文章编号:2096-1472(2023)10-0058-05
DOI:10.19644/j.cnki.issn2096-1472.2023.010.013

基于协同过滤算法的智能岗位分析系统的设计与实现

陈 亮

(大连东软信息学院,辽宁 大连 116023)
✉ chenliang@neusoft.edu.cn

摘 要:针对现有招聘网站种类繁多和求职者无法快速根据自身特点找到适合岗位等问题,基于大数据技术和协同过滤算法,设计和实现了一个智能岗位分析系统。利用 Python 爬虫技术、虚拟化技术、Hadoop 大数据平台及其生态组件进行数据处理,使用协同过滤算法进行智能分析,通过 Axure RP 和 Sugar BI 工具对结果数据进行可视化展示。经试用,系统能够很好地满足求职者的分析需求,提高了求职者精准查询招聘信息的效率。

关键词:岗位分析;智能推荐;数据仓库;可视化
中图分类号:TP311 **文献标志码**:A

Design and Implementation of Intelligent Job Analysis System Based on Collaborative Filtering Algorithm

CHEN Liang

(*Dalian Neusoft University of Information, Dalian* 116023, *China*)
✉ chenliang@neusoft.edu.cn

Abstract: This paper proposes to design and implement an intelligent job analysis system based on big data technology and collaborative filtering algorithm to address the issues of a wide variety of existing recruitment websites and the inability of job seekers to quickly find suitable positions based on their own characteristics. Python crawler technology, virtualization technology, Hadoop big data platform and its ecological components are used for data processing, collaborative filtering algorithm is used for intelligent analysis, and the resulting data is visualized through Axure RP and Sugar BI tools. After trial, the proposed system can well meet the analysis needs of job seekers and improve the efficiency of job seekers to accurately query recruitment information.

Key words: job analysis; intelligent recommendation; data warehouse; visualization

0 引言(Introduction)

据教育部公布的最新数据显示,2022 年高校应届毕业生人数再创新高,突破 1 000 万人,整体社会的就业压力依然巨大[1]。对于求职者来说,招聘网站是其获取求职信息的主要方式,目前市面上已有许多类型的招聘网站,这些网站会定期发布一些企业的招聘需求,但是这些招聘信息的数据量庞大,求职者想要在海量的招聘信息中找到适合自身需求的岗位十分困难,这些网站存在的一个普遍问题是只为企业发布招聘信息和求职者搜索招聘信息提供了一个平台,但是并不能给求职者提供高效、系统性的专业建议和指导,求职者在这些平台上也无法快速、准确的地获取自己需要的企业招聘信息。基于此,本文设计实现了一个基于协同过滤算法的智能岗位分析系统,旨在利用大数据和人工智能技术对海量的招聘信息数据进行智能分析和处理,不仅可以让求职者更加直观地了解目前的就

收稿日期:2023-02-20

业行情与需求,也可以让求职者更快速和便捷地获取适合自己的岗位需求信息。

1　系统数据架构(System data architecture)

智能岗位分析系统整体数据架构包括数据采集、数据存储、数据分析和数据可视化展示等部分。数据源主要来自主流招聘网站上公开的招聘信息,采集技术采用 Python 爬虫框架 Scrapy,原始数据存储在 Hadoop 平台分布式文件系统 HDFS 上,通过 Hive 进行数据查询和处理,得到的数据结果通过 Sqoop 导入 MySQL 数据库,通过机器学习领域的协同过滤算法进行智能化分析,最后通过可视化技术对结果数据进行展示。系统数据架构图如图 1 所示。

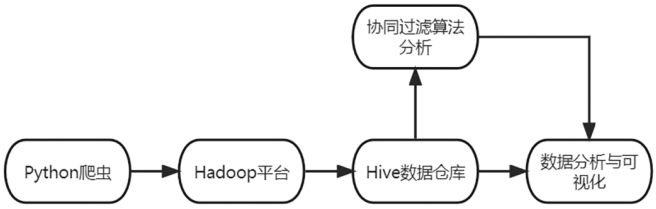

图 1　系统数据架构图

Fig. 1 System data architecture diagram

2　系统设计(System design)

