基于 MTCNN 和 XGBoost 的学者画像构建研究

韩 普[1,2]　杨博凡[1]　仲雨乐[1]　陆豪杰[1]

[1]（南京邮电大学　管理学院，南京 210003）

[2]（江苏省数据工程与知识服务重点实验室，南京 210023）

摘要：学者画像构建对信息检索和推荐系统具有重要意义。为了更有效地抽取用户画像各维度信息，本文提出了一种基于多任务卷积神经网络（MTCNN）、深度残差神经网络（ResNet）和 XGBoost 模型相结合的学者画像构建方法。首先对维基、谷歌等网站中目标学者网页进行预处理，获取目标学者姓名、邮箱、职称和主页网址等属性；其次采用 XGBoost 模型从多源异构的搜索结果网页中识别出学者主页；接着利用 MTCNN 对预测主页图片进行人脸识别以提取学者肖像，然后基于肖像信息使用 ResNet 预测学者性别；最后使用 Flair 框架通过学者姓名预测其所属国籍，进而预测学者使用的自然语言。方法在 CCKS2021: AMiner 学者画像任务评测中获得第二名，验证了方法的有效性及可行性。

关键词：学者画像；实体抽取；MTCNN；Resnet；XGBoost

Research on Scholar Portrait Construction Based on MTCNN and XGBoost

Han Pu[1,2]　Yang Bofan[1]　Zhong Yule[1]　Lu Haojie[1]

[1](School of Management, Nanjing University of Posts & Telecommunications, Nanjing 210003)

[2](Jiangsu Provincial Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023)

Abstract: Scholar portrait construction is of great significance to information retrieval and recommendation systems. In order to extract the information of various dimensions of user profiling more efficiently, this paper proposes a method for constructing scholar profiling based on the combination of multi-task convolutional neural network (MTCNN), deep residual neural network (ResNet) and XGBoost model. Firstly, we preprocess the target scholar's webpage in Wiki, Google and other websites to obtain the target scholar's name, email address, title and homepage URL; Secondly, we use the XGBoost model to identify the scholar's homepage from the multi-source heterogeneous search result webpage; Then we use MTCNN to perform face recognition on the predicted homepage picture to extract the scholar's portrait, in addition, we use ResNet to predict the scholar's gender based on the portrait information; Finally, we use the flair toolkit to predict the nationality of the scholar through the name of the scholar, and then predict the natural language used by the scholar. The method won the second place in the CCKS2021: AMiner Scholar Profiling Task Evaluation, which verified the effectiveness and feasibility of the method.

Keywords: Scholar portrait, Entity extraction, MTCNN, Resnet, XGBoost.

1 引言

随着互联网技术的迅猛发展和人们对科学技术知识传播的需求，网络上积累了海量的学者信息，如何从繁冗、分布零散的学术信息中构建多维度学者精准画像成为了近些年学界关注的热点。学者画像是建立在一系列真实数据上能够客观、全面和精确地呈现精简的目标学者模型，它可以从多个维度呈现学者的基本信息、研究方向和社交关系，对学术同行分类、提升学术影响力、学者网络构建、研究成果共享、信息检索和专家推荐系统具有重要意义[1-3]。

尽管学者画像构建受到了学界广泛关注，但开放互联网中的学者画像面临着数据量巨大，且存在大量数据噪音和数据冗余等新挑战，传统的学者画像理论、模型和方法无法直接移植到开放互联网中的学者画像系统[2]。针对开放互联网中的学者画像构建，不少学者在理论、模型和方法上进行了开拓性的探索。Tang[4]在 2008 年开发的 ArnetMiner 是比较知名的学者画像工具，它通过不同权重的主题表示学者兴趣以构建学者画像。Cruz 等[5]提出了利用本体模型的方法来生成学者画像。池雪花[6]采用基于规则的方法从学者相关网页中筛选出学者个人主页，通过触发词和正则表达式等方式制定规则抽取学者性别、个人照片、邮箱、职位和国籍等属性。张秋颖[7]利用 XGBoost 进行学习和预测学者主页，最终通过 BERT-Bi LSTM-CRF 模型对学者主页进行信息抽取，实验取得了较好效果。Lin 等[8]提出了一种基于 Bi-LSTM-CRF 神经网络的配置文件属性提取模型（PAE-NN），该模型通过循环神经网络自动提取相关学术的实体特征，在 Aminer 数据集上较好地提取网络学术用户的多元异构信息。此外，在开放领域学者画像构建中，Scopus、Web of Science、PubMed 和 Google Scholar 是常用的数据源[3]。

