DOI: 10.54254/3050-2160/2025.21126

基于数据挖掘的互联网行业热门职位分析

卜 淳

（北京邮电大学，经济管理学院，北京，100876；buchun0926@163.com）

摘 要：本文以 51job 网站为例，爬取“大数据”、“5G”、“AI”、“云计算”等关键词的互联网行业招聘信息，通过数据挖掘与可视化技术展开分析。首先，基于 C4.5 算法构建决策树模型，并通过 REP 方法剪枝，分析互联网热门职位的薪资影响因素；其次，利用主成分分析（PCA）对热门地区和技术关键词进行降维，得出地区和新兴技术的综合热门指标排行；最后，基于 Flask 框架调用 E Chatrs 实现数据可视化，为求职者提供全面、直观的招聘信息。研究结果不仅帮助求职者了解行业动态，还为大学生、企业及院校提供了针对性建议，缓解就业压力，优化招聘与人才培养策略。

关键词：数据挖掘；决策树 C4.5 算法；主成分分析；数据可视化

1 引言

1.1 研究背景

在当今时代，企业招聘的方式发生了变化，在过去往往是公司去高校、招聘市场等寻求人才，求职者们通过线下提交简历、面试来获取工作资格。而如今，网络招聘的崛起，如 51job 招聘网站、智联招聘网站、BOSS 直聘等，已经成为公司招聘和求职者寻求工作的主要平台。随着互联网技术飞速发展，越来越多的人加入网络招聘信息平台求职，表明网络招聘信息一定程度上代表着实际工作招聘信息。

新一代信息技术快速发展，应用于各行各业和人们的日常生活。在日常生活方面，如智能家居、疫情行程码、人脸识别等，都与这些新一代信息技术息息相关。与此同时，全球范围内呈现互联网技术与传统行业的融合，如传统农业、制造业和服务业与人工智能、物联网等信息技术融合。对于公司而言，相关人才需求日益增大，因此对互联网行业热门职位，尤其与新一代信息技术相关的招聘信息挖掘分析具有实际意义。近几年来，各类高校不断增加招生名额，2024 界高校毕业生规模预计达到 1179 万人，相比去年增加 21 万人。大量毕业生涌入社会，但企业需要的是技术性人才，由于大多数毕业生自我认知不够精准，技术实践能力普遍较低，使就业压力增加。毕业生的增加随着岗位的需求，网络招聘成为缓解就业问题的重要手段。在大数据时代，充分挖掘网络招聘信息，提炼有价值、有意义的结论，可以帮助毕业生精准定位，符合对职位的个性化需求。进而建立可视化网站，提供充分、全面、直观的职位信息。这些都将一定程度上帮助毕业生降低就业压力，为大学生未雨绸缪。最后，对于很多大学生来说，未来就业前景是迷茫的。在互联网行业与新一代信息技术相关的职位中，哪些新兴技术、地区更热门？哪些新兴技术、地区人才需求更大？学历、工作经验、地区哪个因素对薪资影响最大？不同的条件下薪资状况如何？不同的新兴技术、地区、公司类型、学历、经验下平均薪资是多少？等等一系列问题。这些都是当代大学生的困扰。

1.2 研究意义

理论意义：根据互联网行业热门职位，以新一代信息技术相关职位为例的网络招聘信息，保存到 SQLite 数据库，编写 python 代码，进行数据挖掘和数据可视化。第一，本文提出基于决策树 C4.5 算法和 REP 方法剪枝，构建薪资等级判断系统的决策树模型，结合如今适用性较高的 python 环境，完成一系列编码，该套编码理论上适用于任何行业的职位信息。第二，使用主成分分析，根据计算和 85%提取原则，得到几个主成分，并加权计算出综合评分，对热门职位、地区进行排行。该方法结合了数据库数据调用和 SPSS 平台，来实现主成

分分析（降维），丰富了根据不同需求使用多种软件结合解决问题的方法。第三，在 python 环境中，基于 Flask 框架和 E charts 图表的调用，实现网页的可视化，该网页功能多样，如搜索框，地区分布动态图、各类型平均薪资排行、热门排行等，通过调用 Python 中较为灵活的 Flask 框架，满足对职位信息的需求。

实际意义:第一，根据构建的决策树模型，并将其在网站中可视化，帮助毕业生或在校大学生，了解不同学历、不同经验、不同地区的多种要求下的薪资水平，进一步帮助学生确定目标。第二，由于网站搜索框可以直接访问招聘职位的具体详情网站链接，对于求职者，可以从理论了解直接转换为实际行动，进入详情网站，了解更加细致的职位内容和投递简历，一定程度上缓解毕业生求职者的压力。第三，通过多角度将招聘信息中的浅层知识和深层知识清晰展现，包括各类柱状图、饼状图、动态图等，网站访问者可以了解到当前社会对新一代信息技术相关人才的具体需求。互联网企业可以了解到新一代信息技术相关岗位的需求现状和人才的招聘形势。另外，该网站也为院校提供了互联网行业热门职位的各类要求的具体参考，帮助院校预测未来形式较好的相关职位和专业技能，制定有针对性的培养方案和课程，使毕业生同时满足院校和公司的双重需求，进而缓解就业压力大的形式。

2 研究思路与方法

2.1 研究思路

论文主要研究步骤如图 1 所示，根据研究步骤可以将本文研究内容概括为：确定目标、数据获取与保存、数据挖掘、数据可视化四个部分。

确定目标：在上文提到的研究背景下，结合毕业生、求职者、在校生、企业、院校等实际需求，本文首先确定研究目标是为相关人员提供直观、清晰的浅层知识和通过从网络招聘网站的大数据中挖掘有意义的深层知识的网络招聘信息可视化网站。并丰富网站功能，实现便利，简单的网站操作功能。对于网站的实现方式，本文结合 python 的热门性，选择 python 语言作为全程编程语言，同时可以根据需求使用一系列辅助软件。

数据获取：由于本文研究的是互联网行业热门职位分析，所以选取如今互联网行业比较热门的“物联网、大数据、人工智能、5G、云计算、区块链”新一代信息技术为关键词，获取 51job 招聘网站的相关招聘信息。通过观察总结网站网址规律，编写爬虫代码，将获取的数据进行预处理，最终保存到 SQLite 数据库中。

数据挖掘：本文选择分类决策树模型和主成分分析来进行数据挖掘。首先，将爬取的数据进行清洗，并分为两部分，将 C4.5 算法编写成 python 语言，最后将决策树可视化。并再通过 REP 方法，自下而上的对决策树进行剪枝，降低错误率，最终实现薪资评判的决策树模型。其次，对数据进行热门评分转化，在通过 SPSS 平台，找到热门排行的主成分，最后将主成分进行加权计算综合评分，在根据综合评分得出热门地区和热门新兴技术的排行。

数据可视化：基于 Flask 框架，实现网站的搭建，最终网站实现的功能主要包括：一，根据条件搜索数据库的招聘信息，并可以通过点击职位名称直接访问实际工作详情链接；二，根据不同省份的招聘信息数量，调用 E charts 中的地图动态图，根据颜色的深浅来反映各省份招聘信息数量的多少；三，薪资等级评判系统，根据输入的学历、工作经验、地区给出薪资等级（薪资具体范围）；四，词云图，利用 python 的 wordcloud 库，生成词云图，反映最常见的福利待遇是什么，最常见的地区是哪，最常见的工作名称是什么；五，不同类别下的平均薪资排行柱状图，直观反映平均薪资；