2.1　系统用例设计

本系统主要包括两个角色,分别为管理员和普通用户。满足用户基本业务需求的用例是高层用例,这些用例包括用户基本操作和管理员基本操作。高层用例图如图 2 所示。

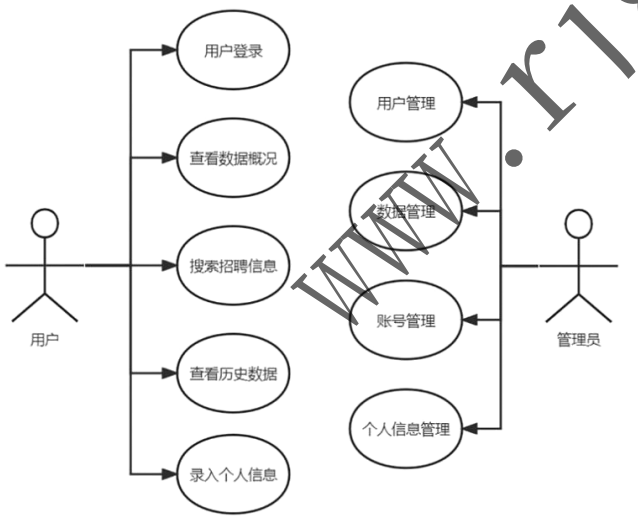

图 2　高层用例图

Fig. 2 High-level use case diagram

2.2　数据采集与清洗

数据采集部分采用 Python 爬虫框架技术 Scrapy,获取主流招聘网站的招聘信息,作为整个系统的原始数据源。Scrapy 是开源快速的网络爬虫框架,可以从网站获取网页数据信息,并从页面中得到用户想要的数据,它的核心是 Scrapy engine 爬虫引擎,通过 Scheduler 调度模块模拟发送 HTTP 请求、Downloader 下载器模块接收并生成页面响应,Spider 爬虫程序模块迭代提取网页中的数据内容,Item Pipeline 数据管道模块对获得的数据进行持久化的存储[2]。Scrapy 爬虫框架如图 3 所示。

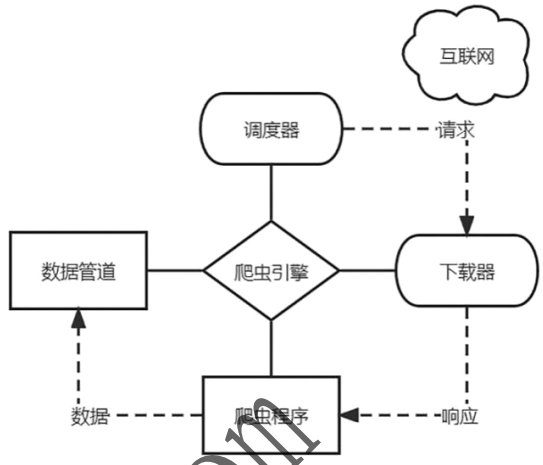

图 3　Scrapy 爬虫框架

Fig. 3 Scrapy crawler frame

2.3　数据平台搭建

平台环境搭建采用虚拟化技术虚拟出三台 Linux 服务器构成集群,主机名分别设为 shixun01、shixun02、shixun03。集群配置 shixun01 CPU 核心数为 4,磁盘空间为 50 GB,内存大小为 8 GB;shixun02 CPU 核心数为 2,磁盘空间为 50 GB,内存大小为 4 GB;shixun03 CPU 核心数为 2,磁盘空间为 50 GB,内存大小为 4 GB[3]。在搭建好的数据平台上安装 Hadoop、MySQL、Hive、Sqoop 等软件工具。数据平台如图 4 所示。

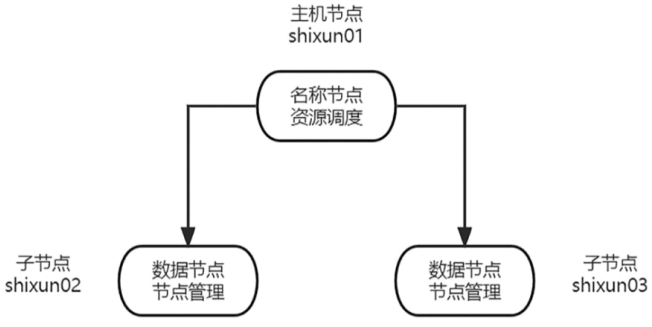

图 4　数据平台

Fig. 4 Data platform

2.4　数据仓库建设与开发

数据仓库使用 Hive 技术进行建设。整体数据仓库架构分为原始数据层、基础数据层、明细数据层、聚合数据层和应用数据层。原始数据层接收采集的原始数据,基础数据层存储经过清洗后的原始数据,明细数据层根据业务场景将基础数据进行细化分类,聚合数据层根据业务主题和需求提前聚合相关统计数据,应用数据层根据需求存储用于产出可视化图表的应用结果数据。数据仓库架构图如图 5 所示。

