其中，研究内容1的特征工程构建过程实现了模式对齐，实体识别预分块，数据清洗，相似度函数库构建，特征向量生成。研究内容2针对自动特征工程中的模型进行超参搜索工作，将研究内容1中生成的特征向量输入给底层的智能实体识别模型。由于数据量大，且数据标注量小，参数搜索空间很大，因此需要设计模型进行自动特征工程的超参搜索。研究内容3针对1，2中的模型和超参搜索过程，利用剪枝，分块等策略进行加速优化，缩小搜索空间，降低时间复杂度，最终形成一个时空有效的大规模实体识别特征工程系统。

研究内容的逻辑结构图如图2所示。这三点研究内容层次较为合理，覆盖了实体识别的关键问题。

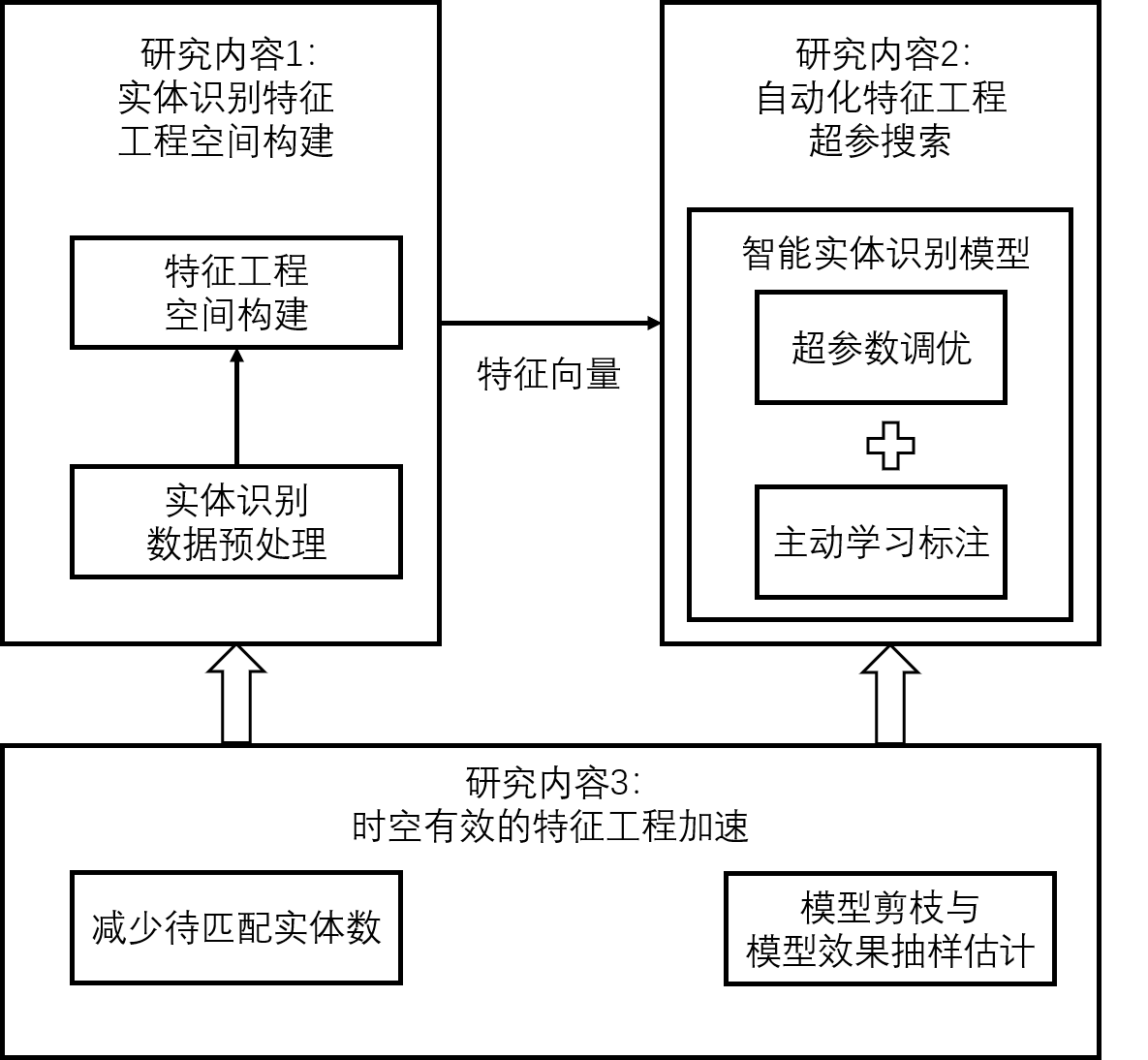


图3 研究内容逻辑结构图

**3.1问题定义**

本研究只讨论关系型数据上的表格数据的实体识别任务，即待匹配的实体是遵循具有属性的不同模式的结构化记录,通过分析实体间的结构化记录，研究二者是否指向现实世界中的同一个实体。表格之间的属性使用现有工具进行对齐处理。每个记录的单元格仅包含与描述每个单元格的属性相关联的信息。属性内文本信息长度有限。

