



1993 至 2019 年人工智能研究的数据分析报告

阴钰骐, 梁昊, 田海煜

Abstract

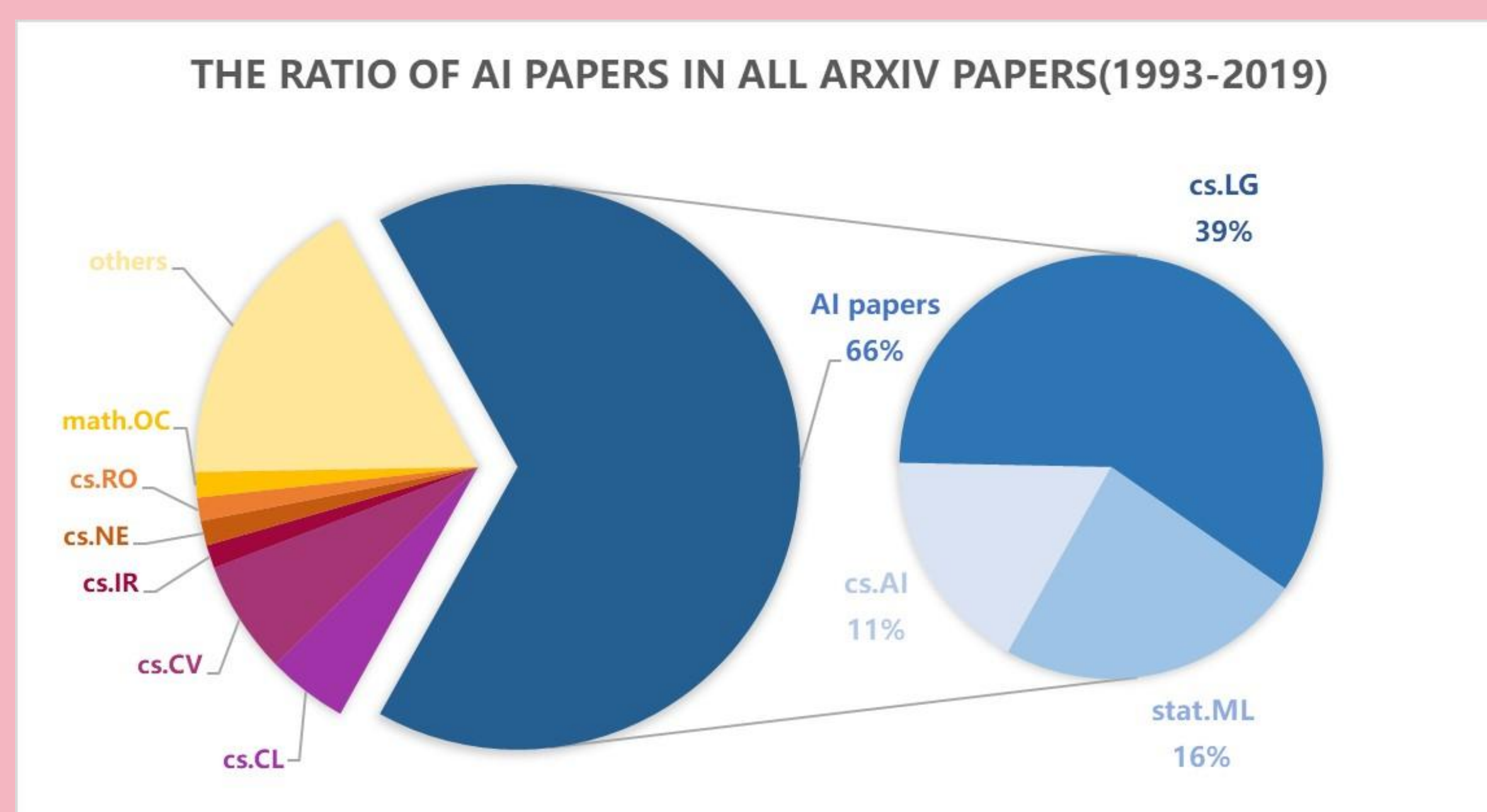
本报告概述了1993-2019年间人工智能领域论文数据集的探索过程及其发现。数据集主要包含人工智能 (cs.AI)、机器学习 (cs.LG) 和统计机器学习 (stat.ML) 三个类别的论文。我们旨在通过对时间序列数据进行可视化, 观察这些年人工智能领域的特点的变化, 比如研究趋势和热点话题。通过一些总的统计性比例指标, 我们也可以分析不同机器学习领域的风格差别、以及人工智能细分领域间的相似性。在数据分析的基础上, 我们借助html, javascript等软件将可视化数据进行美化、并且放到网页上实现了互动效果。