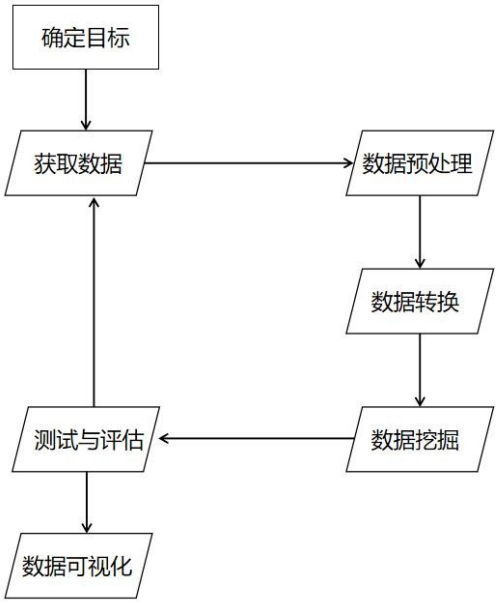

图 1 论文主要研究步骤

本文的整体框架结构如图 2 所示。首先从研究背景出发，提出当今社会中存在的问题，解释了本文的思路来源。其次开始介绍本文拟采用的方法和模型，并给出选用这些方法与模型的原因。再次是本文核心两部分：数据挖掘与数据可视化，通过爬取网络招聘信息，采用数据挖掘方法，得出结论，同时基于 python 代码，创建网站，实现数据可视化。最后，将数据挖掘所得到的结论和数据可视化实现的功能性网站进行总结，并根据结论和网站分别针对学校、企业、大学生、社会求职者给出相关建议。对于论文编写时所发现的问题，提出问题，并给出打算，在未来不断完善论文。

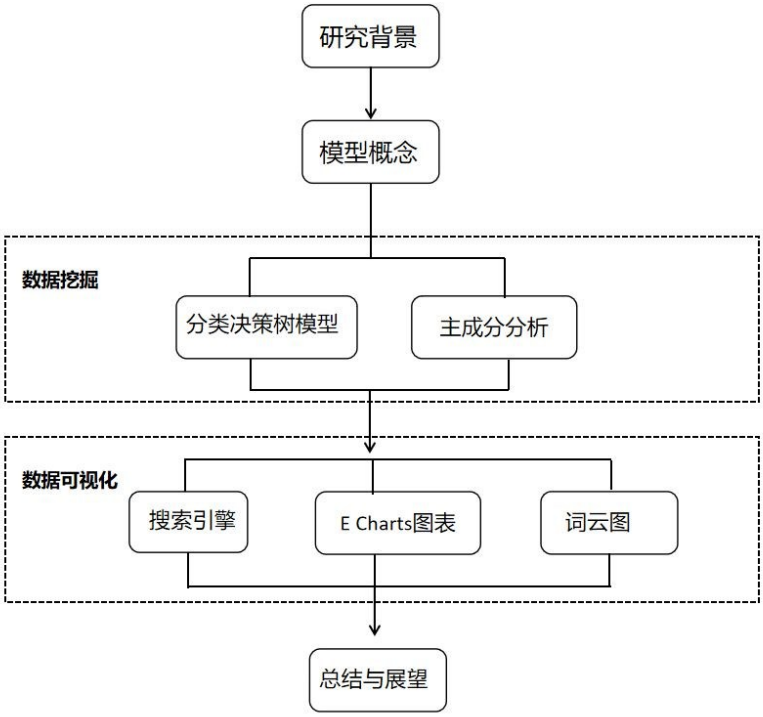

图 2 论文框架结构图

2.2 研究方法

(1)文献研究法：主要指根据已确定的目标，在知网、万方数据库等学术平台，按自身需求搜索相关文章，并将其中的方法和结论进行归纳、总结和学习，为论文的编写做好先前工作。本文通过对于网络招聘信息各类爬取的相关文献进行归纳和学习，编写符合本文需求的 python 代码，获取对应数据并保存。另外，通过阅读数据挖掘中的各种技术在各类领域应用的相关文献，进行总结和比较，根据本文研究的内容和数据，选择适合于本文的数据挖掘方法。最后，阅读各类数据可视化的相关文献，进行功能分析总结，根据本文网站拟实现的功能，来选择本文数据可视化网站的搭建框架和开发平台。

(2)统计分析法：主要指根据原有的数据和研究目的，选择符合需求的统计方法，进行分析。如利用现成的 SPSS 软件或者 python 自编码，进行相关的数学计算，最终得到具有实际意义的结论。

(3)决策树分析法：即利用决策树模型，对问题进行更直观的分析，并得到相关结论。适用于某指标受多种因素影响的分析。主要内容包括根据决策树模型，来比较各因素影响程度的大小，以及在某个影响程度较大的属性的不同分类下，由于其他影响程度较小的属性的差异，导致最终结果不同的分析。

(4)可视化分析法：主要指为了将某些以文字或数字形式的结论更加直观，通过技术实现可视化，以网站、图表、搜索框、动态图等形式将结论清晰的展示出来。主要适用于研究结论形式多样、数量较多的论文。

3 相关概念与理论模型介绍

3.1 网络爬虫技术

截至 2024 年，中国网民规模已超过 11 亿人，网民在各种网页中产生了不计其数的数据。在数据爆炸的时代中，越来越多的人希望在巨大的数据中，用某些手段来获取有价值的结论。在该过程中，获取网页数据成为关键，因此，人们的需求促使了网络爬虫技术的出现。网络爬虫主要指根据人们打算获取的目标数据，来有针对性的编写代码程序，并将这些数据保存下来[1]。在大数据时代，由于部分信息的不公开性、不透明性和商业机密性，根据用户不同的需求定向爬取一些网页信息并进行学术研究已成为主流。网络爬虫框架主要包括：一，获取网页源代码：模拟用户，通过 HTTP 库向目标站点发出请求，将得到的 Response 以 HTML 形式保存。二，网页内容解析：根据网页内容和需求，进行对应正则表达式的编写，将获取的内容以二维列表的方式保存。三，利用 python 编码将内容保存到 SQLite 数据库中。

3.2 SQLite 数据库

目前存储数据主要包括以下四种方式：TXT：超过一万行的数据不好操作；Excel：超过 10 万行速度明显变慢；SQLite：100 万行操作起来方便；MYSQL：亿级以下操作起来非常方便。对于学术研究而言，SQLite 数据库已经完全符合存储数据的需求。

SQLite 数据库可以通过 python 直接调用，且本文数据量远远小于 SQLite 所能处理的数据量，所以选择将爬取的数据保存到 SQLite 数据库中[2]。

3.3 数据挖掘概述

数据挖掘是在数据中寻找有趣结构的过程，它往往用于数据量较多的领域，往往数据量越多，数据挖掘的结论越有说服力和实际意义。如今，往往根据人们的需求，来选择对应的数据挖掘技术，通过 python、r 语言等各类编程语言来实现。数据挖掘主要包括无监督的学习和有监督的学习。有监督学习主要指从已经有预定义的类的数据中构建模型，并对其他数据使用该模型，得到结果，如分类、预测等，无监督学习指的是从没有预定义类的数据中构建模型，如主成分分析、货篮分析等[3]。

文本挖掘也是数据挖掘的一种形式，即对大量无规则的文本数据进行分析，由于各类网站和平台上的文本信息越来越多样丰富，如今文本挖掘已经成为 KDD 领域的热门方向。本文选用的是通过 python 编码，基于词云图的形式，对文本数据的词频大小进行展示，常用的方法有 TF-IDF 算法、LDA 模型和自然语言处理法等[4]。决策树是数据挖掘中分类常见的一种方法。根据训练集生成规则，并按规则产生根节点和叶节点，将数据进行分类。决策树 C4.5 算法是将 ID3 算法进行了改进，把根节点和叶节点的分支判断条件从信息增益改为信

息增益率。另外，在 C4.5 算法构建决策树的过程中可以进行剪枝，避免由于训练集的某些数据导致过拟合现象。

C4.5 算法的步骤：

计算类别信息熵：

$$Info(I) = -\sum_{i=1}^{n} \frac{\#in\ class\ i}{\#in\ I} log\left(\frac{\#in\ class\ i}{\#in\ I}\right)$$

计算每个属性的信息熵：

$$Info(I,A) = \sum_{j=1}^{k} \frac{\#\ in\ class\ j}{\#\ in\ I} info(class\ j)$$

计算信息增益：

$$Gain(A) = Info(A) - info(I,A)$$

计算属性分裂信息度量：

$$Split\ Info(A) = -\sum_{j=1}^{k} \frac{\#in\ class\ j}{\#in\ I} log\left(\frac{\#in\ class\ j}{\#in\ I}\right)$$