首先给出本文最基本的五个概念：实体，笛卡尔积，n元组，关系和关系模式的定义：

定义1（实体）：实体是一个数据对象，指应用中可以区别的客观存在的事物。

定义2（笛卡尔积）：给定一组属性域,这些域中可以有重复。的笛卡尔积定义为。即所有域的所有取值的一个组合，其中不能有重复。

定义3（n元组）：笛卡尔积中每一个元素叫做一个n元组，或简称元组。

定义4（关系）：的子集叫做在属性域上的关系，表示为 ，称为n元关系，其中R为关系名，n为关系的目或度。

定义5（关系模式）：一个具体的关系描述，具有形式R(U, T, I, F), 其中，R是关系名，表示为。U是R的属性集合，A是U中属性的域集合，I是完整性约束集合，F是属性间的函数依赖关系。

定义6（特征）：特征通常是建立在原始数据之上的特定表示，它是一个单独的可测量属性，通常由数据集中的列描绘。

接下来给出本研究中实体识别及其特征工程的定义：

问题1（实体识别）：考虑关系型数据中，D和D'为两个实体元组集合， D和D'中的实体元组可以具有不同关系模式表示。实体识别的目标是找到D和D'之间表示同一真实世界客体的所有实体元组对。

问题2（实体识别的特征工程）：对于具有不同关系模式R的实体对，通过设计特征空间，将实体记录中的属性数据映射到高维空间转换为特征向量v，并输入实体识别模型搜索超参数的过程。

**3.2实体识别特征工程空间构建**

由于智能实体识别模型的输入要求是特征向量的形式，向量内部应为数值型数据，而表格数据中的实体对数据是由属性值构成的文本字符串序列。因此，需要我们设计方法构建特征工程空间，将字符串类型数据转换为数值型数据。

首先，我们利用现有工具进行模式对其，然后对输入的数据进行分块处理，删除不需要匹配的实体对。其次进行初步的数据清洗，其中包括缺失值填充，长文本处理，大小写不一致等，为后续的特征向量构建进行良好的数据准备。最后，需要将字符串类型数据转换为数值型数据。一个合理的想法是参考autoML-EM的思路，按照属性分类后，构造相似度函数库，然后利用相似度函数，将实体对的文本字符串序列转换为实体对间的相似度数值，从而利用相似度数值构建特征向量。

**3.3自动化特征工程的超参搜索**

自动化特征工程的超参搜索，旨在减少参数搜索空间，克服标签数据短缺的困难，提升模型的性能上限。针对参数空间过大问题，可以考虑用贝叶斯优化，遗传算法等方法进行优化。对于标签过少的问题，可以参考autoML-EM提出的主动学习与自训练相结合的方式，降低对初始数据集中标签个数依赖的同时，降低人力参与的需求。此外，在超参数学习的过程中，属性列之间的权重难以确定，可以考虑采用autofuzzyjoin的方法进行列权重系数的计算。

**3.4时空有效的自动特征工程加速**

智能实体识别模型往往训练时间较长，尤其是深度模型，运行时间达到了小时级。时空有效的自动特征工程加速研究旨在通过制定相关规则来删除不能匹配的实体对缩小搜索空间，并采用模型剪枝常见的模型加速策略的方式，加快整体的自动特征工程的运行时间，从而提升整体自动特征工程的时空有效性。

1. 研究方案

**4.1实体识别特征工程空间构建**

研究内容1实现了模式对齐，数据预分块，数据清洗，特征空间设计构建，特征向量生成的一系列流程。具体来看，模式对齐部分要求处理后的数据遵循具有属性的相同模式的结构化记录，这一操作可以利用现有的模式对齐工具处理。数据预分块操作通过将表格数据中所有的实体对进行分块，则只需要对块内的实体对进行匹配比较操作，而块间实体对不需要比较，从而缩短匹配时间。数据清洗部分可以考虑基于规则的方法，初步处理命名方式不一致，格式不一致等问题。例如对于数值类型数据“2000”与“2,000”，可以采用将分隔符去除的方式进行初步的数据预处理。对于字符串类型数据中存在的大小写不一致问题，可以将大小写语义不敏感的属性对应的属性值，统一转换为小写，避免后续相似度函数计算时引入偶然误差。