在已有研究基础上，本文结合深度学习和自然语言处理技术，提出了一种基于 MTCNN、ResNet 和 XGBoost 模型的学者画像构建方法，以深入挖掘海量数据中的学者多维度标签信息，进而实现精准学者画像构建。

2 相关模型和技术

2.1 XGBoost

XGBoost[9]（eXtreme Gradient Boosting）是经过优化的分布式梯度提升库，依据损失函数在梯度下降方向上组合多个 CART 树，能够自动利用 CPU 多线程进行分布式学习和多核计算，适用于处理大规模数据，在多个领域得到广泛应用。鉴于学者主页识别任务可以转换为传统的分类任务，而 XGBoost 模型在文本分类等应用中具有训练速度快而且精度高的特点[10]，本文采用该模型从搜索结果网页中预测学者主页，以提高主页识别效率。

2.2 多任务卷积神经网络

多任务卷积神经网络[11]（Multi-Task Convolutional Neural Network, MTCNN）是一种基于级联架构的多任务 CNN 网络，由 P-Net（Proposal Network）、R-Net（Refine Network）和 O-Net（Output Network）三层网络结构构成，它可以同时实现图像中人脸检测和人脸关键点定位，其结构如图 1。

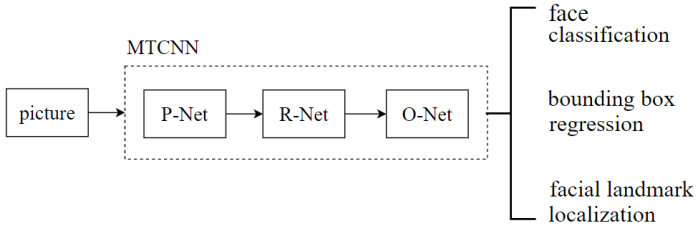

图 1　MTCNN 模型结构

（1）P-Net

P-Net 是一个全卷积神经网络，它首先对图像金字塔输出的图像进行初步特征提取，接着采用 Bounding-Box Regression 调整人脸边框，最后使用非极大值抑制（Non-Maximum Suppression，NMS）技术对边框排除过滤[12]。

在处理时，P-Net 首先将图像特征输入三个卷积层，接着通过人脸分类器判断该区域是否存在人脸，然后使用 Bounding-Box Regression 边框回归确定人脸的大体边框范围，最后通过定位器获取面部关键点的位置信息。P-Net 最终将边框回归得到的多张可能存在人脸区域输入到 R-Net 以进行下一步处理，如图 2 所示。

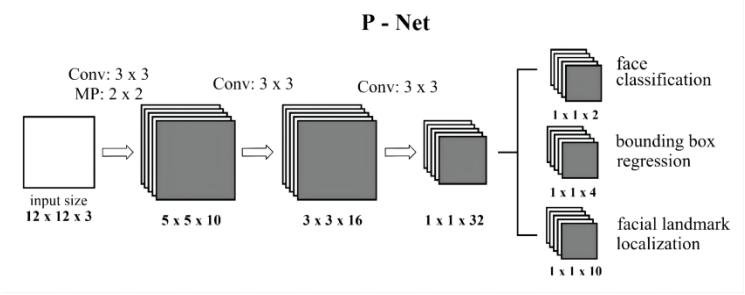

图 2　P-Net

（2）R-Net

相对于 P-Net，R-Net 增加了一个全连接层，能够更严格地筛选数据特征。它首先对 P-Net 输出的多个预测边框进行过滤，接着再次使用 Bounding-Box Regression 和 NMS 进一步优化预测结果。