1 数据分析与预处理

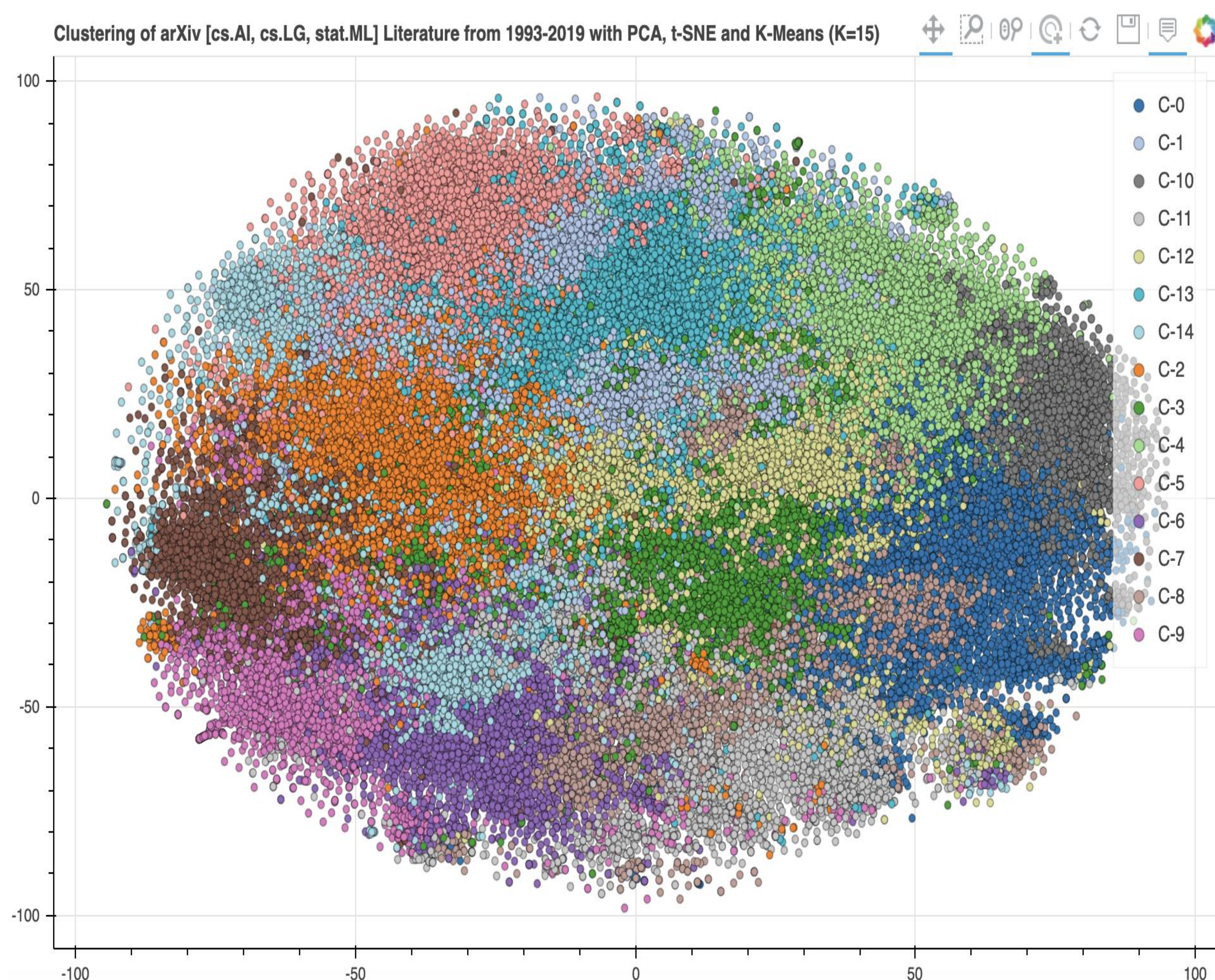
数据集包含了对1993年至2019年间在arXiv.org上发布的研究论文。我们考虑的类别已限定为:

- 计算机科学: 人工智能 [cs: AI]
- 计算机科学: 机器学习 [cs: LG]
- 统计学: 机器学习 [stat: ML]