计算信息增益率：

$$GainRatio(A) = Gain(A)/Split\ Info(A)$$

选择信息增益率最大的为根节点，在子结点当中重复上述过程，直至决策树构建完成[5]。

为了解决决策树构建中可能存在问题，并使模型更具有说服力，对已经构建的决策树模型进行后剪枝。后剪枝即将基于训练集已经构建好的决策模型，根据计算，将某些分支删除，从而达到简化的目的。并且分支转化为节点的类别属性是根据原有分支中各类别属性的多少而决定的，往往选择最多的为该节点类别属性。本文采用的是错误率降低剪枝法，即 REP 方法。具体算法如下：

首先，计算最下层非叶节点的误判样本数：

$$E_r(t) = N(t) - N_c(t)$$

其中，$N(t)$为结点t的样本数，$N_c(t)$为结点t中被正确分类的样本数。

其次，计算该节点子树的误判样本数：

$$E_r(T_t) = \sum_{i \in T_t} E_r(i)$$

最后，判断$E_r(t)$与$E_r(T_t)$大小，若$E_r(t)$<$E_r(T_t)$，则将该子树替换成叶子节点，否则，不变。

重复上述步骤，直到没有任何子树可以进行替换，即完成剪枝过程[6]。

由于本文研究内容其中包括对互联网行业热门职位的薪资影响因素分析，且决策树模型对于影响因素的比较更为直观，所以选择数据挖掘中的分类决策树模型。另外，在决策树常见算法中，C4.5 算法较为常用，与 ID3 算法相比，可以进行剪枝和对连续值变量进行处理，以信息增益率作为节点判断标准，使模型更为准确。在剪枝方法中，错误率降低剪枝法 REP，相比较而言计算量较小，且本文数据量较大，符合 REP 方法适用范围。综上所述，本文选择数据挖掘中的分类决策树模型，采用 C4.5 算法构建树和 REP 方法剪枝。

主成分分析主要指在实际问题分析的过程中，对于决定最终结论的因变量，往往会有很多指标同时对该变量造成影响。如果在分析的过程中同时考虑所有指标，则分析将会十分复杂。为了使分析更加简单和清晰，人们往往会在分析的过程中，使用 PCA，在尽可能不损失原有信息的基础上，从这些指标中提取出几个综合指标，同时要保证原有信息量要保持 85%以上才不会对结论造成影响。主成分分析的步骤如下：

假设有 a 个数据，每个数据有 b 个指标，最后保留 p 个综合指标。

首先，通过标准化处理实现数据无量纲化：

$$x_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}, i = 1,2,\cdots,a, j = 1,2,\cdots,b$$

构建出一个 a × b 的矩阵，接着对数据矩阵进行线性变化，产生新的综合变量，记为P。则原变量指标为$X_1, X_2, \cdots, X_b$，新变量指标为$P_1, P_2, \cdots, P_b$，具体变换公式如下：

$$\begin{bmatrix} x_{11} & \cdots & x_{1b} \\ \vdots & \cdots & \vdots \\ x_{a1} & \cdots & x_{ab} \end{bmatrix}$$

$$\begin{cases} Z_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1b}X_b \\ Z_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2b}X_b \\ \vdots \\ Z_b = u_{b1}X_1 + u_{b2}X_2 + \cdots + u_{bb}X_b \end{cases}$$

其中$Z_1, Z_2, \cdots, Z_b$分别为原变量指标在经过线性变化后，始终保持方差最大的新综合指标，并且这些新综合指标是没有关联性的。称综合变量$Z_1, Z_2, \cdots, Z_b$为主成分，为了尽可能保留原有信息，我们往往提取主成分时，要求累计方差尽可能达到 85%[7]。

由于本文研究内容包括分析哪些地区和新兴技术更为热门，根据数据发现影响热门排行的指标较多，包括薪资大小、招聘数量、公司类型、学历、经验、公司规模六个指标，无法进行比较。所以本文选择 PCA，在尽可能不损失原本信息的情况下，找到几个主成分，进行加权计算，最终实现热门地区和新兴技术排行。

3.4 数据可视化

Flask 网页框架是 python 自带的框架，基于 Flask 的网页实现，可以全部在 Python 的环境下进行。主要流程是先获取前端网页用户所输入的内容，提取到后端，调用数据库，并根据自身功能的需求，在后端进行函数的编写，并将结果输入到前端显示，最后将结果页面反馈给用户。这就是 Flask 网页框架实现的过程[8]。

本文拟搭建的网站所能实现的功能较为基础，且需要连接 SQLite 数据库，所以选用更为灵活且 python 自带的 Flask 框架更为适用。E Charts 具有以下特点：

(1) 具有很强兼容性。可以被谷歌、火狐等大多数浏览器兼容。
(2) 免费。几乎所有数据可视化图表可以免费使用。
(3) 分类多样。除了基本的图形如折线图、柱状图、饼图，还有复杂的雷达图、桑葚图、路径图等。同时也包含静态图和动态图，能够满足绝大多数的图表需求。
(4) 引用配置简单。通过下载并引用封装的 js 文件，对数据和各指标进行配置，根据自身的需求来获取所需的图形的形式[9]。

E Charts 网站中具有较多的图形，且由于本文可视化网站是基于 Flask 框架，可以直接在 HTML 文件中直接引用 E Charts 图表。另外，网站中地图动态图符合本文展示地区招聘信息数量差异的目的，所以本文选择 E Charts 图表可视化。

4 网络招聘信息数据挖掘

4.1 网络招聘相关数据获取

为了获取互联网行业热门职位信息，选取互联网热门新兴技术为关键词，包括：“大数据”、“AI”、“区块链”、“物联网”、“云计算”、“5G”。确定关键词后，对 51job 招聘网站的各个关键词和不同页数网页地址进行总结，得到规律为：“https://search.51job.com/list/000000,000000,0000,00,9,99,(关键词),2,(分页数).html”。基于该规律，编写 pyhon 代码，利用循环语句、urllib 库等，模拟浏览器打开网页，将得到的 Response 即网页源代码，以 HTML 的形式保存，共计 4000 条左右。确定所需数据内容主要包括：工作名称、工作地点、薪资水平、经验要求、公司名称、福利待遇等，观察网页源代码，分别进行不同内容的正则表达式编写。以工作名称为例，在 html 文件中，对于每个招聘信息，工作名称内容以“"job_name":"大数据开发工程师"”形式存在，所以构造正则表达式“findjob_name = re.compile(r'"job_name":"(.*?)"')”，来批量获取对应信息。其他内容以此类推。并以二维列表的方式保存[10]。

4.2 数据预处理

首先进行数据清洗，在爬取数据后，对于薪资属性，以“万/月”、“千/月”、“万/年”形式存在，为了方便后续计算，并且“万/年”的数量占比较少，对于这部分数据进行清洗。最终数据只保留“万/月”、“千/月”两种形式。对于学历要求属性，小部分数据为空，将这些数据进行清洗。数据清洗后进行数据转换，对于薪资属性，以“1-2 万/月”“5-6 千/月”“3 万/月”等形式存在，将其进行数据转换，利用 replace 去掉文字。

对于“1-2 万/月”类型，利用 split 函数，以“-”进行分割，取范围最大值与最小算出平均值。对于“5-6 千/月”，同上，算出平均值，再除以 10。最终数据形式以数值型保存。通过 python 编码，将列表中存放的预处理后的数据保存到 SQLite 数据库中，方便进行数据挖掘、数据可视化等。数据库内容主要包括：数据序号、职位名称、职位详情链接、学历要求、公司类型、公司规模、关键词、工作地点、薪资水平、经验要求、公司名称、福利待遇等，

