AI 前沿论文最新进展

2018.10.29 方建勇

1，随着人工智能（AI）的发展，人工智能应用程序极大地影响和改变了人们的日常生活。在这里，首次提出了一种整合了情感机器人，社交机器人，大脑可穿戴设备和可穿戴设备 2.0 的可穿戴情感机器人。拟议的可穿戴情感机器人适用于广大人群，我们相信它可以在精神层面上改善人类健康，同时满足时尚要求。在本文中，从硬件和算法的角度介绍了一种被称为 Fitbot 的创新可穿戴情感机器人的架构和设计。此外，从硬件设计，脑电数据采集与分析，用户行为感知，算法部署等方面介绍了机器人 - 脑可穿戴设备的重要功能组件。然后，基于脑电图的用户行为认知是实现。通过不断获取深入，广泛的数据，我们提出的 Fitbot 可以逐步丰富用户的生活建模，使可穿戴机器人能够识别用户的意图，进一步了解用户情感背后的行为动机。 Fitbot 中嵌入的生命建模学习算法可以实现更好的用户对情感社交互动的体验。最后，讨论了可穿戴情感机器人的应用服务场景和一些具有挑战性的问题。

2，在美国，国家政策通常从国家法律开始，然后从一个国家传播到另一个国家，直到它们获得成为国家政策的动力。但是，并非每项国家政策都达到国家标准。以前的工作表明，州级政策更有可能成为国家政策，具体取决于其地理来源，立法类别或其发起国家的某些特征，如财富，城市化或意识形态自由主义。在这里，我们通过将这些特征

与州的身份分开，并建立国家政策的预测预测模型成为国家政策，来检验这些假设。使用州级政策及其特征的大型纵向数据集，我们训练模型以预测（i）政策是否成为国家政策，以及（ii）有多少国家在成为国家之前必须通过一项政策。使用这些模型作为组件，我们然后开发一个逻辑增长模型来预测当前正在扩展的州级政策何时可能在国家层面通过。我们的研究结果表明，始发国家的特征与成为国家政策并没有系统地相关，它们既不预测有多少国家必须在政策成为国家之前制定政策，也不预测它最终是否成为国家法律。相比之下，政策的州级采用的累积数量可以合理地预测政策何时成为国家。对于同性婚姻和甲基安非他明前体法律的政策，我们调查后勤增长模型如何能够预测真实国家行动的可能时间范围。我们以数据驱动的预测结束，大麻合法化和“坚持立场”的法律将成为国家政策。

3，从财务数据中提取特征是市场预测领域中最重要的问题之一，已经提出了许多方法。在其他现代工具中，卷积神经网络（CNN）最近已应用于自动特征选择和市场预测。然而，在迄今为止报告的实验中，作为提取特征的可能信息来源，不同市场之间的相关性受到的关注较少。在本文中，我们建议使用具有特别设计的 CNN 的基于 CNN 的框架，该框架可应用于来自各种来源（包括不同市场）的数据集合，以便提取用于预测这些市场未来的特征。建议的框架已用于根据各种初始特征预测第二天标准普尔 500 指数，纳斯达克指数，道琼斯指数，纽约证券交易所和俄罗斯市场指数的走势。与现有技术的基线算法相

比，评估显示预测性能有显着改善。

4，本文为股票移动预测提供了一种新的机器学习解决方案，旨在预测股票价格在不久的将来是涨还是跌。关键的新颖之处在于我们建议采用对抗性训练来改进递归神经网络模型的推广。这里对抗性训练的合理性在于，股票预测的输入特征通常基于股票价格，股票价格本质上是一个随机变量，并且随着时间的推移不断变化。因此，具有固定的基于价格的特征（例如收盘价）的正常培训可以容易地过度拟合数据，不足以获得可靠的模型。为了解决这个问题，我们建议增加扰动来模拟连续价格变量的随机性，并训练模型在小的但有意的扰动下很好地工作。对两个真实世界库存数据的广泛实验表明，我们的方法优于最先进的解决方案，平均 w.r.t 相对改善 3.11%。准确性，验证对抗训练对股票预测任务的有用性。代码将在接受后提供。

5，非接触式用户界面的设计在各种环境中越来越受欢迎。使用这样的接口，即使手脏或不导电，用户也可以与电子设备交互。此外，具有部分身体残疾的用户可以使用这样的系统与电子设备交互。由于 Leap Motion，Kinect 或 RealSense 设备等低成本传感器的出现，这方面的研究得到了极大的推动。在本文中，我们提出了一种基于 Leap Motion 控制器的方法，以便于在显示设备上渲染 2D 和 3D 形状。所提出的方法跟踪手指移动，同时用户在传感器的视野内执行自然手势。在下一阶段，分析轨迹以提取 3D 中的扩展 Npen ++特征。这些特征