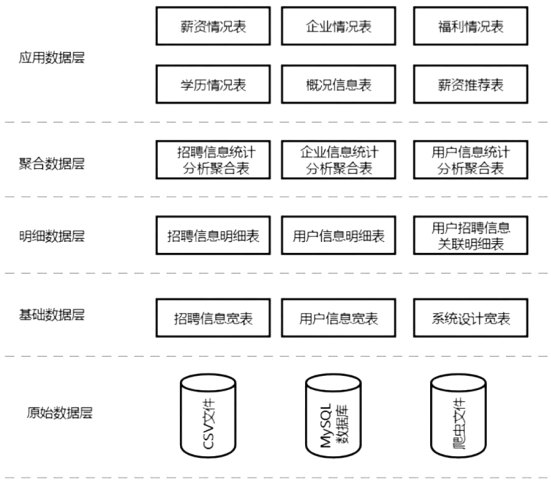

图 5 数据仓库架构

Fig. 5 Data warehouse architecture

2.5 算法分析

预测问题一直是机器学习领域中最重要的问题。很多算法包括回归算法、决策树算法等都是用来解决预测的常用算法。本系统预测算法采用经典的协同过滤算法，首先依据用户属性特征，找到具体相似兴趣的用户，其次根据用户评价矩阵以及对产品的评价结果构建协同过滤算法，进而预测其他未评分的项目或者用户，最后根据预测出的结果对用户进行推荐。

该算法的基本操作步骤如下：①利用已经拥有的用户行为历史数据，构造用户项目评分矩阵；②通过相似度计算公式计算用户之间的相似度，将相似度较高的用户当作目标用户的近邻集；③在进行评分预测后，按照 TOP-N 原则为用户进行推荐[4]。

2.5.1 构建用户项目评分矩阵

构建用户项目评分矩阵 $\boldsymbol{R}_{mn}$，矩阵行中有 m 个用户，用 U 表示，$U=\{U_1,U_2,\cdots,U_m\}$，矩阵列中有 n 个项目，用 I 表示，$I=\{i_1,i_2,\cdots,i_n\}$，R_{ij} 表示用户 i 对项目 j 的实际评分，若用户 i 对项目 j 未评分，则 R_{ij} 为 0，用户项目评分矩阵公式如下：

$$\boldsymbol{R}_{mn}=\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{bmatrix} \tag{1}$$

2.5.2 用户评分相似度计算

用户评分的相似度计算以用户项目评分矩阵为基础，用评分矩阵中的每一行的评分向量表示用户的实际兴趣。所以，计算用户评分的相似度实质上就是计算用户评分向量之间的距离[5]。传统的协同过滤算法中最常用的计算相似度的方法是皮尔逊相似度计算方法，其计算公式如下：

$$sim(a,b)=\frac{\sum_{i\in T}(R_{ai}-R_a)(R_{bi}-R_b)}{\sqrt{\sum_{i\in T}(R_{ai}-R_a)^2}\sqrt{\sum_{i\in T}(R_{bi}-R_b)^2}} \tag{2}$$

在获取用户 a 和其他全部用户的相似度后，将相似度排名最高的前 h 个用户作为该用户的近邻集，应用评分预测公式得出最终的预测评分。评分预测公式如下[6]：

$$P_{a,j}=R_a+\frac{\sum_{b\in Q}sim(a,b)(R_{b,j}-R_b)}{\sum_{b\in Q}|sim(a,b)|} \tag{3}$$

本系统可以实现智能化求职者薪资预测功能，根据用户输入的条件和用户的浏览记录信息等数据，运用传统的协同过滤算法和皮尔逊相似度计算方法，计算出用户评分向量之间的距离，应用评分预测公式得出最终的预测评分，测算出匹配求职者条件和能力的薪资范围，并响应到前端模块。推荐流程图如图 6 所示。

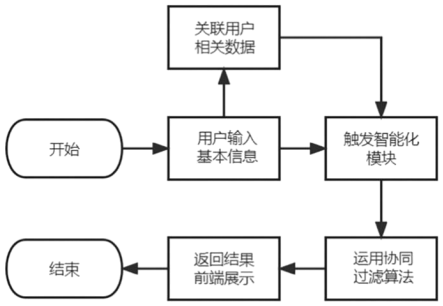

图 6 推荐流程图

Fig. 6 Flow chart of recommendation

2.6 数据的可视化展示

可视化展示部分前端采用 Axure RP 工具，它是一种用来进行原型设计的专业工具，可以快速地创建网站原型和应用软件原型，同时可以定义需求和规格，生成网站和应用软件规格说明文档[7]。网站内的分析图采用第三方可视化工具 Sugar BI，Sugar BI 基于百度 Echarts，能提供丰富的图表组件，开箱即用、零代码操作、不需要 SQL，降低开发成本的同时，还能提高业务对数据的使用效率[8]。Sugar BI 支持多种方式对接数据源，如直连数据库、上传 Excel/CSV 文件、API 接口、静态 JSON 录入等[9]。