在处理时，R-Net 首先对 P-Net 输出的人脸区域进行细化选择，接着对人脸区域进行边框回归和关键点定位，最后输出可信的人脸区域。与 P-Net 使用全卷积输出 1x1x32 特征相比，R-Net 使用全连接层保留了更多的图像特征[12]，其结构如图 3。

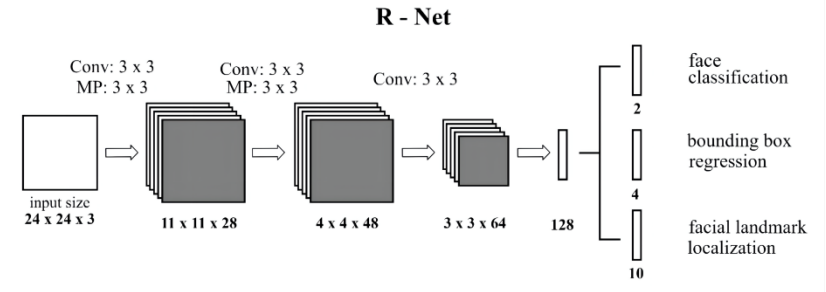

图 3　R-Net

（3）O-Net

与 R-Net 相比，O-Net 增加了一个卷积层，它能够通过更多的数据特征来识别面部区域，并对人的面部特征点进行最后回归，输出精确的人脸面部特征点。

在处理时，O-Net 首先对 R-Net 输出的人脸边框进行判别，接着对正确率最高的人脸边框进行进一步特征定位，最后输出人脸区域左上角和右下角坐标以及人脸区域的五个特征点信息，其结构如图 4。

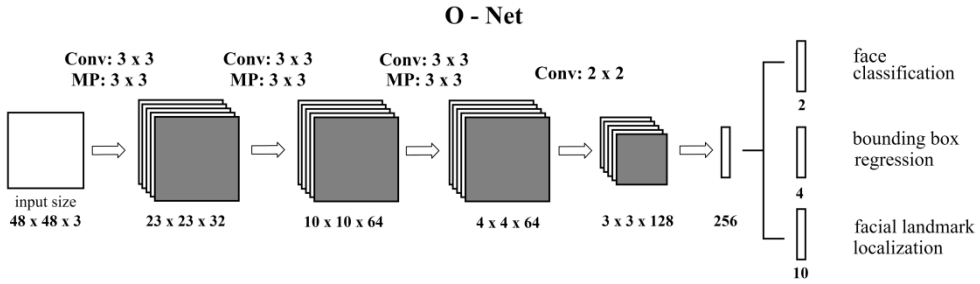

图 4　O-Net

2.3 深度残差网络

深度残差网络（ResNet）于 2015 年由 He[13]提出，在 Imagenet 数据大赛中以压倒性优势取得了冠军。相比卷积神经网络，ResNet 加入了更多网络层数，可以实现更优的网络层次。深度网络一般具有很多冗余层，ResNet 可使这些冗余层完成恒等映射，确保输入和输出完全相同。尽管如此，随着网络深度增加及参数量增加，ResNet 也同样存在训练速度慢[14]等不足。针对该问题，He 等[13]将残差块概念引入 ResNet，把原网络改为多个残差块的叠加，这种设计使得模型在训练时可以根据权重将冗余的网络层设定为恒等层，而不用担心特征数据的损失，从而保证更优网络层次，具体构造如图 5。