表示手势期间的手指运动，并且它们被馈送到单向从左到右的隐马尔可夫模型（HMM）用于训练。提出了手势和形状之间的一对一映射。最后，使用 MuPad 界面在显示器上呈现与这些手势对应的形状。我们已经创建了由 10 名志愿者记录的 5400 个样本的数据集。我们的数据集包含 18 种几何形状和 18 种非几何形状，如“圆形”，“矩形”，“花形”，“锥形”，“球形”等。当使用 5 倍交叉评估时，所提出的方法实现了 92.87%的准确度验证方法。我们的实验表明，扩展的 3D 特征在形状表示和分类的上下文中比现有的 3D 特征表现更好。该方法可用于开发用于智能显示设备的有用的 HCI 应用。

6，分心驾驶是致命的，仅 2015 年在美国就有 3,477 人丧生。尽管在各种条件下对驾驶员的分心行为建模进行了大量研究，但使用多种模态的精确自动检测，尤其是使用语音模态来提高准确性的贡献却很少受到关注。本文介绍了一种新的用于分心驾驶行为的多模态数据集，并讨论了使用三种模态特征的自动分心检测：面部表情，语音和汽车信号。详细的多模态特征分析表明，增加更多模态可以单调增加模型的预测精度。最后，与基线 SVM 和神经网络模型相比，使用多项式融合层的简单有效的多模态融合技术显示出优越的牵引检测结果。

7，机器学习在诸如物体检测之类的计算机视觉任务上已经取得了很多成就，但是传统上使用的模型使用相对低分辨率的图像。记录设备的分辨率逐渐增加，并且对处理高分辨率数据的新方法的需求不断增

加。我们提出了一种注意流水线方法，该方法使用粗略和精细分辨率下的每个图像或视频帧的两阶段评估来限制必要评估的总数。对于这两个阶段，我们使用快速物体检测模型 YOLO v2。我们已经在代码中实现了我们的模型，它在 GPU 之间分配工作。我们保持高精度，同时在 4K 视频上达到 3-6 fps 的平均性能，在 8K 视频上达到 2 fps。

8，对于许多计算机视觉算法而言，人类情感的分类仍然是一项重要且具有挑战性的任务，特别是在人类机器人的日常生活中与人类共存的时代。当前提出的用于情绪识别的方法使用多层卷积网络来解决该任务，该网络没有明确地推断出分类阶段中的任何面部特征。在这项工作中，我们假设一种根本不同的方法来解决情绪识别任务，该方法依赖于将面部标志作为分类损失函数的一部分。为此，我们扩展了最近提出的深度对齐网络（DAN），其中包含与面部特征相关的术语。由于这种简单的修改，我们的名为 EmotionalDAN 的模型能够在两个具有挑战性的基准数据集上超过最先进的情感分类方法，最多可达 5%。此外，我们在做出决策时可视化网络分析的图像区域，结果表明我们的 EmotionalDAN 模型能够正确识别负责表达情绪的面部标志。

9，预测交通条件从在线航线的航班查询是一个具有挑战性的任务作为有许多复杂的相互作用的道路和人群参与。在本文中，我们打算提高流量预测通过适当的集成的三种隐式但重要因素编码在辅助信息。我们做到这一点在一个编码器-解码器序列学习框架中集成了以下数

据：1）脱机地理和社会属性。例如，地理结构的道路或公共社交活动，如国家的庆祝活动; 2）路交叉口的信息。一般情况下，交通拥堵发生在主要路口; 3）在线人群的查询。例如，当许多在线查询发出的相同的目的地由于一个公共的性能，交通各地的目的地将有可能成为较重在这个位置后一段时间。定性和定量实验上的真实世界的数据集百度有表现出的有效性我们的框架。

10，近似最近邻搜索（ANNS）是数据库和数据挖掘中的基本问题。可扩展的 ANNS 算法应该既高效又快速。一些早期的基于图形的方法已经显示出对搜索时间复杂度的有吸引力的理论保证，但是它们都遭受高索引时间复杂度的问题。最近，已经提出了一些基于图的方法，通过近似传统图来降低索引复杂度;这些方法在百万级数据集上取得了革命性的表现。然而，它们仍然没有扩展到十亿节点数据库。在我们的研究中，进一步提高了基于图形的方法的搜索效率和可扩展性。我们首先介绍四个方面：（1）确保图形的连通性; （2）降低快速遍历图的平均出度; （3）缩短搜索路径; （4）减小索引大小。在本文中，我们提出了一种称为单调相对邻域图（MRNG）的新颖图形结构，它保证了非常低的搜索复杂度（接近对数时间）。为了进一步降低索引复杂度并使其适用于十亿节点 ANNS 问题，我们通过近似 MRNG 提出了一种名为 Navigating Spreading-out Graph（NSG）的新图形结构。NSG 同时考虑了这四个方面。大量实验表明，NSG 显着优于所有现有算法。更重要的是，NSG 在淘宝（阿里巴巴集团）的电子商务搜索

