# AI 前沿论文最新进展

2018.10.30 方建勇

1，在教科书或研究论文的 PDF 文档中包含数学表达的 LaTeX 来源，对于“可访问性”考虑具有明确的好处。在这里，我们描述了三种可以完成此操作的方法，完全兼容国际标准 ISO 32000，ISO 19005-3 和即将推出的 ISO 32000-2（PDF 2.0）。两种方法使用嵌入式文件（也称为“附件”），以 LaTeX 或 MathML 格式保存信息，但使用不同的 PDF 结构将这些附件与文档窗口的区域相关联。一个使用结构，因此适用于完全“标记 PDF”上下文，而另一个使用/ AF 标记相关内容。第三种方法根本不需要标记，而是将源代码包括为所谓的“假空间”的/ ActualText 替换。以这种方式提供的信息是通过简单的选择/复制/粘贴操作提取的，并且可用于现有的屏幕阅读软件和辅助技术。

2，标题，摘要，关键词或索引术语，正文，结论，参考等科学文章的元数据在收集，管理和存储科学数据库，学术期刊和数字图书馆的学术数据方面起着决定性作用。从科学论文中准确提取这些数据对于为研究人员和图书馆员组织和检索重要的科学信息至关重要。研究社交网络系统和学术数字图书馆系统提供学术数据提取，组织和检索服务。这些类型的服务大多不是免费或开源的。它们还存在一些性能问题，并且可以提取可以上载到提取系统的 PDF（可移植文档格式）文件数量的限制。在本文中，提出了一个完全免费和开源的基于 Java 的高性能元数据提取框架。此框架提取速度比现有元数据提取系统快 9-10

倍。它还具有灵活性，允许上载无限数量的 PDF 文件。在这种方法中，使用文本的布局特征，字体和大小特征来提取论文标题。使用固定规则集从 PDF 文件中提取其他元数据字段，例如摘要，正文，关键字，结论和参考。提取的元数据存储在 Oracle 数据库和 XML（可扩展标记语言）文件中。该框架可用于在数字图书馆，在线期刊，在线和离线科学数据库，政府研究机构和研究中心制作科学馆藏。

3，作为智慧城市计划的一部分，全球各地的国家，地区和地方政府都有权就如何共享数据采取更加开放的态度。根据这一要求，这些政府中的许多政府都在政府公开数据的保护下发布数据，其中包括来自全市传感器网络的测量数据。此外，许多这些数据作为文档发布在所谓的数据门户中，这些文档可以是电子表格，逗号分隔值（CSV）数据文件或 PDF 或 Word 文档中的普通文档。共享这些文档可能是数据提供者传送和发布数据的便捷方式，但它不是数据使用者重用数据的理想方式。例如，重用数据的问题可能包括难以打开以任何非纯文本格式提供的文档，以及理解文档内部每条知识含义的实际问题。我们的提案通过识别被认为与测量数据相关的元数据并为此元数据提供模式来解决这些挑战。我们进一步利用人工感知传感器网络本体（HASNetO）为城市环境中收集的数据构建架构。我们讨论了使用 HASNetO 和支持基础设施来管理数据和元数据，以支持巴西的大都市区福塔莱萨市。

4，SemEval-2010 基准数据集重新引起了对自动关键短语提取任务的关注。此数据集由自动从 PDF 格式转换为纯文本的科学文章组成，因此需要仔细预处理，以便不可逆转的文本范围不会对关键短语提取性能产生负面影响。在以前的工作中，描述了各种文档预处理技术，但它们对关键短语提取模型的整体性能的影响仍未得到探索。在这里，我们重新评估几个关键短语提取模型的性能，并测量它们对日益复杂的文档预处理级别的稳健性。

5，我们提出了一种从文本中提取测量信息的方法（例如，1370°C的熔点，BMI 大于 29.9kg /平方公尺）。这些提取对于广泛的领域至关重要 - 特别是涉及搜索和探索科学和技术文件的领域。我们首先提出一个基于规则的实体提取器来挖掘测量数量（即，与测量单元配对的数值），它支持大量且全面的常见和模糊测量单元。我们的方法非常强大，即使在将文档格式（如 PDF 格式）转换为纯文本的过程中引入了重大错误，也能正确恢复有效的测量数量。接下来，我们描述了提取被测量的属性的方法（例如，短语“像素间距高达 352μm”中的属性“像素间距”）。最后，我们介绍 MQSearch：实现一个完全支持测量信息的搜索引擎。