3 系统实现(System implementation)

3.1 首页展示

用户进入系统首页，可以进行注册和登录，首页显示可视化展示系统、智能招聘系统和需求分析系统等功能入口。系统首页如图 7 所示。

图 7 系统首页

Fig. 7 System homepage

3.2 数据概况界面

点击进入数据概况界面，界面显示的信息包含公司全称、公司简称、公司规模、融资阶段、区域、职位名称、工作经验、学历要求、薪资、职位福利、经营范围、职位类型。界面上方包含查询功能和搜索功能，用户能更清晰、直观地找到适合的职位。数据概况界面如图 8 所示。

图 8　数据概况界面

Fig. 8 Data overview interface

3.3　可视化模块

为了能让用户更好地分析自己的能力和找到合适的岗位需求信息,系统通过文字云图、柱状图、饼状图、漏斗图、矩形数形图等形式分别对企业发布的薪资情况、企业情况、公司规模分布、学历和工作经验分布等进行了详细的可视化展示。企业发布的薪资概况界面如图 9 所示。

图 9　薪资概况界面

Fig. 9 Salary overview interface

企业概况界面如图 10 所示。

图 10　企业概况界面

Fig. 10 Enterprise overview interface

3.4　智能化模块

目前,系统的智能化模块已完成用户薪资预测功能,用户输入相关信息后,系统就能根据算法模型预测其最低薪资标准,并在前端进行展示。薪资预测功能如图 11 所示。

图 11　薪资预测功能

Fig. 11 Salary forecasting function

3.5　关键技术难点

用户评分矩阵对于协同过滤算法来说,是十分重要的概念,主要作用是计算项目间或用户间的相似度,用户评分矩阵的稀疏程度对预测结果有明显的影响。如果用户评分矩阵特别稀疏,整体的预测和推荐的质量会大幅下降,所以如何解决用户评分矩阵的稀疏性,是提高协同过滤算法预测和推荐质量的核心。

皮尔逊相似度的计算方法在计算的过程中不会使用缺失数据,所以本文使用皮尔逊相似度计算时不用考虑数据稀缺的问题,而是需要着重考虑共同评分项数目不同的问题,可以使用预测数据填充的方法解决未知评分的问题,具体方法如图 12 所示[10]。

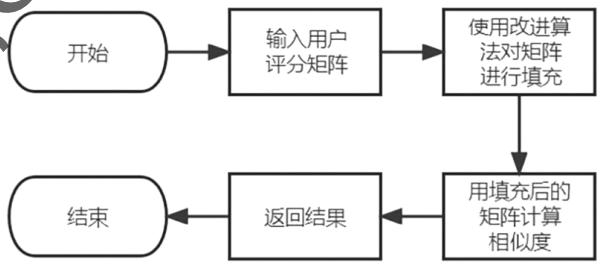

图 12　改进的算法过程

Fig. 12 Improved algorithm process

4　结论(Conclusion)

本系统围绕招聘系统无法聚焦和智能分析的问题,设计了一个集招聘信息可视化展示和智能分析于一体的系统。该系统包括数据采集、数据清洗、平台搭建、数据仓库建设、智能分析以及可视化展示等部分;系统通过 Python 爬虫技术获取主流招聘网站的数据作为原始数据源,通过虚拟化技术和 Linux 操作系统搭建 Hadoop 大数据平台,通过 Hive 技术进行数据仓库建设和数据处理,将得到的结果数据通过 Sqoop 导入 MySQL 数据库,通过协同过滤算法进行智能分析,通过 Axure RP 和 Sugar BI 对结果数据进行可视化展示,可视化展示可以帮助用户清晰直观地看到供需关系,用户点击智能招聘系统,可以按照其所在城市、掌握的技术、工作年限等条件,快速找到符合自己需求的招聘信息,为广大求职者提供了一个方便、准确、快捷的智能岗位分析平台。

目前，系统智能化部分采用的协同过滤算法是经典传统的算法，虽然在解决未知评分的问题上做了相应优化，但是预测评分和精度方面还有待提升，后续会尝试采用更多的协同过滤改进算法进行持续迭代，提高算法的精度。同时，会开发更多的智能化功能，不断满足求职者对招聘信息数据分析方面的需求。

参考文献(References)

[1] 叶雨婷. 名校毕业生去哪儿就业？[J]. 现代青年，2022(3)：38-40.