场景中表现出色，并且已经以十亿节点的规模集成到他们的搜索引擎中。

11，亚马逊，阿里巴巴，Flipkart 和沃尔玛等电子商务网站销售数十亿产品。涉及产品的机器学习（ML）算法通常用于改善客户体验并增加收益，例如产品相似性，推荐和价格估计。在训练 ML 算法之前，需要将产品表示为特征。在本文中，我们提出了一种名为 MRNet-Product2Vec 的方法，用于在电子商务生态系统中创建产品的通用嵌入。我们学习了密集和低维度的嵌入，其中与产品相关的各种信号被明确地注入其表示中。我们训练一个判别式多任务双向递归神经网络（RNN），其中输入是通过双向 RNN 馈送的产品标题，并且在输出处，预测对应于十五个不同任务的产品标签。任务集包括关于产品的若干内在特征，例如价格，重量，尺寸，颜色，流行度和材料。我们定量和定性地评估所提出的嵌入。我们证明它们几乎与稀疏和极高维度的 TF-IDF 表示一样好，尽管 TF-IDF 维数小于 3%。我们还使用多模式自动编码器来比较来自不同语言区域的产品，并显示初步但有前景的定性结果。

12，亚马逊，阿里巴巴和 Flipkart 等电子商务公司每年处理数十亿份订单。但是，这些订单仅占所有合理订单的一小部分。探索所有合理订单的空间可以帮助我们更好地理解电子商务生态系统中各个实体之间的关系，即客户和他们购买的产品。在本文中，我们为电子商务

网站中的订单提出了生成对抗网络（GAN）。一旦经过训练，GAN 中的发电机就可以产生任意数量的合理订单。我们的贡献包括：（a）创建电子商务订单的密集和低维度表示，（b）使用真实订单培训电子商务 GAN（ecGAN）以显示拟议范例的可行性，以及（c）培训电子商务 - 条件-GAN（ec ^ 2GAN）生成涉及特定产品的合理订单。我们提出了几种定性方法来评估 ecGAN 并证明其有效性。 ec ^ 2GAN 用于涉及刚刚引入电子商务系统的产品的可能订单的各种表征。在大多数情景中，所提出的方法 ec ^ 2GAN 的性能明显优于基线。

13，地理分布式数据分析在大型组织中获取有用信息的情况越来越普遍。将现有集群规模数据分析系统简单地扩展到地理分布式数据中心的规模面临着独特的挑战，包括 WAN 带宽限制，监管限制，可变/不可靠的运行时环境以及货币成本。我们在这项工作中的目标是开发一个实用的地理分布式数据分析系统，该系统（1）采用智能机制实现工作，以便跨数据中心有效利用（调整）资源（可变环境）；（2）保证因可能的故障而导致的工作可靠性；（3）通用且灵活，足以运行各种数据分析工作，无需任何更改。为此，我们提出了一个新的通用地理分布式数据分析系统 HOUTU，它由多个自治系统组成，每个系统都在一个主权数据中心运行。 HOUTU 为每个数据中心的地理分布式作业维护一个作业管理器（JM），以便这些复制的 JM 可以单独和协作地管理资源并分配任务。我们对运行在四个阿里巴巴云区域的 HOUTU 原型的实验表明，HOUTU 提供了与现有集中式架构相当的

高效工作性能，并在面临故障时保证可靠的作业执行。

14，出价优化是在线广告中最关键的问题之一。赞助搜索（SS）拍卖，由于用户查询行为和平台性质的随机性，通常采用关键词级出价策略。相反，作为相对简单的拍卖场景，显示广告（DA）利用实时出价（RTB）来提高广告商的表现。在本文中，我们考虑赞助搜索拍卖中的 RTB 问题，称为 SS-RTB。由于随机用户查询行为和基于广告的多个关键字的更复杂的出价策略，SS-RTB 具有更复杂的动态环境。以前的大多数 DA 方法都无法应用。我们提出了一种用于处理复杂动态环境的强化学习（RL）解决方案。尽管已经提出了一些用于在线广告的 RL 方法，但它们都未能解决“环境变化”问题：状态转换概率在两天之间变化。由于观察到两天的拍卖序列在适当的聚合级别共享相似的过渡模式，我们在拍卖数据的小时聚合级别制定了一个强大的 MDP 模型，并为 SS-RTB 提出了逐个模型的框架。我们不是直接生成出价，而是决定每小时展示次数的出价模式，并相应地执行实时出价。我们还扩展了处理多代理问题的方法。我们在阿里巴巴的电子商务搜索拍卖平台中部署了 SS-RTB 系统。离线评估和在线 A / B 测试的实证实验证明了我们的方法的有效性。

