

基于大数据分析的职业画像与求职者画像研究

叶丽娜, 刘明良, 袁超, 赵宣

汉江师范学院, 数学与计算机科学学院, 湖北 十堰

收稿日期: 2023年4月5日; 录用日期: 2023年5月3日; 发布日期: 2023年5月10日

摘要

为推动求职创业服务管理工作提质增效, 推动普通高等学校学生更加全面更高效地就业机会, 我们利用大数据分析的前沿技术, 对公司的招聘人才数据分析的情况实现了智能匹配的数据分析, 一方面能让公司更为清楚地了解到目前国内人才市场的供需情况, 一方面也能更加合理地帮助求职者正确的找到工作与就业机会。本软件采用神经网络爬虫技术, 爬取智联招聘、51job等人力资源网站上所有相应岗位的人才信息, 并抽取出当中的关键数据, 包括但不限于岗位姓名、工资待遇、岗位说明、企业简介、企业发展、企业性质等重要资料。通过对上述数据的挖掘研究, 能够较为准确、清晰地指出应聘者所在专业的薪酬标准、企业可能的福利、以及对企业、专业的选择。通过我们平台后台的计算, 将求职者的求职信息简历信息和公司递交的招聘信息进行特征提取等, 使信息数据化, 方便比较匹配。

关键词

数据爬取, 岗位画像, 求职者画像

Research on Career Portrait and Job Seeker Portrait Based on Big Data Analysis

Lina Ye, Mingliang Liu, Chao Yuan, Xuan Zhao

School of Mathematics and Computer Science, Hanjiang Normal University, Shiyan Hubei

Received: Apr. 5th, 2023; accepted: May 3rd, 2023; published: May 10th, 2023

Abstract

In order to promote the business service management work mass transfer efficiency, promote ordinary higher school students more comprehensive more efficient job opportunities, we use the forefront of big data analysis technology, the company's recruitment data analysis realized the intelligent matching data analysis, on the one hand, can let the company more clear understanding

to the current domestic talent market supply and demand situation, on the one hand also can be more reasonable to help job seekers find the right job and employment opportunities. The software adopts neural network crawler technology to obtain the talent information of all corresponding positions on human resources websites such as Zhaopin and 51 job, and extract the key data, including but not limited to the job name, salary, job description, enterprise profile, enterprise development, enterprise nature and other important information. Through the mining research of the above data, it can accurately and clearly point out the salary standard of the major, the possible benefits of the enterprise, and the choice of the enterprise and the major. Through the background calculation of our platform, the resume information of the job seekers and the recruitment information submitted by the company are extracted, so as to make the information digital and convenient for matching.

Keywords

Data Crawl, Job Portrait, Job Portrait

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 前言

大数据时代背景下，网上平台求职成为主流就业方式之一，通过大数据与前沿科技，对企业内部招聘人员信息的情况进行了智能匹配和统计分析，一方面可使企业更加明确地认识到目前中国就业市场的供求状况，另一方面又能更有效地协助应聘者合理地选择工作和就业。

2. 项目背景

2.1. 业务背景

1) 国家宏观产业发展政策趋利

中共中央、国务院办公厅一直重视大学毕业生就业工作。习近平总书记中央总书记多次就办好大学毕业生就业工作，做出了重大指导批示。国务院政府办公室《“十四五”求职提升计划》明确提出目标，继续抓好高校毕业生就业管理工作。2022 届一般高校毕业生面积、数量创历史发展高峰，求职情况复杂多变艰巨。为深刻地传达党的十九大和十九届二中、三中、四中、五中、六中全会精神，执行党中央、国务院政府办公室决策部署，教务处确定开展“2022 届我国一般高校毕业生就业创业促进行为”，继续规范求职创新推动制度，推动求职创业服务管理工作提质增效，推动普通高等学校学生更加全面更高效地就业机会。

党的十九届五中会议报告明确，要完善就业优先政策措施，千方百计地保持和增加就业机会，继续健全重点人群就业保障体制。积极推动普通高等学校毕业生就业，是就业工作的重要关键。国家教育部已将大学毕业生的就业工作列为中央重要高等教育决策部署督察、各省政府实施高等教育工作评估、大学学科专业评审、地方高等学校领导班子年度考评等工作重点内容。为推动青年大学生就业，我国政府在制度上、策略举措上推出了许多具有突破性的重大改革举措，并获得了相当的正面效应。但怎样继续疏导和拓展大学生就业渠道，是目前需要人们继续思索和研究的问题。