6，我们提出了用于学术搜索的 Bullseye 系统。鉴于一系列研究论文，Bullseye：1）使用任何现成算法识别相关段落; 2）自动检测文档结构并将检索到的段落限制在用户指定的部分; 3）突出显示每个检索到的

PDF 文档的段落。我们在三个方面评估 Bullseye：系统有效性，用户效率和用户努力。在系统盲评估中，要求用户比较使用 Bullseye 的通道检索与忽略文档结构的基线，对于四种类型的评分评估。结果显示系统有效性略有改善，而用户效率和用户努力都显示出实质性改进。用户还报告了对所考虑的两个系统的学术搜索中的段落突出显示非常强烈的需求

7，技术文档包含大量不自然的语言，例如表格，公式，伪代码等。非自然语言可能是混淆现有 NLP 工具的重要因素。本文提出了一种区分非自然语言和自然语言的有效方法，并评估了非自然语言检测对 NLP 任务（如文档聚类）的影响。我们将此问题视为信息提取任务，并构建将非自然语言组件分为四类的多类分类模型。首先，我们通过收集各种格式的幻灯片和论文（PPT，PDF 和 HTML）来创建一个新的带注释的语料库，其中非自然语言组件被注释为四个类别。然后，我们将探索纯文本中可用的功能，以构建可以处理任何格式的统计模型，只要它转换为纯文本即可。我们的实验表明，删除不自然的语言组件可以使文档聚类的绝对改进高达 15%。我们的语料库和工具是公开的。

8，非文本组件（如图表，图表和表格）提供了许多科学文献中的关键信息，但缺乏大型标记数据集阻碍了数据驱动方法的开发，用于科学数字提取。在本文中，我们为大量科学文献中的图形提取任务引入

了高质量的训练标签，没有人为干预。为了实现这一目标，我们利用两个科学文档（arXiv 和 PubMed）的大型网络集合中提供的辅助数据来定位栅格化 PDF 中的数字及其相关标题。我们共享超过 550 万个诱导标签的结果数据集 - 比之前最大的数字提取数据集大 4000 倍 - 平均精度为 96.8%，以便为此任务开发现代数据驱动方法。我们使用此数据集来训练深度神经网络以进行端到端图形检测，从而产生一种模型，与以前的工作相比，该模型可以更容易地扩展到新的域。该模型已成功部署在大规模学术搜索引擎 Semantic Sc holar 中，并用于提取 1300 万份科学文献中的数据

9，随着可移植文档格式（PDF）文件格式的普及，需要研究分析其文本提取和分析的结构。检测标题可以是分类和提取有意义数据的关键组成部分。该研究涉及训练有监督的学习模型，以通过递归特征消除来仔细选择具有特征的标题。表现最佳的分类器的准确度为 96.95%，灵敏度为 0.986，特异度为 0.953。这种对航向检测的研究有助于基于 PDF 的文本提取领域，并可应用于各种专业和基于策略的环境中的大规模 PDF 文本分析的自动化。

10，本文通过{\ it ICDAR 2013 Table Competition}发布的以 PDF 格式发布的 67 个带注释的政府报告，优于最先进的表格识别方法，本文提供了一种新的范例，利用大规模的无标签 PDF 文档打开 - 域表检测。我们将范例整合到我们最新开发的系统（{\ it PdfExtra}）中，通

过来自{\ it ACL Anthology}整个存储库的 9,466 篇学术文章来检测表格区域，其中几乎所有论文都以 PDF 格式存档而没有表的注释。范例首先设计启发式以自动构建弱标记数据。然后，它将各种证据（例如文档和语言特征的布局）提供给不同的规范分类器，这些证据由{\ it Apache PDFBox}提取并由{\ it Stanford NLP}工具包处理。我们最终使用这些分类器，即{\ it Naive Bayes}，{\ it Logistic Regression}和{\ it Support Vector Machine}，对表区域进行协作投票。实验结果表明，与最先进的方法相比，{\ it PdfExtra}实现了巨大的飞跃。此外，我们还讨论了可能影响性能的不同特征，学习模型甚至文档领域的因素。广泛的评估表明，我们的范例足够兼容，可以利用 PDF 文件中的开放域表区域检测的各种功能和学习模型