15，个性化的会话销售代理商可能具有很大的商业潜力。亚马逊，eBay，JD，阿里巴巴等电子商务公司正在试用这类代理商。然而，对该主题的研究非常有限，现有的解决方案要么基于单轮 adhoc 搜索引擎，要

么基于传统的多轮对话系统。它们通常仅在当前会话中使用用户输入，忽略用户的长期偏好。另一方面，众所周知，基于推荐系统可以极大地改善销售转换率，推荐系统基于过去的购买行为来学习用户偏好并且优化诸如转换率或预期收入的面向业务的度量。在这项工作中，我们建议将对话系统和推荐系统中的研究整合到一个新颖统一的深度强化学习框架中，以构建一个个性化的会话推荐代理，优化基于每个会话的效用函数。

16，在亚马逊，淘宝和天猫等电子商务平台上的赞助搜索为卖家提供了一种有效的方式来吸引具有最相关目的的潜在买家。在本文中，我们研究了阿里巴巴移动电子商务平台上的赞助搜索中的拍卖机制优化问题。除了创造收入外，我们还应该保持一个拥有大量优质用户的高效市场，保证广告商的合理投资回报率（ROI），同时为用户提供愉快的购物体验。这些要求实质上构成了约束优化问题。直接优化拍卖参数会产生不连续的非凸问题，否定有效的解决方案。我们的主要贡献之一是原始问题的实际凸优化公式。我们利用离散的代表性实例设计了一种新颖的拍卖机制重新参数化。为了构建优化问题，我们建立了一个拍卖模拟系统，通过重放从真实在线请求记录的拍卖来估计所选参数的结果业务指标。我们总结了实际搜索流量的实验，分析了拍卖模拟保真度的影响，各种约束目标下的效能以及正规化的影响。实验结果表明，通过适当的熵正则化，我们能够在限制给定范围内的其他业务指标的同时最大化收益。

17，仓库规模的云数据中心将具有不同且通常互补的特征的工作负载共同定位，以提高资源利用率。为了更好地了解管理此类错综复杂的异构工作负载所面临的挑战，同时提供有质量保证的资源协调和用户体验，我们分析了阿里巴巴的共址工作负载跟踪，这是第一个公开可用的数据集，其中包含有关每个作业类别的精确信息。两种类型的工作负载 - 长时间运行，面向用户，容器化生产作业，以及瞬态，高度动态，非容器化和非生产批处理作业---正在 1313 台机器的共享集群上运行。我们的多方面分析揭示了我们认为对从事集群管理系统的系统设计人员和 IT 从业人员有用的见解。

18，如今，像亚马逊，阿里巴巴甚至披萨连锁店这样的公司正在推动使用无人机（也称为无人机）来提供服务，例如包装和食品配送。由于政府打算利用无人机必须提供的巨大经济效益，城市规划者正在向智能城市设计中采用所谓的无人机飞行区和无人机高速公路。然而，需要监视无人机的高速移动性和行为动态以检测并随后以恶意目的处理入侵者，流氓无人机和无人机。本文提出了一种无人机防御系统，用于拦截和护送飞行区外的恶意无人机。所提出的无人机防御系统包括防御无人机群，其能够在入侵者检测的情况下自我组织其防御形成，并且将恶意无人机作为网络群追逐。模块化设计原则已用于我们的完全本地化方法。我们开发了一种创新的自动平衡聚类过程，以实现拦截和捕获形成。事实证明，由此产生的网络防御无人机群体可以抵御

通信损失。最后，实现了原型 UAV 模拟器。通过广泛的模拟，我们展示了我们的方法的可行性和性能。

19，在本文中，我们根据这些用户的人口统计（年龄，性别和位置）和社交（友谊，互动和群组成员）信息，挖掘并学习预测用户对视频的兴趣有多相似。我们使用活跃用户的视频访问模式作为基本事实（一种基准形式）。我们采用基于标签的用户分析来建立这个基本事实，并证明为什么使用它而不是基于视频的方法，或许多潜在的主题模型，如 LDA 和协作过滤方法。然后，我们基于结合多个特征的不同机器学习方法，显示不同人口统计和社会特征及其组合和衍生物在预测用户兴趣相似性方面的有效性。我们提出了一种混合树编码的线性模型，用于组合这些特征，并表明它优于其他线性和基于树的模型。当地面实况不可用时，我们的方法可用于预测用户兴趣相似性，例如，对于新用户或其兴趣可能已从旧访问数据更改的非活动用户，对于视频推荐非常有用。我们的研究基于来自腾讯的丰富数据集，腾讯是中国社交网络，视频服务和各种其他服务的热门服务提供商。