随着大数据分析的越来越完善，大数据分析技术不但可以让人类更方便的实现数据分析的收集，更

可以针对应用需要选择数据分析手段,帮助公司创造更大的商业价值。大数据分析 with “云计算”如同一个钱币的正反向一般慢慢勾勒出当今世界的财富发展方向。互联网的产生是因为网络信息技术的高速发展、计算机硬件与软实力的日益增强。随着数字化社会的发展,在“互联网+”语境下数据信息的传播途径将和以往有所不同,不再依赖于传统的电话、短信或书信等模式,而转变更迭智能化终端实现。目前,各种 App、公众号、订阅号层出不穷,高校就业管理模式也有待加深,想要实现大学生就业的双向选择,提高大学生就业率,关键在于用人单位能及时发布就业信息,学生及时获得就业信息,从而使得高校人才输出得到根本保障,人才输出得到了根本保障。因此,依靠“互联网+”技术,建设高校、便捷的大学生智慧就业服务平台,拓宽就业市场领域,已经成为“互联网+”大数据时代的必然趋势。

2) 规制环境不断优化

全国认真贯彻落实党中央的每一次全会精神,紧紧抓住新一轮经济技术变革与产业转型的重大契机,以基础创新和技术方法创新为导向,全面推动新型信息技术和三次产业的结合创新,积极运用“互联网+”推动新科技、新产业、新业态和新方式的融合发展,为全省加速推动经济转型升级带来了强劲动能。

2.2. 技术因素

1) 多领域技术研发比较优势明显

相比较之前的农业与传统制造业,网络企业显示出其巨大的发展机制与成长动力。伴随手机网站、虚拟现实、新一代人工智能、物联网等网络新科技、新模式的不断诞生,从搜索引擎、社区、旅游、电商再到 P2P、O2O、B2B、C2C 的形态发展,以及全新的行业格局、商业形式和新商业模式的产生,中国网络经济步入了全新的发展阶段。

2) 运营商酝酿战略转型

随着 5G 网络的加快部署,移动、电信、联通等运营商都在酝酿实施以数据业务为核心的战略转型,包括优化升级网络承载能力、改进用户体验、强化用户服务等。

3. 问题分析

3.1. 运行模式

目前,中国的互联网招聘市场根据运营方式可以分成综合类招聘网站、地域性信息招聘网站、分类资讯招聘网站、人才查询招聘网站、专业分类招聘网站和社区招聘平台。就全世界范围来说,大数据领域已经进入了硅谷发展曲线的下滑轨道上,即由技术热潮的顶峰期滑落,逐渐进入到了市场战略布局的“低调期”。不过在我国,随着大数据领域的技术发展条件逐步完善,在行业仍然保持着高速发展的态势下,大数据领域概念也得到了投资界的不断重视,针对大数据行业的新资本运作频频涌现。

综合分析,该项目具有可行性。

3.2. 可行性分析

3.2.1. 技术可行性

1) 该系统使用了 Hadoop 分布式文件系统、基于 Hadoop 的数据仓库系统 hive 和大型数据分析的引擎系统 Spark, HDFS 可以进行大吞吐量的数据存取,非常适合于大型数据分析这一需求;

2) Hive 提供了大量工具,用来保存、检索和分析存放在 Hadoop 中的大量数据;

3) Spark 是一种通用引擎,可以用于进行包括:SQL 搜索、文本处理、以及机器学习等的操作,它具有更快的计算速度、高易用性以及通用性。

这三种工具都已经成功地应用到大数据的处理当中,其知识和技术体系已经比较完善,且有成熟的

案例，学习起来比较容易；此外，本系统所用的爬虫技术与数据分析已有十分成熟的知识体系以及参考范例，便于学习，所以在技术上是可行的。

3.2.2. 经济可行性

由于系统开发时所用的 IDE、开发语言及其部分方法都是相对开放的，再加上采用的开发工具有高度通用性，对硬件要求也不高，从而在整体开发的成本方面节省了不少资金，所以在经济上还是比较可以的。经济效益：促进健全中国电商领域的信用与安全管理体系，形成一个相对完善的安全可信的网上商品交易用户利益保障制度。从而规范我国电子商务市场，提高效率的同时创造更多的财富。费用及投入分析：此活动由企业提出相应的立项资金，调查主要是以调研问卷及访谈方式，所以费用较低。

3.2.3. 操作可行性

本系统后台数据不需要用户进行操作，用户所接触到的都是已经经过可视化的数据分析成果和最终推荐展示，用户只需要一台能够联网的电脑就能对系统进行操作，所以对用户而言，系统在操作上是可行的。