11，计算机取证工具中的一个分支是对涉密文档的取证分析。计算机中涉密文档一般以非结构化数据格式存储，数据库无法满足特定的检索要求，必须依赖高效的全文检索技术。Lucene 是一个用 Java 语言实现的高性能、可扩展的开源信息检索框架工具，本文实现的取证检索工具基于 Lucene 以实现索引和搜索功能。

本文首先深入分析 Lucene 源码，采用先提取关键 API 后深入 API 源码的方式，逐步递进剖析索引架构及其文档建立、文本分析、索引建立的过程，并扩展研究了多线程机制下的索引实现。同时分析了搜索查询机制如何选取准确的索引文件、搜索文档相关度的评分准则以及搜索的核心 search 实现机制;其中 search 实现机制从分析查询的权

重计算、得分以及结果收集、处理这两个方面进行。

然后实现一个基于 Lucene 的全文检索取证工具系统，重点实现了文档抽取模块、中文分词模块、索引建立模块以及检索模块。在文档抽取模块中，实现了对非结构化文档如 Pdf、Word、Excel、PPT、XML、HTML 等文件格式的纯文本抽取。在分词模块中，采用对中文支持良好的 Paoding 中文分析器，剖析了其实现机制并扩展了涉密字典，弥补了 Lucene 在中文分词方面的薄弱。在检索模块中，对已建立的索引模块实现了四种方式的检索:单个关键字检索、关键字库检索、多关键字检索以及时间段检索，检索结果按照相关度算法进行排序，并配合用户体验良好的界面输出。最后对系统进行测试与改进，提升检索工具的查全率与查准率，达到了最初的设计目的。

12，Web 应用中数据库交互行为的验证

作者：孙茂华

外文题名：Verifying Database Interactions for Web Applications

学位授予单位：上海大学

导师姓名：缪淮扣

摘要

随着 Internet 的发展，Web 应用已在搜索引擎、电子商务、电子政务和社交网络等领域得到充分发展。同时 B/S 架构模式逐渐取代传统的 C/S 架构模式，使得 Web 应用在更广泛的领域发挥作用。Web 应用是一种非常复杂的、分布式的、多层架构的交互式系统，通常由 Web

浏览器、Web 服务器和应用服务器等组件组成。大多数 Web 应用都离不开数据库管理系统的支持，且包含一个或多个数据库。Web 应用的异构性、动态性、用户多样性以及开发周期短和需求变化快等特性给 Web 应用的验证带来了新的挑战。 目前，学者们大多关注的是从 Web 应用的需求模型到其设计模型的验证，而从需求模型或设计模型到具体实现了的 Web 应用的验证关注得比较少。由于 Web 应用特有的复杂性，使得从设计模型到具体系统实现亦存在着诸多变化的因素。因此即使设计模型满足了需求模型，但是具体实现的 Web 应用是否满足需求模型尚不十分明确，需要进一步验证。 本文以 Web 应用数据库交互行为为研究对象，提出一种基于定理证明的 Web 应用数据库交互行为的验证方法。该方法涉及需求模型和实现模型。需求模型用 Z 语言书写的规格说明表示。本文提出了从 Web 应用数据库交互源代码到 Z 的转换规则，根据转换规则，将这些代码转换为 Z 规格说明表示的实现模型。从 Z 实现模型获取 Web 应用的性质。采用定理证明器 Z/EVES，验证这些性质是否在需求规格说明中得到满足。本文开发了从 Web 应用数据库交互代码段到 Z 的转换工具 DBICODE2Z。它能读入 Web 应用数据库交互源代码，根据转换规则自动生成 LaTeX 格式的 Z 规格说明，LaTeX 的 Z 规格说明可以被定理证明器 Z/EVES 接受并验证。 本文设计了该方法的验证框架，并给出了一个实例，演示了从 Web 应用数据库交互源代码到 Z 的转换以及应用 Z/EVES 进行验证的过程。