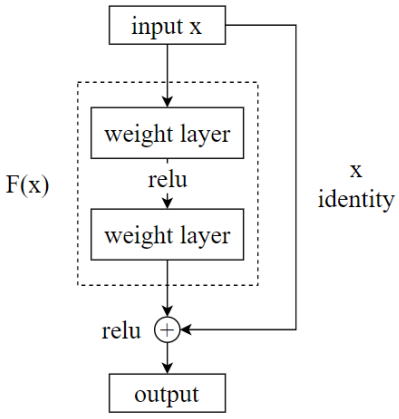

图 5　残差块结构

图 5 中，x 是残差块输入，F(x)是经过第一层线性变换并激活后的输出。值得注意的是，在线性变换之后激活之前，F(x)与该残差块的输入值 x 相加再激活后输出。F(x)输出并激活前加入 x 的路径被称作 shortcut 连接，F(x)在训练的过程中即表示为残差。一个残差块的输

出即为：

$$output = relu(F(x) + x)$$

简单来说，当 shortcut 中间层冗余时，将线性变换层的权重置为 0，则残差输出为 0，从而避免了冗余层造成的梯度消失，完整 ResNet 如图 6。

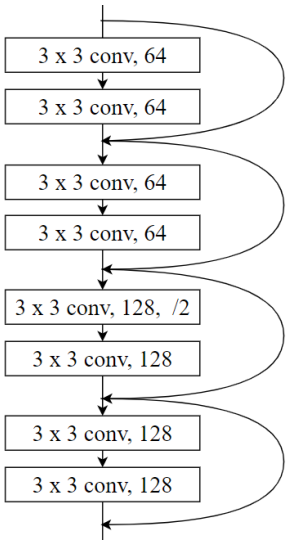

图 6　深度残差网络结构

3 研究设计

3.1 研究框架

根据比赛任务设置，本研究的学者画像通过分析学者信息官网网页结构，抽取学者的详细信息而构建。学者信息抽取分为基本属性（主页、邮箱、职位及头像地址等）抽取和隐式属性（性别和国籍）预测，基本信息可以直接从文本中抽取，隐式属性则是需要预测的属性。具体来说，包含目标学者网页预处理、主页识别与信息抽取、学者肖像识别、学者性别预测和基于学者姓名的国籍预测模块，具体如图 7。

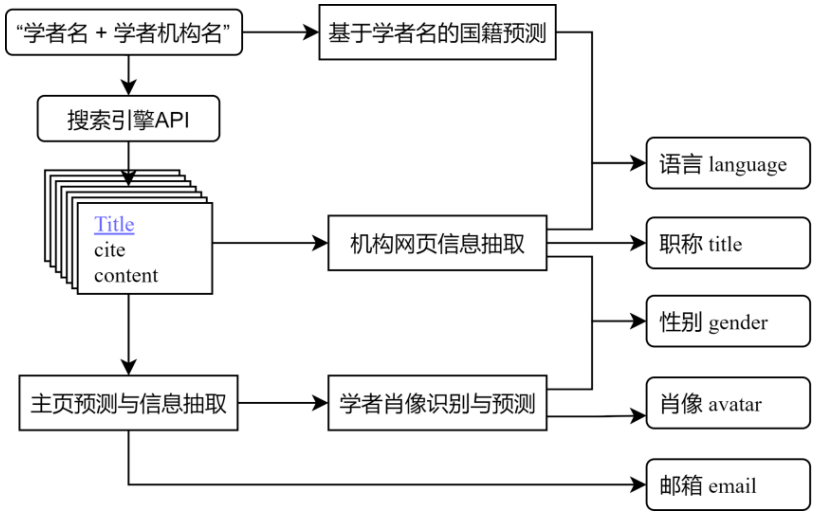

图 7　研究框架

3.2 目标学者网页预处理

学者信息大多包含在个人主页或介绍性网页中，所以准确识别目标学者网页是获取目标学者属性的关键。为了获取较为权威的个人主页，本研究以“学者姓名+学者所在机构名”称为检索词，以 Aminer、Google、Ieeexplore 和 Wikipedia 为数据来源，从搜索结果网页标题、链接和摘要中提取学者姓名、邮箱、职称和主页网址等属性。