20，目前，人工智能（AI）已经获得了前所未有的关注，并且正在成为中国越来越受欢迎的焦点。这种变化可以通过令人印象深刻的学术出版物记录，国家级投资的数量以及全国范围内的参与和投入来判断。在本文中，我们重点讨论了中国人工智能工程的进展。我们首先介绍了中国学术界对人工智能的关注，包括超级计算脑系统，寒武纪神经

网络超级计算机和生物识别系统。然后，介绍了工业界人工智能的发展以及百度，腾讯，阿里巴巴等公司的最新 AI 产品布局。最后，我们引入了中国主要知识分子关于人工智能未来发展的观点和论点，包括如何审视人性与科学技术之间的关系。

21，由于它们的低分辨率和嘈杂的表现，检测小物体是众所周知的挑战。现有的对象检测管道通常通过学习多个尺度的所有对象的表示来检测小对象。然而，这种 ad hoc 架构的性能增益通常限于偿还计算成本。在这项工作中，我们通过开发一个单一的体系结构来解决小对象检测问题，该体系结构内部将小对象的表示提升为“超分辨”对象，实现与大对象类似的特性，从而更加区分检测。为此，我们提出了一种新的感知生成对抗网络（Perceptual GAN）模型，该模型通过缩小小对象与大对象的表示差异来改进小对象检测。具体来说，它的生成器学习将小对象的感知不良表示转移到与现实大对象足够相似的超分辨表示，以欺骗竞争的鉴别器。同时，其鉴别器与发生器竞争以识别所生成的表示并且施加额外的感知要求 - 生成的小对象的表示必须有益于检测目的 - 在发生器上。对具有挑战性的清华 - 腾讯 100K 和加州理工学院基准测井的广泛评估证明了感知 GAN 在检测包括交通标志和行人在内的小物体方面的优越性，而不是完善的现有技术水平。

22，识别两个文本对象之间的关系是许多自然语言处理任务的核心研

究问题。为文本匹配提出了广泛的深度学习方案，主要集中在句子匹配，问答或查询文档匹配上。我们指出现有方法在匹配长文档方面表现不佳，这对于例如基于 AI 的新闻文章理解和事件或故事形成是至关重要的。原因是这些方法要么省略要么不能在长文档中充分利用复杂的语义结构。在本文中，我们提出了一种文本匹配的图表方法，特别是针对长文档匹配，例如识别两篇新闻文章是否在现实世界中报告相同的事件，可能具有不同的叙述。我们提出概念交互图以产生文档的图形表示，其中顶点表示不同的概念，每个概念是文档中的一个或一组连贯的关键字，并且边缘表示不同概念之间的交互，通过文档中的句子连接。基于文档对的图形表示，我们进一步提出了一种 Siamese 编码图形卷积网络，该网络通过 Siamese 神经网络学习顶点表示，并通过图形卷积网络聚合顶点特征以生成匹配结果。基于腾讯为其智能新闻产品创建的两个标记的新闻文章数据集，对所提出的方法进行了广泛的评估，结果表明，建议的长文档匹配图表方法明显优于广泛的最新方法。

23，我们描述了我们在腾讯实施新闻内容组织系统的经验，该系统以大量突发新闻发现事件，并以在线方式发展新闻故事结构。与先前关于主题检测和跟踪（TDT）以及事件时间线或图形生成的研究相比，我们的真实世界系统具有不同的要求，因为我们 1）需要准确快速地从大量长文本文档流中提取可区分的事件。多样化的主题，包含高度冗余的信息，2）必须以在线方式开发事件故事的结构，而不是反复

重组以前形成的故事，以保证一致的用户观看体验。在解决这些挑战时，我们提出了 Story Forest，这是一套在线方案，可以自动将流媒体文档聚合到事件中，同时连接成长树中的相关事件来讲述不断变化的故事。我们基于 60 GB 的真实中文新闻数据进行了广泛的评估，尽管我们的想法不依赖于语言，并且可以通过详细的试用用户体验研究轻松扩展到其他语言。结果表明，与多个现有算法框架相比，Story Forest 能够准确识别事件并将新闻文本组织成一个吸引人类读者的逻辑结构。