筛选后得到clean_data.csv

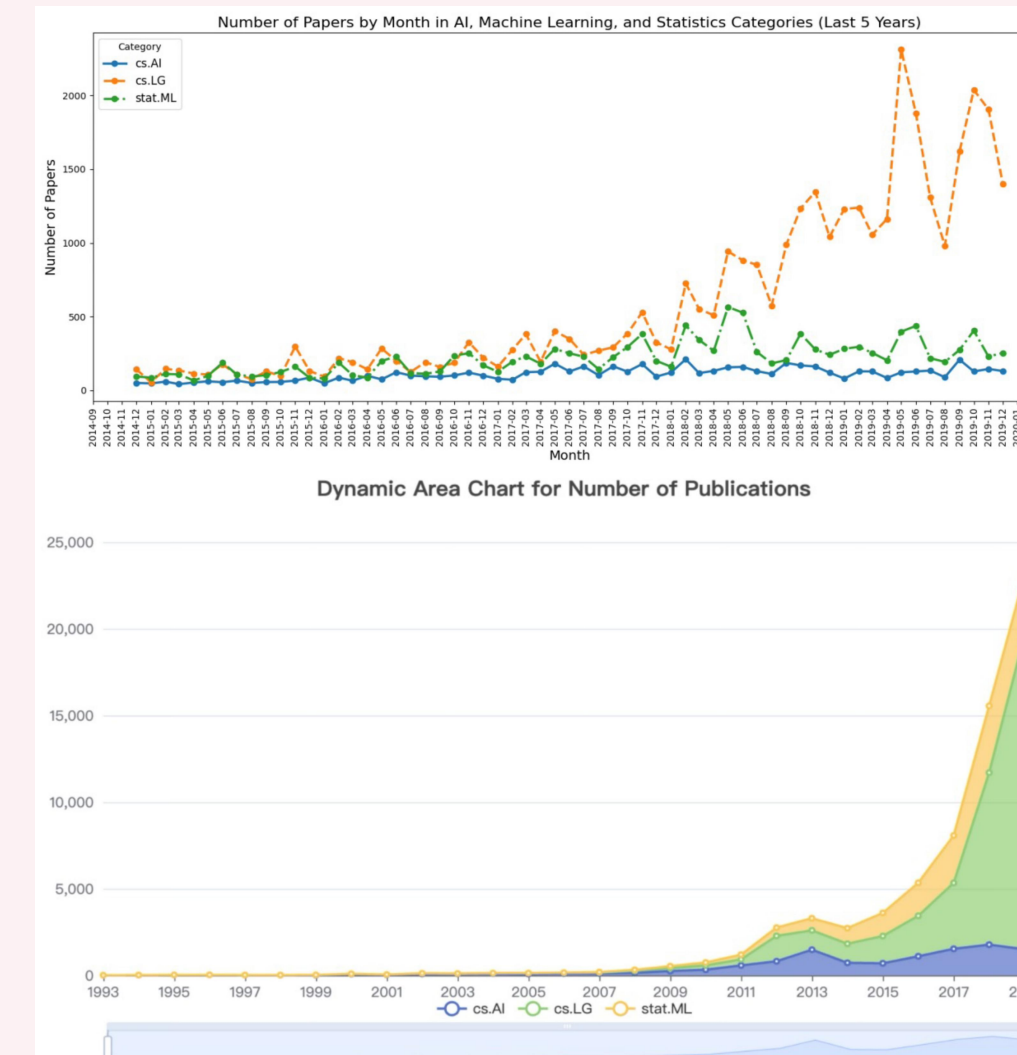


我们观察到在1993-2019 年Arxiv里的所有论文中, AI领域及其相关论文占了较大部分。造成这一现象的主要原因有以下几个方面:对于实验性学科而言, 大多数研究者会将注意力放在各类期刊上, 鲜少有人会选择Arxiv这类预印本网站作为论文发表的平台;而对于数理以及传统计算机科学而言, 尽管有不少研究者青睐于这一平台, 但由于学科的热度相对AI领域较为“不温不火”, 因此论文更新速度远不如AI领域快速, 研究者并不需要在Arxiv上提前登出论文以便宣示自己的创意“主权”。故而导致AI与相关领域论文占比高达Arxiv上全部论文的2/3。