4.3 基于 C4.5 算法的薪资影响因素分析

首先，结合实际情况，我们选取工作地点、学历要求、经验要求作为薪资影响因素。将数据库的“薪资”、“工作地点”、“经验要求”、“学历要求”进行调用和预处理。

对于“薪资”，由于我们获取的职位信息地区包含全国各地，所以不考虑一线城市与偏远城市之间消费差距，进而导致收入差距。本文为了降低决策树的复杂度，仅划分 A、B、C 三个等级，根据实际生活经验，认为年薪大于 20 万/年为高收入，年薪在 10 万/年到 20 万/年为中收入，年薪低于 10 万/年为低收入。故将大于 1.6 万/月转换为“A”，将大于 0.8 万/月且小于等于 1.6 万/月转换为“B”，将小于等于 0.8 万/月转换为“C”，即将薪水水平评为“A”、“B”、“C”三类。对于“经验要求”，将大于等于 3 年经验要求的转换为经验要求高即“yes”，将小于 3 年经验要求的转换为经验要求低“no”，共两类。对于“工作地点”，我们将广东省、江苏省、浙江省、上海转换为“沿海城市”，将北京各区转换为“北京市”，将剩余地区转换为“其他城市”，三种类型；对于“学历要求”，根据占比分为“本科”、“硕士”、“大专”、“其他”四类；基于训练集构建决策树模型，通过选取“薪资”为类标签属性，以“学历要求”、“经验要求”、“地区分布”作为节点，进行决策树建模。根据 C4.5 算法编写 python 代码，得到决策树结果，并将其可视化。代码主要包括以下六部分函数：

第一，创建数据集。调用数据库 C45 表中数据，并以列表形式保存数据集。

第二，计算数据原始熵。

第三，将数据按某个类别的不同属性进行分割。

第四，选择最优的分类特征。计算按不同类别分类后的熵，得出信息增益与信息增益率，比较信息增益率的大小，选取最大的类别进行分类。

第五，建树。最终输出结果以字典形式存在。

第六，决策树可视化。

最终得到基于 C4.5 算法的决策树模型，如图 3 所示。

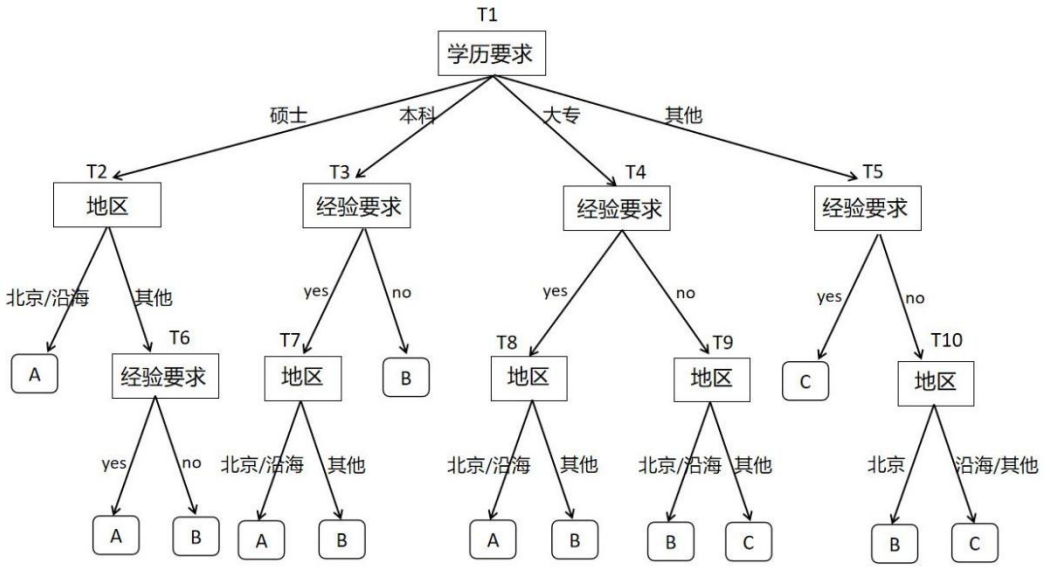

图 3 剪枝前的决策树模型

但基于训练集所生成的决策树模型容易出现过拟合问题，我们选取一个月后的 51job 网络招聘信息中相应关键词数据约 1000 条，作为测试集，来判断决策树模型的准确度，并进行后剪枝。我们采用 REP 方法，自上而下的判断非叶子节点是否剪枝。

首先，根据大多数原则，即选该节点中占比最大的类别为结果，计算出原决策树模型最下方的非叶子节点的误判样本数和错误率。其次，计算出对应子树的误判样本数和错误率。最后，判断剪枝后错误率是否降低，来决定是否剪枝[11]。

对于节点 T6，剪枝前误判样本数为 8 个，错误率为 53.33%，剪枝后误判样本数为 6 个，错误率为 40%，40%<53.33%，故进行剪枝。对于节点 T7，剪枝前误判样本数为 110 个，错误率为 27.99%，剪枝后误判样本数为 91 个，错误率为 23.16%，23.16%<27.99%，故进行剪枝。对于节点 T8，剪枝前误判样本数为 23 个，错误率为 37.70%，剪枝后误判样本数为 24 个，错误率为 39.34%，39.34%>37.70%，故不剪枝。对于节点 T9，剪枝前误判样本数为 48 个，错误率为 40.34%，剪枝后误判样本数为 64 个，错误率为 53.78%，53.78%>40.34%，故不剪枝。对于节点 T10，剪枝前误判样本数为 10 个，错误率为 32.26%，剪枝后误判样本数为 10 个，错误率为 32.26%，故为了简化决策树，进行剪枝。

得到剪枝后的决策树模型，如图 4 所示。

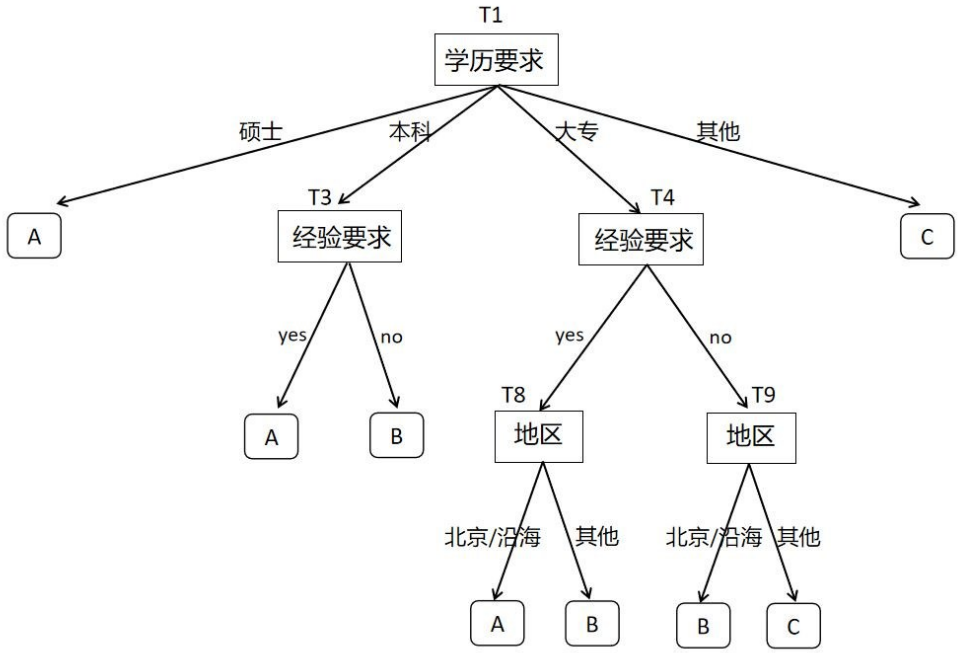

图 4 剪枝后的决策树模型

对于求职者而言，学历、经验要求、地区等对薪资水平都有一定程度的影响。结合常识与数据表格证明，我们知道学历越高薪资水平越高、经验要求越高薪资水平越高、一线城市薪资水平更高等。但哪个因素对薪资水平影响最大？对于不同条件下的职位薪资水平是多少？这部分缺少实验证明。