13，伴随着信息化技术不断地发展，科学文献以电子档的形式出现的需求越来越多，如何实现科学文献的电子化得到更加广泛的关注和深入的研究。数学公式是许多科学文献的重要组成部分，对文献的理解往往起着至关重要的意义，所以数学公式的电子化尤为重要。中学数学智能解答中题目的输入是一个重要的研究内容，题目中也包含了不少数学公式。当前的 OCR(Optical Character Recognition)技术可以很好地识别中英文字符以及数学字符，但由于数学公式结构的复杂性、符号的多样性以及符号的歧义性等原因，使得 OCR 对数学公式的识别变得较为困难，识别准确率低。另一方面数学公式手工输入比较困难，从而自动、高效的数学公式识别技术是必须突破的研究。研究数学公式字符识别技术的研究，是数学公式处理研究中的一部分，和数学公式定位、数学公式分析以及数学公式输出一起构成整个数学公式处理。针对的是印刷体文档中的数学公式识别问题，主要研究的对象是数学公式图像。数学公式的结构不是简单的一维的，而是复杂的二维的;字符出现在不同的位置所表示的意义是不一样的，字符没有统一的大小;数学公式中包含的字符有数字、字母、运算符号等，种类繁多。以上这些原因给数学公式符号的分割和识别都带来了一定的难度。数学公式识别系统主要研究数学公式中的字符分割和字符识别两个部分。

本研究在对数学公式图像进行分割前，对图像进行了预处理工作。预处理工作包括图像滤波去噪、图像二值化、图像倾斜校正和图像细化。数学公式符号分割采用的是投影法和连通域分割法相结合的方法，设计的算法可以很巧妙地分割出单个符号。对分割得到的单个符号做

归一化处理，为后续的特征提取和识别做了充分的准备。针对当前识别的低准确率和常见混淆符号的难识别性，提取三组具有代表性的特征:横纵交截特征、基于像素的网格特征和孔洞特征。特征相互之间存在一定的互补性，将这些特征输入条件随机场中进行训练，从中学习得到对应的条件随机场，并对测试数据集做识别测试。基于特征融合训练的条件随机场，对符号识别的正确率达到了的 97.1%，比传统的识别方法具有更好的识别效果。

14，随着计算机技术和互联网技术的快速普及与迅猛发展，用户对书籍及文献资料的电子信息化需求越来越大，书籍及文献资料的电子信息化不仅包括电子化存储，还包括对内容的分析与理解等。随着字符识别技术的发展，光学字符识别对于电子信息化的书籍文献中的英文字符和数字等具有很好的识别效果，但是由于书籍文献中的数学公式符号存在种类复杂、尺寸变化大、二维嵌套结构等难点，数学公式的定位与识别方法的精度还不能满足实际需求。

本文以文本中数学公式的精确定位与识别为目标，研究了不同版面下的公式定位以及公式中的数字、运算符号、希腊字符、英文字符的特征提取和识别。论文的主要工作如下:

(1)对数学公式的图像进行了预处理。预处理包括去噪、倾斜校正、图像细化以及毛刺去除等操作，为公式符号的分割及识别打下了基础。

(2)分析了书籍文献的版面结构特点和文本中数学公式的排版位置特征等，给出了一种基于投影法的数学公式定位方法，该方法能够

准确地将文本中的公式进行定位提取。

(3)投影法是数学公式符号分割普遍采用的方法，但该算法只对于结构简单、无角标、无层次结构的数学公式分割有效。为了分析处理复杂的二维嵌套结构数学公式，本文研究并给出了一种改进的基于连通域的数学公式符号分割方法，该方法实现了嵌套结构数学公式中字符的精确分割。

(4)特征提取与分类器是数学公式符号识别的关键环节。考虑到公式字符的多样性，本文给出了一种多特征融合的特征提取算法，融合的特征包括孔洞特征、交截特征、网格区域特征、不变矩特征。为寻求公式字符的最佳分类效果，本文分别采用了模板匹配、人工神经网络、SVM 三种分类方法。实验结果表明，基于多特征融合和 SVM 的分类方法精度较高。另外，对于相似字符采用基于模板匹配的二次分类方法，有效的提高了公式字符识别的精度。