24，自动驾驶汽车上的交通信号灯和标志探测器是道路场景感知不可或缺文献中有丰富的深度学习网络，可以检测灯光或标志，而不是两者，这使得它们不适合实际部署，因为嵌入式系统上的图形处理单元（GPU）内存和功率有限。此问题的根本原因是没有公共数据集包含交通灯和标志标签，这导致难以开发联合检测框架。我们提出了一个深层次结构与小批量提案选择机制相结合，允许网络检测交通灯和来自单独交通信号灯和标志数据集的培训标志。我们的方法解决了重叠问题，即一个数据集中的实例未在另一个数据集中标记。我们是第一个提供对交通信号灯和标志进行联合检测的网络。我们在清华 - 腾讯 100K 交通标志检测基准和博世小交通灯交通灯检测基准上测量我们的网络，并显示它优于现有的博世小交通灯最先进的方法。我们专注于自动驾驶汽车，并且由于其低内存占用和实时图像处理时间，我们的网络比其他网络更合适。可以在 https://youtu.be/_YmogPzBXOw

查看定性结果

25，同样，从人口统计到情绪，在社交网络中，无论是离线还是在线，都会产生联系。然而，随着音乐流媒体服务的蓬勃发展，在线音乐听力是否存在同质性仍不清楚。在本研究中，分别在网易音乐和微博中建立了同一组活跃用户的两个在线社交网络。通过呈现多个相似性度量，可以明显地证明在两个在线社交网络的音乐收听中确实存在同音异义。微博意想不到的音乐相似性也意味着来自通用社交网络的知识可以自信地转移到面向域的网络，用于上下文丰富和算法增强。进一步探讨了可能在形成同性恋中起作用的综合因素，并且深刻揭示了许多有趣的模式。结果发现，女性朋友在音乐聆听方面更加同质化，积极而充满活力的歌曲显着拉近了用户。我们的方法和研究结果将为在线音乐服务中的现实应用提供信息

26，中国是世界上最大的 Android 市场之一。由于中国用户无法访问 Google Play 购买和安装 Android 应用程序，因此出现了许多独立的应用程序商店，并在中国应用程序市场上展开竞争。一些中国应用程序商店是预先安装的特定于供应商的应用程序市场（例如，华为，小米和 OPPO），而其他应用程序商店则由大型科技公司（例如，百度，奇虎 360 和腾讯）维护。这些应用程序商店的性质和通过它们提供的内容差别很大，包括其可信赖性和安全性保证。截至今天，研究界尚未深入研究中国 Android 生态系统。为填补这一空白，我们推出了首个

大型比较研究，涵盖了从 16 个中国应用市场和 Google Play 下载的 600 多万个 Android 应用。我们的研究主要集中在应用商店的目录相似性，它们的功能，发布动态以及各种形式的不当行为（包括虚假，克隆和恶意应用程序的存在）的普遍存在。我们的调查结果还表明，在代码维护，第三方服务的使用等方面，跨应用商店的异构开发人员行为。总体而言，中国应用程序市场在采取积极措施保护移动用户和合法开发人员免受欺骗性和滥用行为者影响时表现更差，显示恶意软件，假冒和克隆应用程序的流行程度明显高于 Google Play。

27，本文对中世纪中国手稿（Or.8210 / S.3326）中的星图集进行了分析，该手稿于 1907 年由考古学家 Aurel Stein 在丝绸之路敦煌镇发现，现在在英国图书馆举行。虽然少数中国学者对其进行了部分研究，但在西方世界从未充分展示和讨论过。这组天空地图（准圆柱投影中的 12 小时角度地图和方位角投影中的极地地图）显示了从北半球可见的完整天空，这是迄今为止最古老的完整保存星图集来自任何文明。它也是中国星座准整体的第一个已知图形表示。本文描述了物理对象的历史 - 一卷用墨水绘制的薄纸。我们分析每张地图的恒星内容（1339 颗星，257 颗星）以及与地图相关的文本。我们建立绘制地图的精度（最亮的恒星为 1.5 到 4 度）并检查所用投影的类型。我们得出结论，使用精确的数学方法来制作地图集。我们还讨论了手稿的年代及其可能的作者，并根据现有证据确认 649-684（早期唐朝）的日期是最有可能的。这与先前估计值+940 左右不一致。最后，我们将对中国和欧洲

的后期天空图进行简要比较。

28，唐（公元 618 - 907 年）和宋（960-1279）朝代是中国文学发展的两个非常重要的时期。唐宋时期最具影响力的诗歌形式分别是诗词和词典。唐诗和宋词奠定了中国文学的重要基础，其对中国社会文学作品和日常生活的影响一直持续到今天。我们可以从不同的角度分析和比较完整的唐诗和完整的宋词。在本演示文稿中，我们报告了我们关于其词汇表差异的发现。有趣的新词开始出现在宋词中并继续用于现代汉语中。色彩是诗歌意象的重要组成部分，我们讨论唐诗和宋词中出现的最常见的色彩词。