本文利用爬取的数据，进行了实验证明。使用决策树 C4.5 算法生成决策树和 REP 方法进行剪枝，生成最终的决策树模型，如图 4。

根据决策树模型，我们发现在互联网行业热门职位中，对于薪资水平影响最大的因素是学历，对于硕士而言，绝大多数薪资等级为 A 即大于 20 万/年。对于其他（初中、高中、中专等），薪资等级为 C 即小于 10 万/年。

其次影响因素为经验要求，对于本科生而言，当经验要求大于 3 年时，薪资水平大于 20 万/年，当经验要求小于等于 3 年时，薪资水平在 10 万/年-20 万/年之间。

最后影响因素为地区，对于大专学历的学生，当经验要求大于 3 年时，如果工作地点在北京和沿海城市时，薪资大于 20 万/年，如果工作地点在其他地区时，薪资水平在 10 万/年-20 万/年之间；当经验要求小于等于 3 年时，如果在北京和沿海城市时，薪资水平在 10 万/年-20 万/年之间，如果在其他地区，薪资小于 10 万/年。

4.4 基于主成分分析的人才需求分析

首先对数据进行预处理，便于进行热门地区和职位排行。“薪资”、“职位关键词”保留原始数据库类型；“职位地点”中城市全部转换为对应省份，如“南京-鼓楼区”转换为“江苏”；“公司类型”中民营、上市公司、国企共占 80%，表明这三种类型较为热门，将其根据占比进行排序评分，如表 1；“经验要求”中“3-4 年”、“2 年”、“5-7 年”占比 70%，表明这三类较为热门，根据占比进行排序评分，如表 2；“公司规模”中“150-500 人”、“50-150 人”、“1000-5000”对人才需求较大，另外，在该类属性中，有 182 个空白数据，计算其余数据的得分平均值为 5 分，将其对空白数据进行填充，如表 3；“学历要求”中“本科”、“大专”、“硕士”总占比 94%，表明这三类学历较为热门，根据占比进行排行评分，如表 4。其余属性不考虑。将其保存到 SQLite 数据库 PCA 表中，为后续数据分析做准备。

表 1 “公司类型”评分统计

公司类型	数量	占比	得分
民营	2097	59.91%	6
上市公司	394	11.26%	5
国企	386	11.03%	4
外资	326	9.31%	3
合资	183	5.23%	2
其他	114	3.26%	1
	3500	1	

表 2 “经验要求”评分统计

经验要求	数量	占比	得分
3-4 年	1375	39.29%	6
2 年	681	19.46%	5
5-7 年	584	16.69%	4
1 年	492	14.06%	3
无需经验	275	7.86%	2
8-9 年	64	1.83%	1
10 年以上	29	0.83%	1
	3500	1	

表 3 “公司规模”评分统计

公司规模	数量	占比	得分
150--500	802	24.17%	7
50--150	792	23.87%	6
1000-5000	514	15.49%	5
500-1000	456	13.74%	4
小于 50	435	13.11%	3

表 3 (续表)

大于 10000	207	6.24%	2
5000-10000	112	3.38%	1
空白	182	平均值	5
	3500	1	

表 4 “学历要求”评分统计

学历要求	数量	占比	得分
本科	2341	66.89%	5
大专	722	20.63%	4
硕士	311	8.89%	3
中技/中专	47	1.34%	2
高中	46	1.31%	2
博士	25	0.71%	1
初中及以下	8	0.23%	1
	3500	1	

第一，对热门地处综合指标进行排行，在 PCA 表中主要包括：北京、上海、浙江、江苏、广州、沈阳等 17 个省份，共计 3333 条，其余省份共计 167 条，占比不足 5%，不进行分析。编写 python 代码，调用 SQLite 数据库，分别计算 17 个省份的招聘信息数量、平均薪资、平均公司类型得分、平均公司规模得分、平均经验要求得分、平均学历要求得分，并将其保存到 Excel 表格中。为了获取热门地区排行，基于 SPSS 软件，进行 17 省份及其六个影响因素的主成分分析。

首先，使用 SPSS 软件功能，将数据标准化，再进行降维（主成分分析）。得到解释的总方差和成分矩阵，如表 5、表 6。

表 5 解释的总方差

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	2.043	34.051	34.051	2.043	34.051	34.051
2	1.566	26.096	60.146	1.566	26.096	60.146
3	0.899	14.987	75.133	0.899	14.987	75.133
4	0.749	12.477	87.61	0.749	12.477	87.61
5	0.452	7.533	95.143			
6	0.291	4.857	100			

表 6 成分矩阵

	成分			
	1	2	3	4
Zscore(number)	0.659	0.558	0.329	0.394
Zscore(salary)	0.73	0.502	0.036	0.021
Zscore(companysize)	-0.037	-0.559	0.798	-0.029
Zscore(companytype)	-0.531	0.646	0.144	-0.566
Zscore(experience)	0.623	-0.15	0.129	0.344
Zscore(education)	0.636	-0.501	-0.341	0.304

由表 5 可知，上述标准化的数据在进行主成分分析后，得到四个主成分，其累计方差百分比为 87.61%，大于 85%，符合主成分分析方差提取原则。

用成分矩阵中数值除以相应成分的特征值的开方，得到特征向量矩阵表 7。

表 7 特征向量矩阵

Z1	Z2	Z3	Z4
0.46	0.45	0.35	0.46
0.51	0.40	0.04	0.02
-0.03	-0.45	0.84	-0.03
-0.37	0.52	0.15	-0.65
0.44	-0.12	0.14	0.40
0.44	-0.40	-0.36	0.35

计算 17 个省的各个主成分的得分值，将表 7 的特征向量与标准化的数据相乘。则有 F1、F2、F3、F4 的主成分表达式为：

$$F1 = 0.46ZX_1 + 0.51ZX_2 - 0.03ZX_3 - 0.37ZX_4 + 0.44ZX_5 + 0.44ZX_6$$
$$F2 = 0.45ZX_1 + 0.40ZX_2 - 0.45ZX_3 + 0.52ZX_4 - 0.12ZX_5 - 0.40ZX_6$$
$$F3 = 0.35ZX_1 + 0.04ZX_2 + 0.84ZX_3 + 0.15ZX_4 + 0.14ZX_5 - 0.36ZX_6$$
$$F4 = 0.46ZX_1 + 0.02ZX_2 - 0.03ZX_3 - 0.65ZX_4 + 0.40ZX_5 + 0.35ZX_6$$

计算热门地区综合指标的具体算法为令每个主成分的特征值占总特征值的比例作为权重[12]，得到公式：

$$F = \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}F1 + \frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}F2 + \frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}F3 + \frac{\lambda_4}{\lambda_1+\lambda_2+\lambda_3+\lambda_4}F4$$

按 F 的大小进行排序，得到热门地区综合指标排序如表 8。

表 8 热门地区综合指标排行

area	F1	F2	F3	F4	F	排名
广东	1.73	2.06	1.47	1.24	1.71	1
上海	2.86	0.67	0.05	1.85	1.58	2
北京	1.66	1.07	-0.66	0.67	0.95	3
江苏	1.04	0.33	-0.35	1.02	0.59	4
浙江	1.20	-0.34	0.07	0.74	0.48	5
湖北	-0.69	0.91	0.67	-0.82	0.00	6
江西	1.82	-3.88	-0.24	2.92	-0.07	7
四川	-0.43	0.05	0.38	-0.44	-0.15	8
陕西	-0.52	0.59	-0.51	-0.70	-0.21	9
湖南	-0.93	-0.22	1.49	-0.95	-0.31	10
云南	-0.45	-1.26	0.72	0.06	-0.42	11
重庆	-0.32	0.47	-2.34	-0.35	-0.43	12
山东	-0.94	0.38	-1.26	-0.44	-0.53	13
安徽	-0.90	-0.48	-0.20	-0.93	-0.66	14
福建	-1.36	-0.45	0.85	-1.13	-0.68	15
河南	-2.09	0.58	-0.39	-1.54	-0.92	16
辽宁	-1.68	-0.49	0.26	-1.19	-0.93	17