15，针对印刷体数学公式中的结构分析,提出将"自下而上"和"自上而下"相结合的策略。自上而下是针对特殊结构的分析,特殊结构包括根号、矩阵、上下标等。自上而下是对公式整体结构的分析,并且用递归的方式对各个子表达式采用同样的分析方法。结构分析成功后,用树形结构表示整个公式的二维空间布局。实验结果表明,此种分析策略有效地提高了印刷体数学公式的结构分析成功率。

16，用于识别字母字符和数学符号的人工神经网络和模糊逻辑

作者：GiuseppeAiròFarulla，Tiziana Armano，Anna Capietto，Nadir Murru，Rosaria Rossini

摘要：光学字符识别软件（OCR）是获取可访问文本的重要工具。我们建议使用人工神经网络（ANN）来开发能够识别正常文本和公式的模式识别算法。我们提出了反向传播算法的原始改进。此外，我们描述了一种新颖的图像分割算法，该算法利用模糊逻辑来分离触摸字符。

17，当前的关键词自动提取研究大多基于候选词的词频、文档频率等统计信息,往往忽略了侯选词所在的学术文本的内在结构,导致关键词提取的效果不佳。本文将学术文本看作是 5 个结构功能域的集合,提出了融合学术文本结构功能特征的多特征组合提取方法,并利用学术文本的章节标题对其结构功能进行识别,然后通过 SVM 二分类和 LambdaMART 学习排序算法分别在计算机语言学领域的文献集上进行了实现。实验结果表明,本文提出的组合特征方法相比基准特征在关键词提取的效果上取得了较大的提升,尤其在分类实验中准确率的相对提升上达到 10.75%,证明了学术文本结构功能特征在关键词自动提取上的重要性。

18，共享经济的增长是由共享平台的出现推动的，例如 Uber 和 Lyft，它们希望与想要租用它们的客户分享他们的资源。这种平台的设计是经济学和工程学的复杂混合，如何“优化”设计这样的平台仍然是一个悬而未决的问题。在本文中，我们将重点放在共享平台的价格和补贴

设计上。我们的结果提供了对收入最大化价格和社会福利最大化价格之间权衡的见解。具体而言，我们引入了一种新的共享平台模型，并描述了该模型中利润和社会福利最大化的价格。此外，我们将效率损失限制在利润最大化的价格下，表明在实际环境中利润与效率之间存在强烈的一致性。我们的结果强调，由于供应短缺，平台的收入可能在实践中受到限制;因此，平台有强烈的动机鼓励通过补贴分享。我们提供了这种补贴何时有价值的分析特征，并说明如何优化所提供补贴的规模。最后，我们使用来自中国最大的共乘平台滴滴出行的数据，验证了我们分析的见解。

19，本文考虑了一个闭路排队网络模型的共乘系统，如滴滴出行，Lyft和优步。我们专注于空车路线，这是一种机制，通过该机制我们控制网络中的车流以优化系统范围的实用功能，例如，乘客到达时空车的可用性。我们将排队网络的流程级和稳态收敛建立到大型市场体系中的流体限制，其中对汽车的乘坐和供应的需求倾向于无限，并且使用该限制来研究基于流体的优化问题。我们证明了基于流体的优化获得的最优网络效用是有限汽车系统中用于任何静态和动态路由策略的效用的上限，在该策略下闭合排队网络具有静态分布。在基于流体的最优路由策略下渐近地实现该上限。滴滴出行发布的实际数据的仿真结果证明了与其他各种策略相比，使用基于流体的最优路由策略的好处。

20，出租车服务和产品交付服务对我们现代社会起着重要作用。由于共享经济的出现，Uber，Didi，Lyft 和 Google 的 Waze Rider 等乘坐共享服务正变得越来越普遍，并成为我们日常生活中不可或缺的一部分。然而，这些服务的效率受到供需之间的次优和不平衡匹配的严重限制。我们需要一个通用的框架和相应的有效算法来解决有效匹配问题，从而优化这些市场的性能。现有的出租车和送货服务研究仅适用于单边市场的情况。相比之下，这项工作研究了市场经济中的出租车和交付服务（缩写为“出租车和交付市场”）的高度概括模型，可广泛用于双边市场。此外，我们为不同的应用提供有效的在线和离线算法。我们在实际设置下通过理论分析和跟踪驱动模拟来验证我们的算法。