特征向量构建与生成部分，需要将实体对的文本序列字符串，转换为数值型数据，以供下层智能实体识别模型训练。设处理的表格数据属性数为，本研究内容首先获取数据的全部属性，然后为每种属性构建相似度函数,其中。从而，每个实体对对应了一个由相似度函数取值构成的特征向量，其中向量的每一维由实体对对应的属性，在对应位置计算得出的相似度数值构成。向量的最终形式为,其中由第个属性对应的相似度函数集合中每一个元素（相似度函数）计算得出的间相似度构成，即：

最终，向量总长度为。

综上，对于研究内容1，我们拟确定以下两个研究点：

（1）实体识别数据预处理：利用现有工具对表格数据进行模式对齐，填充数据缺失值。（可选）

（2）特征工程空间设计构建：利用多种相似度函数，构建属性上的相似度函数库，计算实体对间的相似度数值，自动化构建特征向量。

**4.2自动化特征工程的超参搜索**

本研究内容旨在通过制定自动化特征工程的超参搜索策略，通过缩小超参搜索范围和设计方法降低对人为标注数据需求这两个角度，试图降低数据量过大，标注数据过少这两个问题对于特征工程的影响。

可以采用遗传算法，贝叶斯优化，网格搜索等搜索方法进行超参数的优化，并结合主动学习，自训练技术，计算预测出的标签数据置信度，对于低置信度的数据人工标注，高置信度数据直接赋予预测的标签进行迭代式的训练。对于多属性列权重的制定问题，Autofuzzyjoin提出了计算列权重的方法可供参考。

综上，对于研究内容2，我们拟确定以下两个研究点：

（1）超参数调优：通过遗传算法，贝叶斯优化等方法，缩小超参数搜索范围。

（2）主动学习标注：通过采用主动学习技术计算标签置信度，并人为标注少量低置信度数据相结合的方式，克服数据标注量少的问题。

**4.3时空有效的自动特征工程加速**

本部分我们对研究内容1，2进行优化，针对模型整体如何进一步加速，提升时空有效性展开研究。首先，我们通过制定实体对排除方法例如blocking的方式，将可能需要匹配操作的实体放在同一块中，从而不同块间实体对无需匹配，降低了匹配时间。对于模型训练方面，可以采用模型剪枝技术，例如基于相似度的剪枝，基于匹配准确率，召回率估计的剪枝等，进一步降低模型运行时间，提升时空有效行性。最终形成具有时空有效性的大规模实体识别自动特征工程系统。

综上，对于研究内容3，我们拟确定以下两个研究点：

（1）减少待匹配实体对数量：通过制定实体对排除方法如blocking等方式，减少可能需要匹配的实体数量，降低模型输入的数据量

（2）模型剪枝与模型效果估计：制定模型剪枝策略，抽样估计模型效果，提升模型运行速度，提升时空有效性。

1. 进度安排，预期达到的目标

**5.1进度安排**

2022.10-2022.11 进行场景调研和定义完善。

2022.12-2023.3 设计模型进行初步实验，参加中期答辩。

2023.3-2023.5 对自动特征工程和实体识别模型进行优化，准备撰写毕业设计论文。

2023.5-2023.6 撰写毕业设计论文，准备毕设答辩。

**5.2预期达到的目标**

完成模型设计，通过实验测试大规模实体识别特征工程效率，完成毕业论文。

6．课题已具备和所需的条件、经费

(1)硬件方面

实验验证自动特征工程针对于大规模关系型表格数据的效果时，可能需要使用大内存的服务器，该条件已具备，服务器由海量数据计算研究中心提供。

(2)文献方面

研究过程中需要的参考文献已具备:由哈工大图书馆中文数据库与外文数据库提供(http://www.lib.hit.edu.cn/databasenav)。

(3)经费方面

除必要的书籍等参考资料购置费用外，无其他经费需求。

1. 研究过程中可能遇到的困难和问题，解决的措施

预计本课题的可行性较高，在研究过程中不会遇到极端困难障碍。为了高质量地完成课题，将可能遇到的问题及解决途径总结如下：

(1)算法设计上出现思路局限的问题。在思考停滞不前的时候，不局限于实体识别和特征工程的方法，多思考自动机器学习，自然语言处理等其他领域的解决办法。同时阅读相关书籍，以达到拓展思路的目的。

(2)实验数据上的问题。大规模实体识别数据可以从经典的实体识别baseline论文中获取，可以参考论文中实验部分的数据集处理使用。

8．主要参考文献

[1] AnHai Doan, Alon Y. Halevy, Zachary G. Ives: Principles of Data Integration. Morgan Kaufmann 2012, ISBN 978-0-12-416044-6, pp. I-XVIII, 1-497

[2] Dong, X. L.; and Srivastava, D. 2013. Big data integration. In ICDE, 1245–1248. IEEE

[3] Fellegi, I.P. & Sunter, A.B., 1969. A Theory for Record Linkage. Journal of the Americal Statistical Association, 64(328), pp.1183–1210.