我们发现广东、上海、北京、江苏、浙江依次是在互联网行业热门职位中最为热门的城市。

第二，对热门新兴技术综合指标进行排序，在 PCA 表中主要包括大数据、5G、AI、云计算、区块链、物联网六个职位关键词，共计 3500 条。同上，分别计算 6 个职位关键词的招聘信息数量、平均薪资、平均公司类型得分、平均公司规模得分、平均经验要求得分、平均学历要求得分，并将其保存到 Excel 表格中。使用 SPSS 软件功能，将数据标准化，再进行降维（主成分分析）。得到解释的总方差和成分矩阵，如表 9、表 10。

表 9 解释的总方差

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	3.206	53.439	53.439	3.206	53.439	53.439
2	1.525	25.415	78.854	1.525	25.415	78.854
3	0.846	14.094	92.948	0.846	14.094	92.948
4	0.41	6.836	99.784			
5	0.013	0.216	100			
6	1.11E-16	1.85E-15	100			

表 10 成分矩阵

	成分		
	1	2	3
Zscore(number)	-0.484	-0.133	0.864
Zscore(salary)	0.537	-0.82	0.089
Zscore(companysize)	-0.762	0.641	0.059
Zscore(companytype)	0.695	0.486	0.071
Zscore(experience)	0.921	0.214	0.228
Zscore(education)	0.878	0.375	0.178

由表 10 可知，提取三个主成分，其累计方差百分比为 92.948%，大于 85%，符合主成分分析方差提取原则。同上得到特征向量矩阵表 11。

表 11 特征向量矩阵

Z1	Z2	Z3
-0.27	-0.11	0.94
0.30	-0.66	0.10
-0.43	0.52	0.06
0.39	0.39	0.08
0.51	0.17	0.25
0.49	0.30	0.19

将表 11 的特征向量与标准化的数据相乘，得到 F1、F2、F3 主成分表达式，并令每个主成分的特征值占总特征值的比例作为权重，计算出 F，按 F 的大小进行排序，得到热门职位关键词综合指标排行，如表 12。

表 12 热门职位关键词综合指标排行

keyword	F1	F2	F3	F	排行
区块链	2.99	-1.23	-0.33	1.33	1
物联网	0.27	2.06	-0.16	0.69	2

表 12 (续表)

大数据	0.38	0.19	1.74	0.53	3
云计算	-0.04	0.35	-0.88	-0.06	4
5G	-1.33	-0.11	-0.52	-0.87	5
AI	-2.27	-1.26	0.15	-1.63	6

我们发现，区块链、物联网、大数据依次是互联网行业中更为热门的职位关键词。

5 网络招聘信息可视化

如今网上招聘信息复杂多样，不少求职者没有办法快速获得一些有实际价值和符合自己需求的信息。为了让信息更直观、更清晰的展示，本文将通过建立网站系统，将数据挖掘和分析的结果，用数据可视化的方式表现出来。

网站功能主要包括：按条件查询招聘信息（可直接访问职位详情地址）、招聘信息数量地区分布动态图、薪资排行柱状图（不同地区平均薪资排行、不同关键词平均薪资排行、不同学历平均薪资排行等）、三类词云图、按条件查询薪资等级等功能。

5.1 基于 Flask 框架的网页构建

本文选用 python 中的 Flask 框架，因为 Flask 更为灵活，并且扩展包很多样。Flask 本身对所使用数据库没有要求，所以选择本文已经保存数据的 SQLite 数据库。

前后端交互主要实现过程包括：当用户访问网站，并打算获取相关信息时，可以在页面的输入框内输入相关内容。在 Flask 框架中，在后端会获取用户所输入的信息，进行函数处理，本文可视化网站的后端函数主要包括：主页关键词搜索-结果页面、薪资判断系统-结果页面、职位分布动态图、词云图等。最后在根据获取用户在网站中输入的信息，经过函数计算，将结果通过 jinja2 模板引擎对结果页面实现渲染，并将结果页面展示给用户。

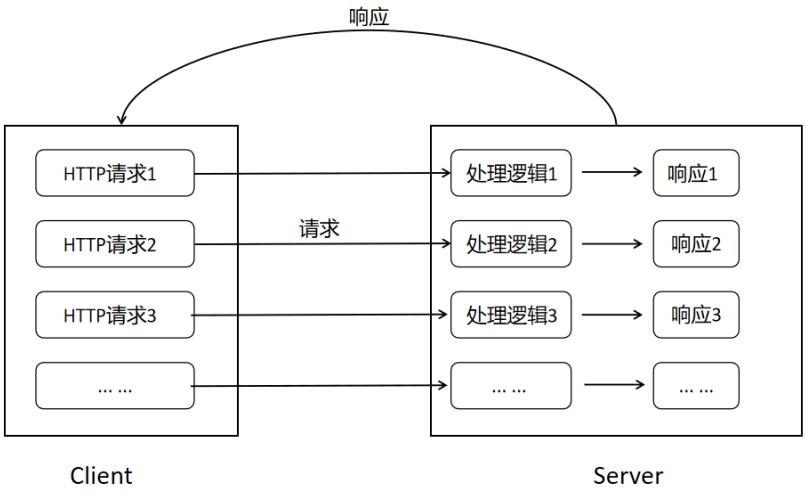

图 5　Flask 网页框架

5.2 平均薪资排行界面

对于互联网相关专业的毕业生，在找工作的工程中，薪资的高低一定程度上影响着其选择。针对本文所爬取的数据，根据不同的地区、职位关键词、公司类型、学历要求和经验要求的平均薪资排行进行了可视化，如下图所示。

图 6 地区平均薪资排行

如图 6 可以看出，在互联网行业热门职位中，传统一线城市上海、北京等平均薪资明显高于其他城市，可见相比于其他城市，大数据、云计算、AI、区块链、物联网、5G 等新兴技术在最发达的一线城市发展更好。对于求职者来说，北上广深有更好的机遇、待遇更好。但随着越来越多的人涌入，竞争压力越来越大，求职者还应根据自身需求来选择工作城市。

图 7 职位关键词薪资排行

如图 7 可以看出，在互联网行业热门新兴技术中，各类技术本身平均薪资相差不大，均在月薪 2 万/月左右。但在所有行业中，互联网行业的相关职位薪资水平依然名列前茅。虽然互联网行业薪资水平较高，但随着越来越多的人加入该行业，并且技术也在不断升级换代，导致该行业的相关职位从业时间较短，对于求职者而言要想清楚在选择。

图 8 公司类型平均薪资排行

如图 8 可以看出，在互联网行业热门职位中，外资（欧美）公司平均薪资超过 2 万/月，位居第一。这是由于新兴技术在欧美国家中发展较快，对人才需求更大。但与其他公司类型薪资水平基本相差不多。

图 9 学历要求平均薪资排行

图 10 经验要求平均薪资排行

如图 9、图 10 可见看出，薪资水平与经验和学历成正比，经验要求越高平均薪资越高，学历越高平均薪资越高。

对于求职者来说，选择提高学历或丰富经验，都是提高薪资水平最直接的途径。目前考研人数逐年增加，竞争压力越来越大，当大学生选择保研、考研或读博时，要努力坚持实现目标。对于考研失败或不想考研的学生来说，多参加实习和提高实践能力，也是一种更好的选择。

5.3 三类关键词词云图

首先，调用 python 中的 jieba 库进行中文分词、matplotlib 库进行绘图和数据可视化、wordcloud 库实现词云图。主要步骤包括：调用数据库 job 表中“工作待遇”、“工作地点”、“职位名称”三类属性的数据，并分别将其处理为一个字符串。进行中文分词，生成词云图，并保存。

根据地域分布词云图，我们可以发现深圳南山区、上海浦东新区、广州天河区、北京海淀区在互联网行业职位中是最为热门的地区。根据职位名称词云图，可以发现开发工程师、大数据、算法工程师、物联网等在互联网行业中比较热门。根据福利地域词云图，我们发现在互联网行业各种职位中，福利待遇最常见的依次分别是五险一金、定期检查、餐饮补贴、员工旅游等。