[2] 杜鹏辉，仇继扬，彭书涛，等. 基于 Scrapy 的网络爬虫的设计与实现[J]. 电子设计工程，2019，27(22)：120-123，132.

[3] 蔡春花，黄思远，高继梅. 基于 Hadoop 的学习行为数据云存储平台的设计与实现[J]. 软件工程，2022，25(10)：50-53，49.

[4] FANG J，LI B C，GAO M X. Collaborative filtering recommendation algorithm based on deep neural network fusion[J]. International Journal of Sensor Networks，2020，34(2)：71.

[5] WU Y T，ZHANG X M，YU H，et al. Collaborative filtering recommendation algorithm based on user fuzzy similarity[J]. Intelligent Data Analysis，2017，21(2)：311-327.

[6] 王英博，韩国森，王铭泽. 基于子空间聚类的协同过滤推荐算法[J]. 计算机工程与应用，2022，58(3)：127-134.

[7] 金泓. 基于情境学习理论的计算机软件学习研究：以 Axure RP 快速原型设计工具为例[J]. 软件导刊(教育技术)，2017，16(5)：18-19.

[8] 王子毅，张春海. 基于 ECharts 的数据可视化分析组件设计实现[J]. 微型机与应用，2016，35(14)：46-48，51.

[9] 百度. 数据可视化 Suger BI[EB/OL]. (2018-04-12)[2022-12-15]. https://cloud.baidu.com/product/sugar.html.

[10] 王根龙. 基于用户可信度的抗攻击协同过滤算法的研究与应用[D]. 重庆：重庆大学，2014.

作者简介：

陈　亮(1987-)，男，硕士，讲师。研究领域：大数据处理与分析。

(上接第 14 页)

4　结论(Conclusion)

为解决溢油检测过程中油膜标签数据较少的问题，本文采用 mRMR 提取油膜标签的有效特征，然后在半监督决策树学习模型中引入自适应置信度，采用基于模糊聚类的方法衡量样本预测的置信程度，最终获得具有较好泛化能力的分类器。在不同的标签样本比例下分别进行分类实验的结果表明，采用自适应置信度的半监督决策树模型能够有效地提高油膜识别准确率；在标签样本比例较低时，模型的提升效果更为明显。

参考文献(References)

[1] DELPECHE-ELLMANN N C，SOOMERE T. Investigating the Marine Protected Areas most at risk of current-driven pollution in the Gulf of Finland，the Baltic Sea，using a Lagrange transport model[J]. Marine Pollution Bulletin，2013，67(1/2)：121-129.

[2] AL-RUZOUQ R，GIBRIL M B A，SHANABLEH A，et al. Sensors，features，and machine learning for oil spill detection and monitoring：a review[J]. Remote Sensing，2020，12(20)：3338.

[3] ALVES T M，KOKINOU E，ZODIATIS G，et al. Multidisciplinary oil spill modeling to protect coastal communities and the environment of the Eastern Mediterranean Sea[J]. Scientific Reports，2016，6：36882.

[4] XU J，WANG H X，CUI C，et al. Oil spill monitoring of shipborne radar image features using SVM and local adaptive threshold[J]. Algorithms，2020，13(3)：69.

[5] 陈彦彤，李雨阳，吕石立，等. 基于深度语义分割的多源遥感图像海面溢油监测[J]. 光学精密工程，2020，28(5)：1165-1176.

[6] FINGAS M，BROWN C E. A review of oil spill remote sensing[J]. Sensors，2017，18(1)：91-98.

[7] XU J，JIA B Z，PAN X X，et al. Hydrographic data inspection and disaster monitoring using shipborne radar small range images with electronic navigation chart[J]. PeerJ Computer Science，2020，6：e290.

[8] SUN X F，LIN X G，SHEN S H. High-resolution remote sensing data classification over urban areas using random forest ensemble and fully connected conditional random field[J]. ISPRS International Journal of Geo-Information，2017，6(8)：245.

[9] 周慧，陈澎. 利用最小冗余最大相关和 SVM 的 SAR 图像海上溢油识别[J]. 电讯技术，2018，58(8)：895-899.

[10] TEMITOPE YEKEEN S，BALOGUN A L，WAN YUSOF K B. A novel deep learning instance segmentation model for automated marine oil spill detection[J]. ISPRS Journal of Photogrammetry and Remote Sensing，2020，167：190-200.

作者简介：

刘宏宇(2002-)，男，本科生。研究领域：大数据，人工智能。

周　慧(1983-)，女，硕士，教授。研究领域：计算机视觉，深度学习。