2 数据可视化

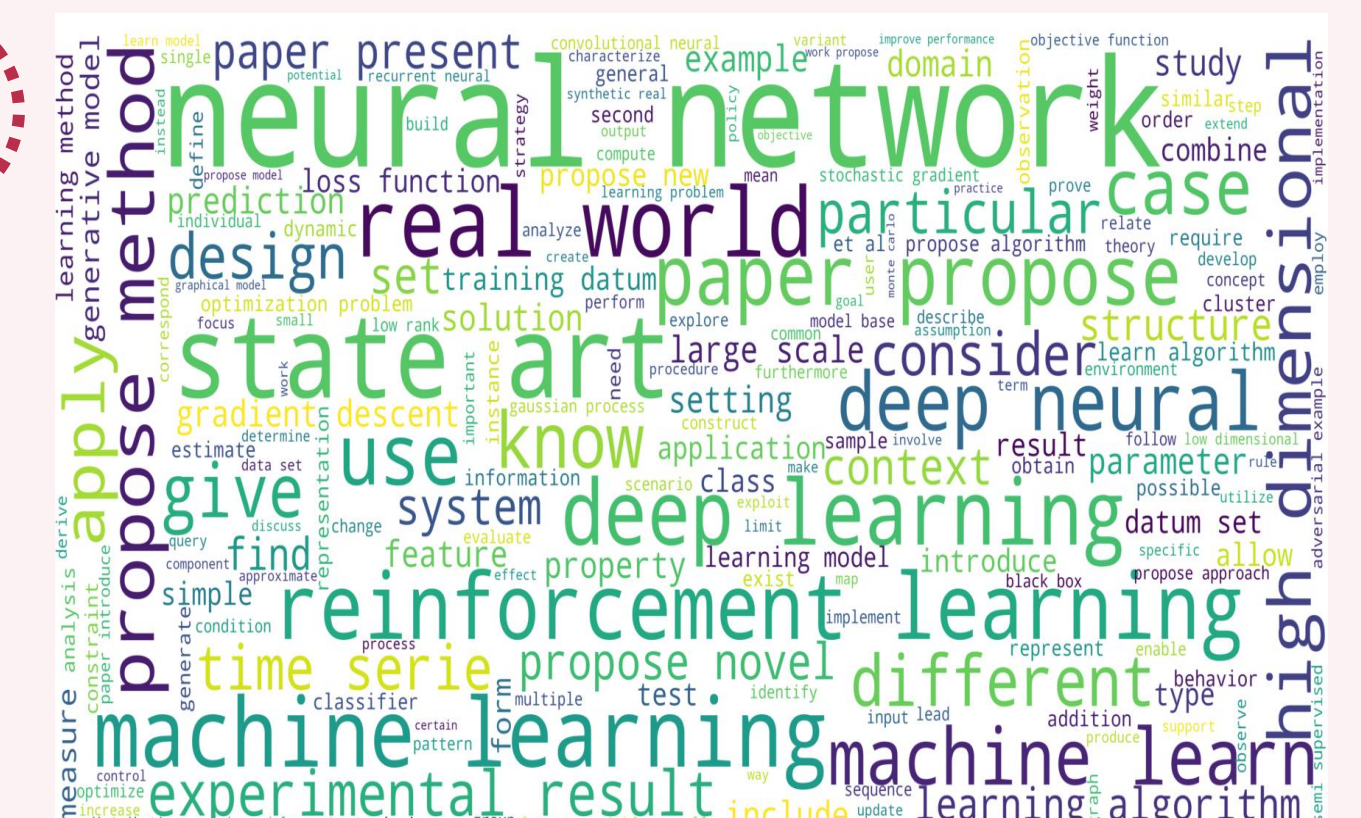
2.1 论文数量可视化



时间序列分析显示, 近五年来, cs.LG类别的论文数量显著增加, 这可能反映了机器学习算法在学术界的兴起, 确实也与我们最近几年感受到的人工智能算法大爆炸相符合。