图 11 地域分布词云图

图 12 职位名称词云图

图 13 福利待遇词云图

5.4 关键词搜索引擎

关键词搜索引擎主要包括两部分：搜索页面和结果页面。

首先，搜索页面导入 HTML 文件模板，其中搜索框主要输入内容包括职位关键词、工作地区、薪资，如图 14。

图 14 搜索框

通过 request 将输入数据保存下来，调用数据库，使用查询语句进行相关词查询，并将结果传入到结果页面，最终以表格的形式展示查询结果。表格包括相关搜索的所有招聘信息，并且可以点击职位名称直接跳转到该招聘信息详情页面，如图 15。

相关职位详情

关键词:大数据　地区:北京　薪资:万/月

职位名称	地区	薪资	公司名称	公司类型	公司规模	福利待遇
大数据开发工程师	北京-海淀区	1.5-3万/月	北京汉诺威尔能源科技有限公司	民营公司	少于50人	五险一金 员工旅游 定期体检 年终奖金 绩效奖金 节日福利 商业保险 人文关怀
大数据开发工程师(j11102)	北京-海淀区	1.5-2万/月	北京先进数通信息技术股份公司	上市公司	1000-5000人	五险一金 餐饮补贴 绩效奖金 年终奖金 定期体检 每年调薪
大数据工程师	北京	20-40万/年	海达保险经纪有限公司	国企		
中级大数据开发工程师	北京-东城区	1.5-2.5万/月	国研大数据研究院有限公司	国企	50-150人	五险一金 补充医疗保险 餐饮补贴 通讯补贴 弹性工作 定期体检
大数据架构师	北京	1.3-2.3万/月	北京市华都峪口禽业有限责任公司	合资		

图 15 查询结果

5.5 招聘信息地区分布动态图

为了更加清晰地展示出互联网行业职位地区分布数量，通过 E charts 进行数据可视化。主要步骤包括：首先，在 HTML 文件中引入 E charts 的中国地图模版，并设置显示区域。其次，初始化 E charts，指定配置项和数据。通过查询数据库中各地区的数量，将结果数据传入地图各省数据集中。由于数据库中共 4000 条左右招聘信息，所以对于各省设置 0-500 范围，依次按蓝黄红表示数量程度。最后，设置配置项显示图表。如图 16。

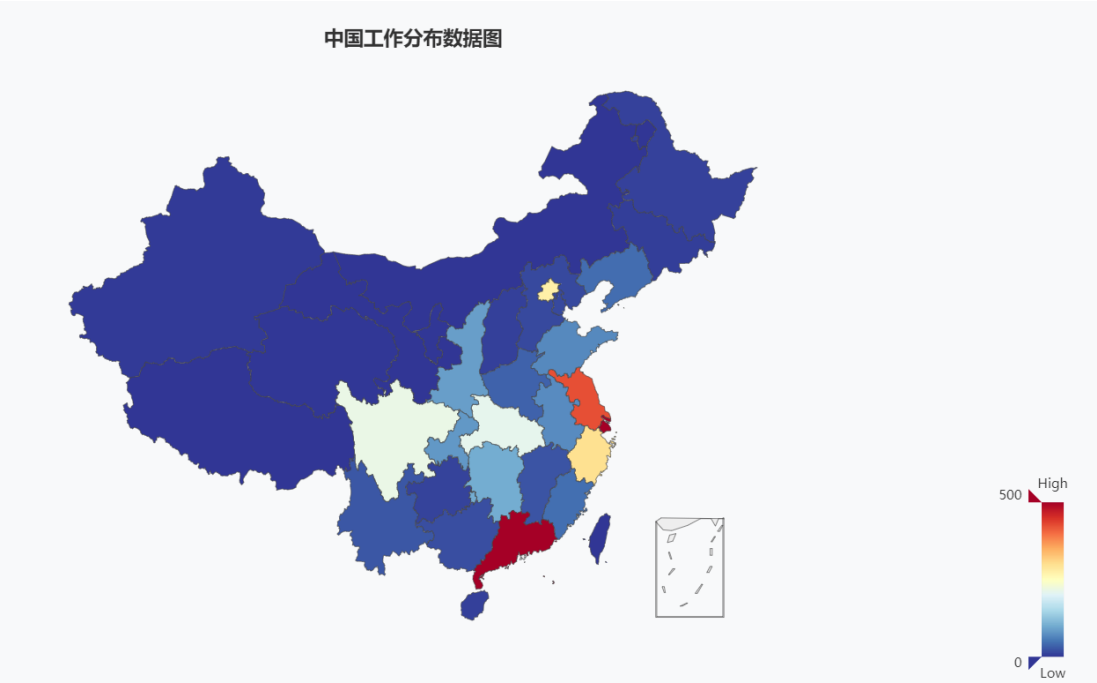

图 16 互联网行业招聘信息地区分布

5.6 基于决策树模型的薪资等级评判系统

根据已经建立的决策树模型，在后端函数中编写相应 IF-ELSE 语句。通过获取前端页面输入数据，带入判断语句中进行判断。输入页面如图 17。

图 17 薪资评判系统输入框

最终将薪资等级结果传入结果页面，展示给用户，结果页面如图 18。

图 18 薪资评判系统结果页面

通过 python 代码的编写，基于 Flask 框架的前后端交互，实现了薪资等级评判系统。访问者可以根据自身实际条件或目标条件，输入该系统，得到基本的薪资等级。该系统的实现，有利于帮助求职者了解不同条件下的薪资等级，部分求职者在无法改变某个条件时，为了获得更高的薪资，从而努力改变其他条件。

如大专生为例，在已经无法改变自身学历条件时，根据系统可以发现，当具有较高的经验水平时，并且在北京、沿海一线城市时，还是可以获得“A”级别的薪资水平。

6 总结与展望

6.1 总结

本文从实际需求出发，旨在帮助互联网行业的求职者进一步了解网络招聘信息和其内在联系。故本文以 51job 招聘网站为例，爬取互联网行业新一代信息技术“大数据”、“5G”、“AI”、“云计算”、“物联网”、“区块链”为关键词的招聘信息，进行预处理，并保存到 SQLite 数据库中，为接下来的数据挖掘与数据可视化做准备。

由于数据挖掘手段和方法丰富多样，本文根据实际需求和方法适用性，选择较为常见的分类决策树模型和主成分分析进行数据挖掘，最终得到有效模型和结论，并建立网站实现数据可视化。以下是本文主要研究结果：

第一，本文根据数据库的数据，基于 python 代码，利用决策树 C4.5 算法和 REP 剪枝，构建出决策树模型，对互联网行业热门职位的薪资影响因素进行了分析。结论是：对于互联网行业热门职位的薪资来说，学历为第一影响因素，经验为第二影响因素、地区为第三影响因素。首先，硕士学历平均薪资大于 20 万/年，初高中平均薪资小于 10 万/年；本科学历且具有 3 年以上经验平均薪资大于 20 万/年，否则在 10 万/年到 20 万/年之间；大专学历且具有 3 年以上经验并在一线城市平均薪资大于 20 万/年，有经验但不在一线城市和没经验在一线城市平均薪资均在 10 万/年到 20 万/年之间，没经验且不在一线城市平均薪资小于 10 万/年。

第二，本文利用 PCA 进行了热门地区和热门新兴技术关键词综合指标排行。结论是，在互联网行业热门职位中，广东、北京和上海仍是最热门的省份；区块链、物联网、大数据是最热门的新兴技术。

第三，本文利用文本挖掘技术，创建福利待遇词云图、职位地点词云图、职位名称词云图。在互联网行业热门职位中，在职位福利待遇中“五险一金”最常见，在职位地点中“深圳南山区”、“上海浦东区”最常见，在职位名称中“大数据”、“开发工程师”最为常见。