29，唐，明，清两代王朝的中国陵墓，令人震惊的纪念碑，构思和建造，以确保皇帝在地球上来世和永恒的名声不朽。为此目的，在为这些纪念碑选择的丧葬景观中体现了一系列认知元素，包括天文学，一般地形和中国传统风水。利用卫星图像，我们以一般方式在此研究此问题。特别是，我们开发并应用严格的方法来调查是否在规划这些纪念碑时使用了磁罗盘。

30，研究表明，序列到序列的神经模型，特别是那些具有注意机制的神经模型，可以成功地产生中国古典诗歌。然而，神经模型不能产生符合特定风格的诗歌，如唐代着名诗人李白的冲动风格。这项工作提出了一个记忆增强神经模型，以产生风格特定的诗歌。关键思想是一

种记忆结构，用于存储人类如何生成具有所需风格的诗歌，并使用类似的片段来调整生成。我们证明了所提出的算法能够生成具有灵活风格的诗歌，包括特定时代的风格和个体诗人。

30，在本文中，我们开发了一种低于字符的特征嵌入，称为基本嵌入，并将其应用于 LSTM 模型，用于前现代汉语文本的句子分割。这些数据集包括来自 3 个不同朝代的 150 多本经典中文书籍，包含不同的文学风格。 LSTM CRF 模型是序列标记问题的现有方法。我们的新模型增加了激进嵌入的组件，从而提高了性能。基于上述中文书籍的实验结果表明，与早期的句子分割方法相比，特别是在 Tang Epitaph 文本中具有更好的准确性

31，像报纸档案这样的纵向语料库对历史研究具有极大的价值，时间作为历史学家的一个重要因素，对这些档案馆中的搜索行为产生了强烈的影响。在搜索随时间发布的文章时，关键的偏好是从重要的时间点检索涵盖重要方面的文档，这些文档与标准搜索行为不同。为了支持这种搜索策略，我们引入了历史查询意图的概念，以明确地建模历史数据库的搜索任务，并定义新闻档案的方面时间多样化问题。我们提出了一种新的算法，HistDiv，它基于历史学家的信息搜寻行为明确地模拟方面和重要的时间窗口。通过结合基于出版时间和时间表达的时间先验，我们在方面和时间维度上进行多样化。我们通过构建基于纽约时报收集的测试集合来测试我们的方法，其中包含 30 个手动评

估的历史意图查询。我们发现 HistDiv 在副主题召回方面优于所有竞争对手，但精度略有下降。我们还提供定性用户研究的结果，以确定这种精度下降是否对用户体验有害。我们的结果显示，用户仍然更喜欢 HistDiv 的排名。

32，高考是中国大陆大学入学的年度学历资格考试。高考由各省级行政区（PAR）组织，于 6 月初在全国同时举行。为了在 9 月份入读大学，学生必须在 7 月份带高考并提交入读 PAR 高考办公室的常用申请，列出他们打算参加和学习的少数几个大学和专业。每年约有 950 万高中毕业生参加高考，而高考的成绩仅为一年。如果学生的高考分数优于他们在申请中选择的大学的录取分数，则学生很有可能被录取。然而，在填写申请时，大学的录取分数是未知的，这些申请将在录取过程中动态确定，并且可能每年波动。为了增加他们接受最适合大学的机会，学生需要预测他们感兴趣的每所大学的录取分数。早期预测方法是经验性的，没有深入数据研究的支持。我们通过基于 PAR 中高考分数的排名提供经过充分测试的数学模型来填补这个空白。我们表明，我们的方法明显优于教师和专家常用的方法，并且可以在 750 分评分等级的考试中在 7 分的范围内预测准确率为 91%的入学分数。

33，随着两岸形势的迅速演变，作为社会科学研究主题的“中国大陆”最近在知识分子中引发了“反思中国研究”的声音。本文试图将自动内容分析工具（CATAR）应用于“中国大陆研究”（1998-2015）期刊，以

便根据期刊中每篇论文的标题和摘要中的文本聚类来观察研究趋势。 。结果显示，该期刊发表的 473 篇文章分为七个主题。从每个主题的出版物编号（包括“出版物的数量”,“出版物的百分比”）来看，该期刊有两个主题，而其他主题随着时间的推移而变化很大。本研究的贡献包括：1。我们可以将每个“独立”研究分组为一个有意义的主题，因为小规模实验证实该主题聚类是可行的。本文揭示了台湾期刊“中国大陆研究”的重大研究课题及其发展趋势。 3.确定了各种主题关键词，便于访问过去的研究。 4.所确定主题的年度趋势可被视为未来研究方向的标志。