3.3 主页识别与信息抽取

在学者主页识别任务中，为了提升学者主页识别的准确率和效果，将学者主页识别转化为网页分类任务；同时根据任务实际需求和分类任务场景，采用 XGBoost 模型预测学者主页。

3.3.1 数据集构建

将学者主页识别视为机器学习分类任务，训练集质量成为主页识别效果好坏的关键。本研究首先对目标学者网页进行预处理，提取搜索结果页面中的特征信息生成样本数据；接着人工标记搜索结果，并分别用 0、1 和 2 来标记。其中，0 表示该页面不是对应的学者主页；1 表示该页面为学者的详细介绍，但不是学者个人主页，如百度百科中的页面；2 表示该页面为学者个人主页或官方页面。具体来说，提取的特征包括：排列次序、学者是否就职于教育机构、积极词在搜索结果中的包含情况、消极词在搜索结果中的包含情况、网址类型、标题长度、简介长度、机构名长度、学者姓名是否出现在标题和简介中。

3.3.2 XGBoost 模型评估

参照已有研究，本实验按照 8:2 划分训练集和测试集。由于正向标签（为主页）在标签集中平均仅占到 10%，所以实验采用召回率评估模型预测性能。在人工标注数据集上 XGBoost 召回率为 80%，表明该模型能够从搜索结果网页中有效地提取特征，进而可以较好地识别目标学者主页。

3.4 学者肖像识别

学者肖像识别是指对学者人脸进行检测，进而提取单张人脸的图片作为学者肖像。目前常用的人脸检测模型主要有 CCF[15]、MTCNN[16]、FaceBoxes[17]和 SRN[18]等。其中 MTCNN 可以兼顾人脸检测与人脸对齐任务，网络结构轻量精简、速度快且召回率高而得到广泛应用。本研究首先从预测的学者主页中提取图片，然后采用基于特征融合的 MTCNN 对图片进行人脸检测，提取出单张人脸的图片作为学者肖像。

为了提升人脸检测准确率和训练速度，在 MTCNN 人脸检测中主要使用到了图像金字塔、边框回归和非最大值抑制等技术。具体来说，首先构建图像金字塔，对输入图像进行尺度变换，以适应不同大小的模型输入要求，将图片转换为输入卷积神经网络的固定大小；其次将金字塔化处理后的图像输入 P-Net 生成人脸的候选边框，并利用非极大化抑制算法 NMS

（Non-Maximum Suppression，NMS）校准边框去除多余边框；然后将 P-Net 得到的候选框和原图输入 R-Net，并对 P-Net 输出的图像进行进一步筛选，以过滤重复且不符合要求的候选框，再利用 NMS 做候选框合并处理；最后将人脸的双眼、鼻子和嘴部两端五个关键特征点作为检测重点，最终得到人脸候选边框以及相应特征点位置。

基于特征融合的 MTCNN 模型的人脸检测效果如图 8。实验结果表明 MTCNN 预测性能良好，检测精度高。

图 8 人脸抓取

3.5 学者性别预测

在学者肖像识别基础上，采用 ResNet 对 MTCNN 识别的学者肖像进行性别预测。首先为了提高模型精度，从 Kaggle 上下载“肖像图片与性别数据集”[1]进行训练；接着采用 MTCNN 识别学者肖像，通过测试得到准确率为 97%；同时经过统计发现，女性学者仅占比 10%，因此将女性看作是正向标签，最终得到性别属性准确率 81%，表明即使图片数据存在大量噪声，ResNet 依然可取得较好预测效果。

3.6 基于学者姓名的国籍预测模块

在学者国籍预测模块，为了提升模型效果，调用 Flair 框架中 TextClassifier 分类器，从 Kaggle 上下载“姓名与国籍数据集”[2]进行训练，对来自 40 多个国家的姓名进行了国籍分类，得到测试集的准确率为 78%，由此可知姓名与国籍具有较强关系。本研究中，通过姓名预测国籍模型取得了 75%的准确率，表明通过学者姓名预测国籍可以取得较好效果。