21，在本文中，我们提出了机器学习方法，用于表征和预测按需乘车服务的短期需求。我们建议需求的时空估计是与交通，定价和天气条件相关的可变效应的函数。关于该方法，使用各种统计数据（例如，使用各种统计数据）对单个决策树，自举聚合（袋装）决策树，随机森林，增强决策树和用于回归的人工神经网络进行了调整和系统地比较。 R 平方，均方根误差（RMSE）和斜率。为了更好地评估模型的质量，他们使用中国主要的按需乘车服务提供商 DiDi Chuxing 的数据进行了实际案例研究。在目前的研究中，已经提取了 199,584 个描述时空乘车需求的时隙，其聚合时间间隔为 10 分钟。所有方法都是根据该数据集中的两个独立样本进行训练和验证的。结果显示，提升的决策树提供了最佳的预测准确度（RMSE = 16.41），同时避免了过度拟

合的风险，其次是人工神经网络（20.09），随机森林（23.50），袋装决策树（24.29）和单一决策树（33.55）。

22，Uber 和 Didi 等租赁平台在全球越来越受欢迎。然而，利用共享经济的未经授权的采购活动可能会严重损害这一新兴产业的健康发展。作为规范按需乘坐服务和消除黑市的第一步，我们设计了一种基于轨迹从汽车池中检测出租车的方法。由于法律问题，一些城市可能无法公开获得许可的外包汽车痕迹并且可能完全缺失，我们转而将公共交通开放数据（即出租车和公共汽车）的知识转移到普通车辆的检测中。我们提出了一个两阶段转移学习框架。在第 1 阶段，我们将出租车和公交车数据作为输入，使用出租车/公共汽车和其他车辆共享的轨迹特征来学习随机森林（RF）分类器。然后，我们使用 RF 标记所有候选汽车。在第 2 阶段，利用前一阶段高可信标签的子集作为输入，我们进一步学习用于 Ridesourcing 检测的卷积神经网络（CNN）分类器，并通过 co 迭代地改进 RF 和 CNN 以及特征集。- 训练过程。最后，我们使用由此产生的 RF 和 CNN 集合来识别候选池中的 ridesourcing 汽车。真实汽车，出租车和公共汽车轨迹的实验表明，我们的转移学习框架，无需预先标记的租赁数据集，可以达到与监督学习方法类似的准确性

23，短期乘客需求预测对于按需乘车服务平台非常重要，该平台可以激励空置汽车从供过于求的地区转移到过度需求的地区。然而，需要

同时考虑空间依赖性，时间依赖性和外生依赖性，这使得短期乘客需求预测具有挑战性。我们提出了一种新颖的深度学习（DL）方法，称为融合卷积长期短期记忆网络（FCL-Net），以在一个端到端学习架构中解决这三个依赖关系。该模型由多个卷积长短期记忆（LSTM）层，标准 LSTM 层和卷积层堆叠和融合。卷积技术和 LSTM 网络的融合使得所提出的 DL 方法能够更好地捕获解释变量的时空特征和相关性。使用定制的空间聚合随机森林来对解释变量的重要性进行排序。然后将排名用于特征选择。建议的 DL 方法适用于中国杭州按需乘坐服务平台的乘客需求短期预测。在 DiDi Chuxing 提供的真实数据上验证的实验结果表明，FCL-Net 比传统方法（包括经典时间序列预测模型和基于神经网络的算法（例如，人工神经网络和 LSTM））实现更好的预测性能。本文是首批通过检验时空相关性来预测按需乘坐服务平台的短期乘客需求的 DL 研究之一。