第四，本文通过 python 代码，基于 Flask 框架，调用 E Charts 图表，创建网站，将数据可视化，为求职者提供全面且细致的浅层和深层信息。功能包括：薪资排行柱状图（不同地区平均薪资排行、不同关键词平均薪资排行、不同学历平均薪资排行、不同经验平均薪资排行、不同公司类型平均薪资排行）、按条件查询招聘信息（可直接进入职位详情地址）、招聘信息数量地区分布动态图、三类词云图、按条件查询薪资等级。

6.2 建议

对大学生和求职者的建议：第一，对于想要追求高薪资的学生，可以考虑在上海、北京这类一线城市从事区块链、AI 相关的工作。另外，从薪资影响因素分析模型中得知，对于大学生来说，学历是首要影响因素，学历越高，薪资水平越高，所以选择保研、考研、读博等都是一种比较好的选择。其次，是经验水平，当学历不占优势时，可以通过多年的实习、工作的经验累积，同样使自己获得较高水平的薪资，所以对于那些本科毕业就想就业的大学生来说，在大学生活中多参加实习、多积累经验、多提高技术也是十分重要。最后，是工作地区，虽然北京、沿海一线城市的薪资明显高于其他地区，但一线城市消费水平高、工作压力大也是实际存在的问题，所以就业的大学生应该充分考虑，结合自己实际情况来进行最后的选择。

第二，对于想要追求竞争压力小的学生而言，根据热门地区和热门职位关键词的 PCA 综合指标排行，可以选择一些排名相对靠后，同时满足自己需求的地区和职位。比如辽宁省和福建省，相较而言没有那么热门，竞争压力较小，同时平均薪资水平中等，适合这些学生选择。

第三，对于社会中的求职者来说，可能自身部分条件已经无法改变，如学历、地区等，但同样可以根据薪资影响因素模型来提高薪资等级。以大专生为例，当进入社会后学历已经无法在更改，根据模型，如果具备三年以上经验且在北京沿海等一线城市，同样薪资水平可以达到 20 万/年，甚至与某些研究生毕业工资相同，这

个现象也说明了如今企业更多招的是专业性人才，更看重的是实际操作技能，而不是空有学历的大学生。所以求职者可以结合地区，丰富工作经验，都能取得一个较为满意的薪资水平。

第四，对于在校大学生而言，根据网站信息，在校大学生可以提前规划目标，并为之努力，不断丰富经验或提高学历，在如今就业压力巨大的市场环境中，只有清晰的自我定位，才能在茫茫求职者中脱颖而出，抓住时代发展的机遇，迎合新一代信息技术的发展。

总而言之，无论对于学生还是社会人员来说，在求职过程中，要清楚薪资高低并不能完全决定生活质量，薪资也不应该衡量工作的唯一标准。福利待遇、企业文化、工作地点、领导风格等都应该作为考虑的因素，找一份适合自己的工作是更重要的。另外，想要找到一份待遇更好、薪资更高的工作，更重要的是要提高自己的专业技能和实战经历，高学历和丰富的经验都一定程度代表者求职者的能力。作为求职者只有不断的努力，不断完善理论知识，不断提高自身实践能力，最重要的是要找到自己最擅长的方向，精比广更重要，主动迎合如今企业对专业性人才的需求方向，最终才能实现一个更好的生活。

对于企业和高校的建议：对于企业，根据网站所提供的信息，更好的了解如今新一代信息技术相关的招聘形势。根据热门地区排行，在结合公司自身状况，有条件的在热门地区开展分公司和人才招聘。比较相关类型其他的公司的薪资和福利待遇，提高自身，使企业在行业中更具竞争力。另外，可以多与高校开展合作，助力培养专业性人才，为公司引进做好先决条件。

对于高校，如今社会中招聘岗位越来越注重求职者的专业能力，如今的绝大多数大学生仅仅停留在书本知识中，实践经历微乎其微。为了学校提高就业率和缓解当今时代的就业压力，院校应该有针对的开设专业性课程，有针对的开展培训丰富实践经历。结合如今的企业招聘的形式，来制定既满足学校学习要求又满足企业专业能力要求的培养方案。例如，学校可以有针对的增加毕业生毕业要求，可以将 1-3 个月实习经历作为毕业硬性要求。同时，学校可以与企业合作，为企业提供实习生，或学生自行寻找实习机会。再例如，院校可以根据社会企业实际专业要求，聘请企业培训人员，开展课程，提高学生在企业工作中的实践能力。

6.3 未来展望

第一，本文研究的仅仅是互联网行业热门职位的网络招聘信息，但所有模型与分析均适用于其他行业，可以根据需求爬取不同行业的网络招聘信息数据保存到 SQLite 数据库中，并且同时可以爬取除了 51job 网站之外的其它招聘网站信息。未来可以随着招聘信息的更新，对数据库不断进行更新，将除互联网行业的其他行业招聘信息都录入数据库中，将可视化页面适用各行各业的求职者。

第二，对于数据挖掘技术在网络招聘信息中的应用，本文仅采用了分类决策树模型和主成分分析，其他数据挖掘技术，如关联算法、神经网络模型等均可以使用。另外可以对文本数据进行充分挖掘，采用 TF-IDF 算法和 LDA 模型等，挖掘文本中有价值的信息。

第三，在未来，对于可视化网站，可以继续完善其他行业的招聘信息，继续完善网站的功能，并基于云服务器进行网站搭建，实现其他用户的访问。另外可以完善后端，将从已爬取完成的数据库调取数据，转换成随用户访问时间而实时爬取的方式。

第四，本文在使用 C4.5 算法构建决策树时，为了降低决策树的复杂度，仅选择学历、经验、地区作为薪资影响因素。未来可以考虑更多的影响因素，如公司类型、职位关键词等，虽然剪枝前的决策树可能较为复杂，但可以通过剪枝来进行决策树简化。当考虑多个影响因素后，对人群划分更细致，对薪资评级更为准确。另外，薪资等级仅包括 A（大于 20 万/年）、B（大于 10 万/年且小于 20 万/年）、C（小于 10 万/年）。未来可以根据更加具体的研究，来进行级别的划分，增加级别类型，降低级别范围。生成的决策树可以提供更具体的结果。

第五，本文在进行决策树模型剪枝时，测试集共 1000 条，但某些部分类别数量占比较小，如硕士学历仅 87 个，在采用 REP 方法时，容易出现过度修剪现象，模型普适性降低。在未来可以增加测试集数量，提高到 10000 条。另外也可以采用 PEP 方法来进行剪枝，PEP 方法不需要额外的测试集，仅用训练集即可，是目前决策树后剪枝方法中精度较高的算法之一。

参考文献

[1] 孙立伟,何国辉,吴礼发.网络爬虫技术的研究[J].电脑知识与技术,2010,6(15):4112-4115.

[2] 万玛宁,关永,韩相军.嵌入式数据库典型技术 SQLite 和 Berkeley DB 的研究[J].微计算机信息,2006(02):91-93+272.

[3] 王光宏,蒋平.数据挖掘综述[J].同济大学学报(自然科学版),2004(02):246-252.

[4] 周雪忠,吴朝晖.文本知识发现:基于信息抽取的文本挖掘[J].计算机科学,2003(01):63-66.

[5] 蒲海坤,高鑫,桑鑫.基于 C4.5 数据挖掘算法研究与实现[J].科学技术创新,2021(23):55-56.

[6] 路逸行. 决策树误差降低剪枝算法的改进研究[D].山东大学,2020.

[7] 虞晓芬,傅玳.多指标综合评价方法综述[J].统计与决策,2004(11):119-121.

[8] 叶锋.Python 最新 Web 编程框架 Flask 研究[J].电脑编程技巧与维护,2015(15):27-28.

[9] 崔蓬.ECharts 在数据可视化中的应用[J].软件工程,2019,22(06):42-46.

[10] 甯文龙,毛红霞.基于 Python 爬虫技术的 51job 网站内容爬取[J].信息与电脑(理论版),2021,33(04):180-182.

[11] 王黎明. 决策树学习及其剪枝算法研究[D].武汉理工大学,2007.

[12] 刘根霞.基于主成分分析法的上市公司综合竞争力排行——以电子产业为例[J].财会通讯,2012(35):15-17.