4 总结

为了精准抽取开放互联网中的学者信息，本研究利用深度学习和自然语言处理技术，首先对目标网页进行预处理以抽取学者姓名、邮箱、职称和主页网址等属性；其次使用 XGBoost 模型识别学者主页；接着利用 MTCNN 对预测主页中的图片进行人脸识别；最后使用 Flair 框架通过姓名预测学者国籍，从而精准构建多维度学者画像，通过实验评测取得了较好效果。

---

[1] https://www.kaggle.com/maciejgronczynski/biggest-genderface-recognition-dataset.

[2] https://www.kaggle.com/bryanpark/nana-dataset.

本研究的主要工作体现在以下几个方面：

（1）在主页预测与信息抽取任务中，使用 XGBoost 模型预测学者主页。实验在人工标注数据集上取得了召回率为 80%的结果。

（2）在学者肖像识别与性别预测任务中，本研究首先从预测出的学者主页中抽取图片，接着利用基于特征融合的 MTCNN 对图片进行人脸识别，进而预测学者肖像图片。

（3）在学者性别预测任务中，本研究使用 ResNet 对 MTCNN 识别出的学者肖像进行性别预测，准确率达到 97%，验证了基于学者肖像预测其性别的可行性和有效性。

（4）在学者姓名的国籍预测任务中，采用 Flair 框架预测国籍准确率达到 78%，表明姓名与国籍具有较强的相关关系。

5 参考文献

[1] Silva T, Ma J, Yang C, et al. A profile-boosted research analytics framework to recommend journals for manuscripts[J]. Journal of the Association for Information Science and Technology, 2015, 66(1): 180-200.

[2] 袁莎,唐杰,顾晓韬.开放互联网中的学者画像技术综述[J].计算机研究与发展,2018,55(9):1903-1919.

[3] Gasparyan A Y, Nurmashev B, Yessirkepov M, et al. Researcher and author profiles: opportunities, advantages, and limitations[J]. Journal of Korean medical science, 2017, 32(11): 1749-1756.

[4] Tang J, Zhang J, Yao L, et al. Arnetminer: extraction and mining of academic social networks[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008: 990-998.

[5] Cruz I, Bravo M, Reyes-Ortíz J A. Ontology-based Population and Enrichment of Researcher Profiles[J]. Computers & Operations Research, 2019, 148(3): 181-194.

[6] 池雪花. 学者精准画像的自动构建研究[D].南京：南京理工大学,2019.

[7] 张秋颖. 学者主页的判别与信息抽取[D].上海：上海交通大学,2020.

[8] Lin W, Xu H, Li J, et al. Deep-profiling: a deep neural network model for scholarly Web user profiling[J]. Cluster Computing, 2021: 1-14.

[9] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[10] SeyfioLu M ， Demirezen M . A Hierarchical Approach for Sentiment Analysis and Categorization of Turkish Written Customer Relationship Management Data[C]// Computer Science & Information Systems. IEEE, 2017:361-365.

[11] Zhang K ， Zhang Z ， Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016,

23(10):1499-1503.

[12] 周航,蔡茂国,吴涛,等.一种改进的多任务级联网络人脸检测算法研究[J].智能计算机与应用,2021,11(3):172-176.

[13] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[14] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.

[15] Yang B, Yan J, Lei Z, et al. Convolutional channel features[C]//Proceedings of the IEEE international conference on computer vision. 2015: 82-90.

[16] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[17] Zhang S, Zhu X, Lei Z, et al. Faceboxes: A CPU real-time face detector with high accuracy[C]//2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017: 1-9.

[18] Zhang S, Chi C, Lei Z, et al. Refineface: Refinement neural network for high performance face detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. arXiv: 1909.04376.