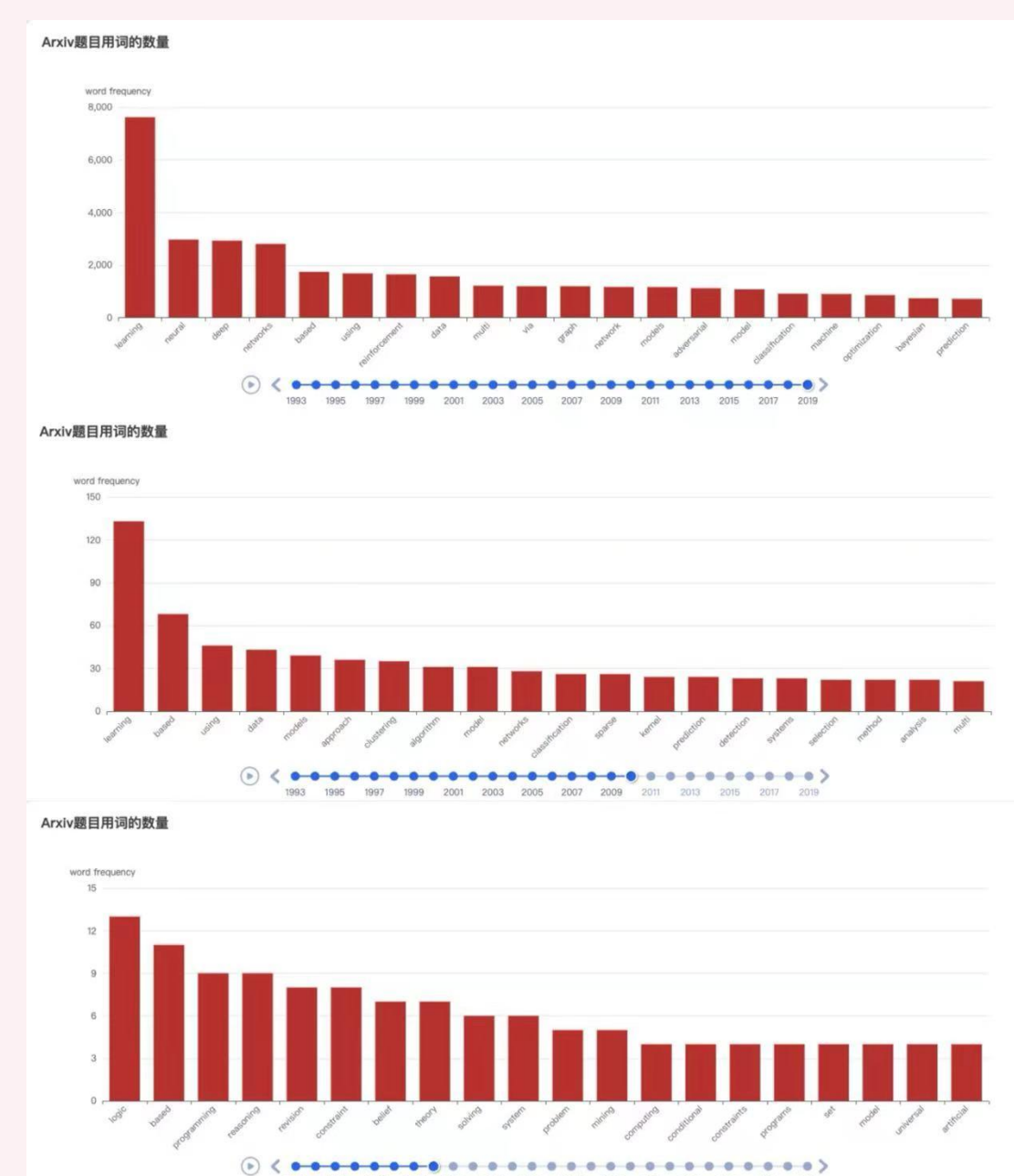
2.2 论文关键词可视化

词云分析, 观察到“neural network(神经网络)”、“machine learning(机器学习)”、“deep learning(深度学习)”、“reinforcement learning(强化学习)”等词汇的高频出现表明了这些是当前研究中的热点