[4] O. Benjelloun, H. Garcia-Molina, D. Menestrina.et al. Swoosh: A generic approach to entity resolution. VLDB J. 2009,18(1): 255–276.

[5] J. Wang, G. Li, J. X. Yu, and J. Feng. Entity matching:How similar is similar. PVLDB. 2011, 4(10): 622-633

[6] A. Chandel, O. Hassanzadeh, N. Koudas, M. Sadoghi, and D. Srivastava. Benchmarking declarative approximate selection predicates. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pages 353–364. ACM, 2007

[7] Wang, J., Li, G., Kraska, T., Franklin, M.J., Feng, J. Leveraging transitive relations for crowdsourced joins. SIGMOD Conference. 2013: 229–240.

[8] N. Vesdapunt, K. Bellare, and N. Dalvi. Crowdsourcing algorithms for entity resolution. Proc. VLDB Endow.2014, 7(12): 1071-1082.

[9] C. Chai, G. Li, J. Li, D. Deng, and J. Feng, A partial-order-based framework for cost-effective crowdsourced entity resolution. VLDB J.2018, 27(6): 745–770.

[10] P. Konda et al. Magellan: Toward building entity matching management systems.Proc. VLDB Endow.2016, 9(12): 1197-1208.

[11] V. Meduri, L. Popa. P. Sen, M. Sarwat.et al. A Comprehensive Benchmark Framework for Active Learning Methods in Entity Matching. In Proc. SIGMOD.2020: 1133-1147, 2020

[12] S. Chaudhuri, B. Chen, V. Ganti, and R. Kaushik. Example-driven design of efficient record matching queries. In VLDB07, pages 327–338, 2007.

[13] W. E. Winkler. The state of record linkage and current research problems. In Statistical Research Division, U.S. Census Bureau, 1999

[14] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In KDD, 2003.

[15] Kunho Kim, C. Lee Giles: Financial Entity Record Linkage with Random Forests. DSMM@SIGMOD 2016: 13:1-13:2

[16] R. Wu,S. Chaba, S. Sawlani.et al. ZeroER: Entity Resolution using Zero Labeled Examples. SIGMOD Conference. 2020: 1149-1164.

[17] Pei Wang and Weiling Zheng and Jiannan Wang and Jian Pei, Automating Entity Matching Model Development, 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021, 1296—1307

[18] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq R. Joty, Mourad Ouzzani, Nan Tang. Distributed Representations of Tuples for Entity Resolution. Proc. VLDB Endow. 11(11): 1454-1467 (2018)

[19] S, Mudgal, H, Li, T, Rekatsinas.et al. Deep learning for entity matching: A design space exploration. SIGMOD Conference.2018: 19–34.

[20] A.Vaswani, N.Shazeer, N.Parmar.et al. Attention is All you Need. NIPS.2017: 5998-6008.

[21] U, Brunner, K. Stockinger.Entity Matching with Transformer Architectures-A Step Forward in Data Integration. EDBT.2020: 463-473.

[22] Li Y, Li J, Suhara Y, et al. Deep Entity Matching with Pre-Trained Language Models. Proc. VLDB Endow.2020, 14(1): 50-60.

[23] Ralph Peeters and Christian Bizer. Dual-Objective Fine-Tuning of BERT for Entity Matching. PVLDB, 14(10): 1913 - 1921, 2021.

[24] Sunita Sarawagi and Anuradha Bhamidipaty. 2002. Interactive deduplication using active learning. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). Association for Computing Machinery, New York, NY, USA, 269–278.

[25] P. Wang, W. Zheng, J. Wang and J. Pei, "Automating Entity Matching Model Development," 2021 IEEE 37th International Conference on Data Engineering (ICDE), 2021

[26] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep Learning for Entity Matching: A Design Space Exploration. In Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18). Association for Computing Machinery, New York, NY, USA

[27] Bilenko, M., Mooney, R. J.: Adaptive duplicate detection using learnable string similarity measures. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 39–48. ACM (2003)

[28] Feng Li, Xiang Cheng, Xu Chu, Yeye He, and Surajit Chaudhuri. 2021. Auto-FuzzyJoin: Auto-Program Fuzzy Similarity Joins Without Labeled Examples. In Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21). Association for Computing Machinery, New York, NY, USA, 1064–1076