24，本文的目标是提出一个端到端的数据驱动框架来控制自动移动点播系统（AMoD，即自动驾驶车辆的车队）。我们首先使用时间扩展网络对 AMoD 系统进行建模，并提出计算最佳再平衡策略（即抢先重新定位）的公式以及给定旅行需求的最小可行车队规模。然后，我们调整此公式以设计模型预测控制（MPC）算法，该算法利用基于历史数据的短期需求预测来计算再平衡策略。我们使用最先进的 LSTM 神经网络测试该控制器的端到端性能，以预测 DiDi Chuxing 的客户需求和真实客户数据：我们证明这种方法可以很好地适用于大型系统（事

实上， MPC 算法的计算复杂性不依赖于客户和系统中车辆的数量），并且通过将平均客户等待时间减少高达 89.6%来优于最先进的再平衡策略。

25，出租车需求预测是智能城市智能交通系统的重要组成部分。准确的预测模型可以帮助城市预先分配资源以满足旅行需求，并减少街道上的空车，这会浪费能源并加剧交通拥堵。随着优步和滴滴出行（中国）等出租车服务的日益普及，我们能够不断收集大规模的出租车需求数据。如何利用这些大数据来改进需求预测是一个有趣且关键的现实问题。传统的需求预测方法主要依赖于时间序列预测技术，这些技术无法模拟复杂的非线性空间和时间关系。深度学习的最新进展通过学习复杂特征和大规模数据的相关性，在传统上具有挑战性的任务（如图像分类）方面表现出色。这一突破激发了研究人员探索交通预测问题的深度学习技术。然而，现有的业务量预测方法仅考虑了空间关系（例如，使用 CNN）或时间关系（例如，使用 LSTM）。我们提出了一个深度多视图时空网络（DMVST-Net）框架来模拟空间和时间关系。具体而言，我们提出的模型包括三个视图：时间视图（通过 LSTM 建模未来需求值与近时间点之间的相关性），空间视图（通过本地 CNN 建模局部空间相关性）和语义视图（共享相似时间的区域之间的建模相关性）模式）。对大型真实出租车需求数据的实验证明了我们的方法相对于最先进的方法的有效性。

26，本文提出了一种随机模型预测控制（MPC）算法，该算法利用短期概率预测来调度和重新平衡自主移动点播系统（AMoD，即自动驾驶车辆的车队）。我们首先根据时间扩展的网络流模型提出核心随机优化问题。然后，为了改善其易处理性，我们提出了两个关键的松弛。首先，我们用样本平均近似（SAA）替换原始随机问题，并描述性能保证。其次，我们将控制器分成两个独立的部分，以解决将车辆分配给与重新平衡不同的优秀客户的任务。这使问题能够作为两个完全单模式的线性程序来解决，因此可以容易地扩展到大问题大小。最后，我们基于实际数据在两种情况下测试所提出的算法，并表明它优于先前的最先进算法。特别是，在使用 DiDi Chuxing 客户数据的模拟中，与现有技术的非随机算法相比，此处提供的算法显示客户等待时间减少了 62.3％

27，人口迁移是有价值的信息，可以在城市规划战略，大规模投资和许多其他领域做出正确决策。例如，城市间迁移是一个后证据，可以看出政府对人口的限制是否有效，而社区间移民可能是房地产价格上涨的先前证据。通过及时的数据，也无法比较哪个城市对人民更有利，假设城市发布了不同的新规定，我们还可以比较不同房地产开发集团的客户，他们来自哪里，他们可能会去哪里。不幸的是，这些数据不可用。在本文中，利用滴滴定位团队产生的数据，我们提出了一种新的方法，及时监测从社区规模到省级规模的人口迁移。可以在一周内检测到迁移。它可能更快，一周的设置是出于统计目的。开发了一个

监测系统，然后在全国范围内应用，本文将介绍该系统的一些观测结果。这种新的迁移感知方法源于当今人们大多数人都使用其个人接入点（AP）（也称为 WiFi 热点）移动的洞察力。假设 AP 移动到人口迁移的比例是不变的，对比较人口迁移的分析是可行的。通过少量样本研究和模型回归，也可以进行更精确的定量研究。处理数据的过程包括许多步骤：消除伪迁移 AP 的影响，例如口袋 WiFi 和二手交易路由器;通过移动公司来区分人口迁移;通过指纹簇等识别 AP 的移位