34，最近的研究表明，社交媒体平台能够影响股价走势的走势。然而，现有的作品主要关注美国股市，并且缺乏对中国等某些新兴国家的关注，而中国则是散户投资者主导市场。在这方面，由于散户投资者容易受到新闻或其他社交媒体的影响，因此从社交媒体平台中提取的心理和行为特征被认为可以很好地预测中国市场的股票价格变动。中国投资者社交网络的最新进展使得能够从网络规模数据中提取这些特征。在本文中，基于来自专注于投资者的中国流行 Twitter 社交平台雪球的推文，我们分析了集体情绪和股票相关感知的特征，并通过采用非线性模型预测股票价格变动。感兴趣的特征证明在我们的实验中是有效的。

35，传记数据库包含有关个人的各种信息。人名，出生信息，职业，

朋友，家庭和特殊成就是个人记录中的一些可能项目。个人之间的关系，如亲属关系和友谊，提供了关于隐藏社区的宝贵见解，这些社区并未直接记录在数据库中。我们证明了一些简单的矩阵和基于图的操作对于推断个体之间的关系是有效的，并用中国传记数据库（CBDB）说明了主要思想。

36，由于快速的城市化，中国遇到了严重的土地流失问题以及城市扩张。在不考虑复杂的时空异质性的情况下，以前的研究无法很好地提取大规模的城市转型规则。本研究提出了一种基于随机森林算法（RFA）的细胞自动机（CA）模型，以模拟 2000 年至 2030 年间中国的城市扩张和农田流失。本研究的目的是：1）不同均质的城市转换规则。经济发展区; 2）以高空间分辨率（30 米）模拟中国的城市扩张过程和农田损失。首先，根据官方统计数据，我们在中国聚集了几个同质的经济发展区域。其次，我们构建了一个基于 RFA 的 CA 模型来挖掘复杂的城市转换规则，并对每个同区的城市扩张和农田损失进行了模拟。该模型在位于中国广州的天河一号超级计算机上实现。精度评估表明，基于 RFA 的 CA 模型的仿真结果更符合实际土地利用变化。本研究证明，中国农田流失的主要因素是 2000 年以来的快速城市化，预计农田流失率将逐步放缓，并将从 2010 年到 2030 年稳定下来。这表明中国能够保护 120 万公里的农田。在未来 20 年内没有越过“红线”，但情况仍然严峻

37，在线社交媒体，如 Twitter 或其变体微博，如何与股票市场互动，以及它是否可以成为预测股市的有力代理，多年来一直争论不休，特别是对中国而言。正如行为金融学中的传统理论所说，个体情感可以影响投资者的决策，从网络情感的角度系统地进一步探讨这些有争议的话题是合理的，这些话题在社交媒体中得到了大量的推文。通过对超过 1000 万股与股票相关的推文和来自微博的 300 万投资者的深入研究，发现情绪波动较大的缺乏经验的投资者对市场波动比经验或机构投资更为明智，他们的主导职业也表明与西方同行相比，中国市场可能更具情感。然后，相关性分析和因果关系检验都表明，中国股票市场的五个属性可以通过各种在线情绪进行预测，如厌恶，喜悦，悲伤和恐惧。具体而言，所呈现的预测模型明显优于基线模型，包括以纯金融时间序列作为输入特征的基线模型，在 K 均值离散化下预测股票市场的五个属性。我们还在现实的在线应用场景中使用了这种预测模型，并进一步证明了其性能。

38，识别相同语言的不同变体比不相关的语言识别更具挑战性。在本文中，我们提出了一种方法，用于区分中国大陆，香港，台湾，澳门，马来西亚和新加坡，大中华区（GCR）的语言品种或方言。当应用于 GCR 的方言识别时，我们发现常用的字符级或单词级单字符特征不是非常有效，因为存在几个特定问题，例如单词中的歧义和依赖于上下文的特征。 GCR 的方言。为了克服这些挑战，我们不仅使用字符级 n-gram 等一般功能，还使用许多新的字级功能，包括基于 PMI 和

基于字对齐的功能。来自维基百科的新闻和开放域数据集的一系列评估结果显示了所提出方法的有效性。

39，唐宋诗歌之间的大规模比较揭示了诗人如何使用和分享词语，搭配和表达方式。有些词只用在唐诗中，有些只用在宋诗中，可以引发有趣的语言学研究。唐宋诗歌中最常见的颜色不同，提供了一个朝代不断变化的社会环境的痕迹。当前工作的结果与词典编纂，语义学和社会转型的研究课题相关联。我们讨论了我们的发现，并提出了我们的算法，以便在诗歌之间进行有效的比较，这对于在可接受的时间内完成数十亿次比较至关重要。