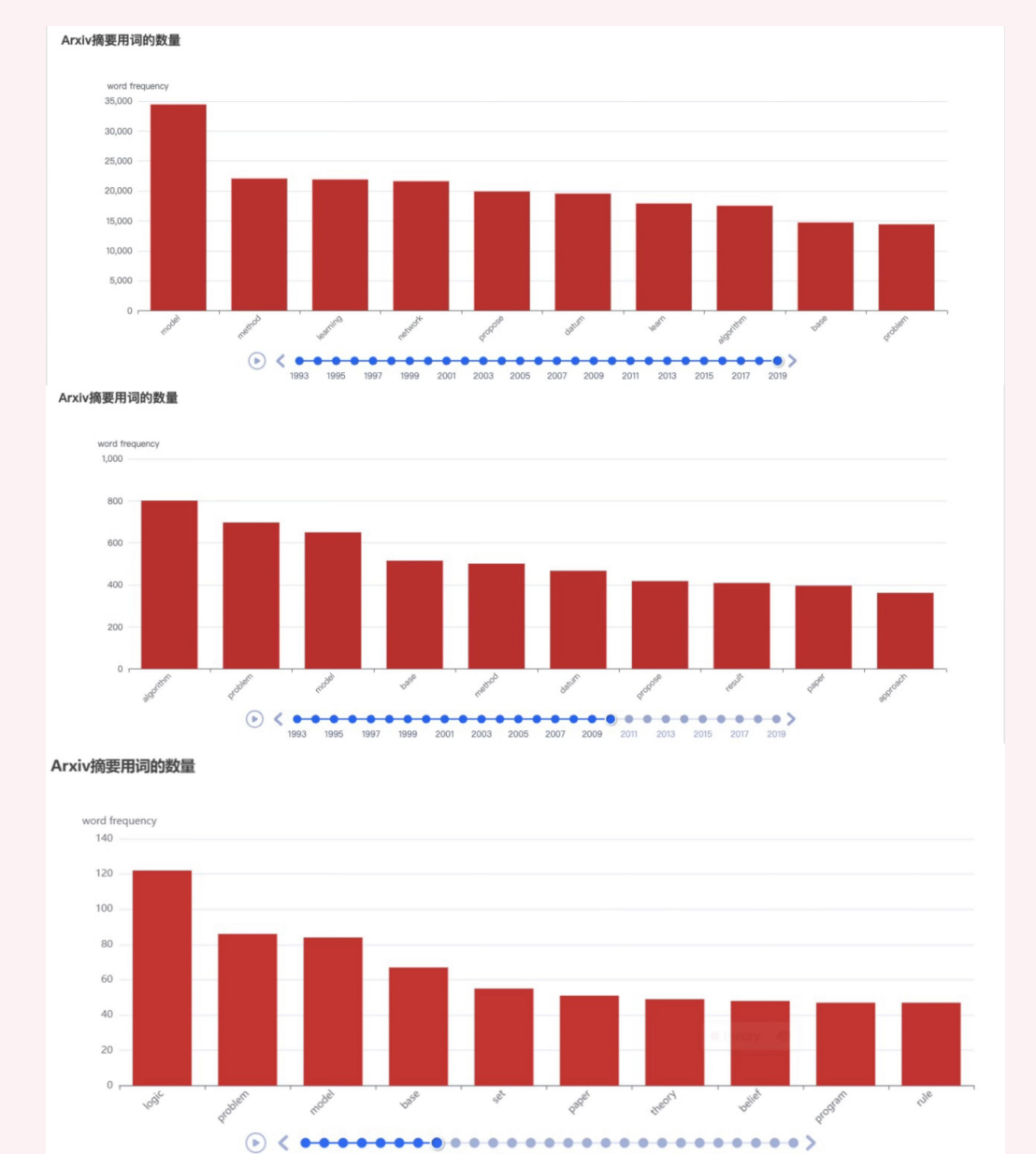
题目关键词

我们通过三个时间点: 2019年, 2010年以及2000年来说明人工智能研究热点的变化。我们可以看到在2000年的时候, 人们更多关注的是模型的logic, 而到了2010年, 人们的关注点就从logic转变成了learning, 到了2019年, 大家则更关心learning, data这些和神经网络模型学习以及数据相关的学术词汇。



摘要关键词

我们通过三个时间点: 2019年, 2010年以及2000年来说明人工智能研究热点的变化。我们可以看到在2000年的时候, 人们更多关注的是模型的logic, 而到了2010年, 人们的关注点就从logic转变成了algorithm, 到了2019年, 大家则更关心model, method, learning, network这些和神经网络模型相关的学术词汇。



论文关键词可视化额外功能

我们的交互模型支持查看关键词在不同年份中的使用率排名, 该功能覆盖了题目关键词和摘要关键词。

如右图所示, 题目关键词learning与摘要关键词model均在多年前就很重要, 随着时间的推移, 重要性越来越强, 直到最近若干年每年霸占使用率排名第一。



3 摘要分析网页交互

设计

1. 首先筛选有效的摘要;
2. 将摘要转化成向量, 使用Bert Base模型;
3. 我们对摘要进行聚类分析, 采用KMeans聚类方法进行聚类, 用Davies Bouldin score来对聚类效果进行评估;
4. 我们将聚类结果通过Bokeh库来生成网页文件。

使用

1. 可以针对感兴趣的关键词搜索相关的论文并可视化关键词和摘要;
2. 可以选定一个可视化的点并查看相应论文的关键词和摘要;
3. 可以查看一个聚类类别的关键词有哪些, 方便大家更清晰的了解一个领域。