28，最近，在线汽车服务业 Didi 已成为共享经济的领导者。乘客和司机广泛使用，对于汽车服务提供商来说，最小化乘客的等待时间和优化车辆利用率变得越来越重要，从而改善整体用户体验。因此，供需估计是高效在线汽车服务的必不可少的组成部分。为了提高估算结果的准确性，分析了本文中兴趣点（POI）与供需缺口之间的隐含关系。不同类别的 POI 对估计有正面或负面影响，本文提出了 POI 选择方案并将其纳入 XGBoost 以获得更准确的估计结果。我们的实验表明，与现有方法相比，本方法提供了更准确的估计结果和更稳定的估计结。

29，我们通过文本挖掘历史地方志，difangzhi 来展示扩大中国传记数据库内容的结果。数据库的目标是通过亲属关系，社会关系以及他们所服务的地点和办公室来了解人们如何联系在一起。地名录是宋至清时期最重要的名称和办公室。虽然我们从地方官员开始，但我们最终将包括当地考试候选人名单，当地人在政府服务，以及着名的当地人

物传记。我们收集的数据越多，出现的连接就越多。进行系统文本挖掘工作的价值在于，我们可以识别直接提供信息的相关联系，或者在没有深入的历史研究的情况下可以变得有用。中央研究院正在为明清两代中央政府的官员开发一个名称数据库。

30，我们通过文本挖掘历史地方志，difangzhi 来展示扩大中国传记数据库内容的结果。数据库的目标是通过亲属关系，社会关系以及他们所服务的地点和办公室来了解人们如何联系在一起。地名录是宋至清时期最重要的名称和办公室。虽然我们从地方官员开始，但我们最终将包括当地考试候选人名单，当地人在政府服务，以及着名的当地人物传记。我们收集的数据越多，出现的连接就越多。进行系统文本挖掘工作的价值在于，我们可以识别直接提供信息的相关联系，或者在没有深入的历史研究的情况下可以变得有用。中央研究院正在为明清两代中央政府的官员开发一个名称数据库。

31，人名和地名是识别用文学中文撰写的历史文献中的事件和社交网络的重要基石。我们率先探讨了基于语言模型和基于条件随机场的方法在历史研究中对文学中文命名实体进行算法识别的研究，并将我们的工作扩展到挖掘历史文献中的文档结构。对从 220 多卷当地地名录（地方志）中提取的文本进行了实际评估。 Difangzhi 是一个庞大而且最重要的收藏品，其中包含有关中国历史上当地政府官员的信息。我们的方法在这些实际测试中表现得非常好。从文本中确定了数以千

计的名称和地址。提取的名称的很大一部分与目前在哈佛大学中国传记数据库（CBDB）中记录的传记信息相匹配，其他许多名称可以由历史学家验证，并将成为 CBDB 的新增内容。

32，中国文学文本实体关系分类的结构规则化神经网络

作者：纪文，徐孙，任宣成，齐苏

摘要：关系分类是自然语言处理领域的一项重要语义处理任务。在本文中，我们提出了中国文学文本关系分类的任务。建立了一个新的中国文学文本数据集，以促进该任务的研究。我们提出了一种新的模型，称为结构正则化双向递归卷积神经网络（SR-BRCNN），以识别实体之间的关系。所提出的模型沿着从结构正则化依赖树中提取的最短依赖路径（SDP）学习关系表示，这具有降低整个模型的复杂性的益处。实验结果表明，该方法显着提高了 F1 得分 10.3，优于中国文学文本的最新方法。

33，中国文学文本的话语层命名实体识别与关系抽取数据集

作者：徐晶晶，季文，孙孙，齐苏

摘要：中文文本的命名实体识别和关系提取被认为是一个非常棘手的问题，部分原因在于缺乏标记集。在本文中，我们从数百篇中国文献文章中构建了一个话语层面的数据集来改进这一任务。为了构建高质量的数据集，我们提出了两种标记方法来解决数据不一致的问题，包括启发式标记方法和机器辅助标记方法。基于该语料库，我们还介绍

了几种广泛使用的模型来进行实验。实验结果不仅显示了所提出的数据集的有用性，而且为进一步的研究提供了基线。该数据集可在 https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset 获得。