3.3. 用户分析

3.3.1. 在求职准备阶段

求职者的主要活动包括：

1) 逐步掌握了相关职业选择与职业生涯规划等方面的基本知识，并开始实际地考虑职业选择与职业发展的问題；

2) 早期的职业活动，如学习、打工、社会实践等，人们从这种活动中可以获得对职业生活的感性认识；

3) 通过进行的各类专业技能训练或试验，获得了教育或职业能力等方面的考试合格证明；

4) 掌握一定就业方面的理论知识与经验，尝试撰写就业申请书。

就目前应聘中存在的问题而言，主要体现在如下方面：

一是存在着一定程度的自卑心理许多毕业生在面临职业选择时，都或多或少存在着自卑的情况。表现在以下方面，即一是在对方招聘而自己去应聘考试时，担心自己不符合要求；

二是在对方招聘的岗位有众多竞争者的情况下，担心自己不能取胜；

三是在对方没有招聘而自己主动去求职的情况下，担心对方对自己不感兴趣。

3.3.2. 在应聘筛选阶段

求职者的主要活动包括：

1) 积极搜集并利用偶然获得的人才信息，并对这些信息加以研究，确定能否应聘；

2) 通过相应的方式投寄个人简历和求职申请书；

3) 参与用人单位所要求的测试筛选；

4) 取得用人单位的录取通知书并与用人单位达成就业合同。

3.3.3. 在进入组织阶段

求职者的主要活动包括：

1) 根据雇佣合同或其他协议规定，向用人单位报到，并处理报到的有关事宜；

2) 订立劳动合同，确定工作岗位和薪资待遇等；

3) 进行技术培训，适应组织环境，参加或担任管理工作；

4) 进行试用期评估，直到最后通过考核。从毕业生就业的角度，只要顺利通过了试用期审核，其就

业程序才结束。

4. 解决方案

4.1. 业务分析

拟解决的问题有：

- 1) 应聘者常面临的问题：对自己专业能力与具体的职业匹配有误区；就业准备形式性；对求职意向不清楚；对公司的了解甚少；对简历投递随意性过大；对自身没有足够的了解(没信心或者过度自信)等。
- 2) 企业招聘存在的问题：向大学生传递的招聘信息不准确；获得应聘者的信息不全面；大学生短期离职率较高；很难选出优秀且符合标准的人才；面试程序复杂等。

拟解决的关键问题、方案与途径、实现目标如表 1 业务分析表所示：

Table 1. Business analysis table

表 1. 业务分析表

关键问题	方案与途径	实现目标
对自身专业及相应的岗位匹配有误解	我们提供自身画像与岗位画像匹配的算法	实现求职者更加了解自己的岗位的目标
求职准备形式化，求职意向不明确	我们提出了对自身简历的分析，为目标用户提出职位建议	实现求职者明确自己的求职意向的目标
企业向大学生传递招聘信息不准确	提供了官方正式的平台对招聘信息进行校准	实现双方对招聘信息的准确认知的目标
企业面试过程复杂	提供了真实信息的核对匹配，让 HR 提前对招聘者的全方面信息进行了了解	实现减少不必要的认知过程的目标

验证方法：功能演示，专家评审，实际应用
为求职者提供详细的企业岗位画像(包括岗位要求，待遇等)

4.2. 解决思路

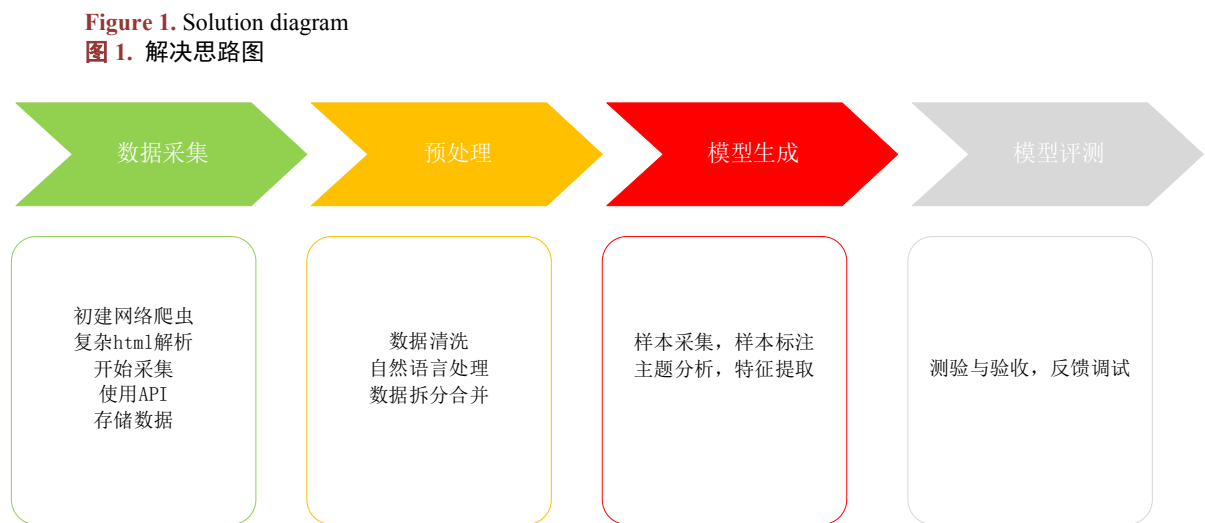
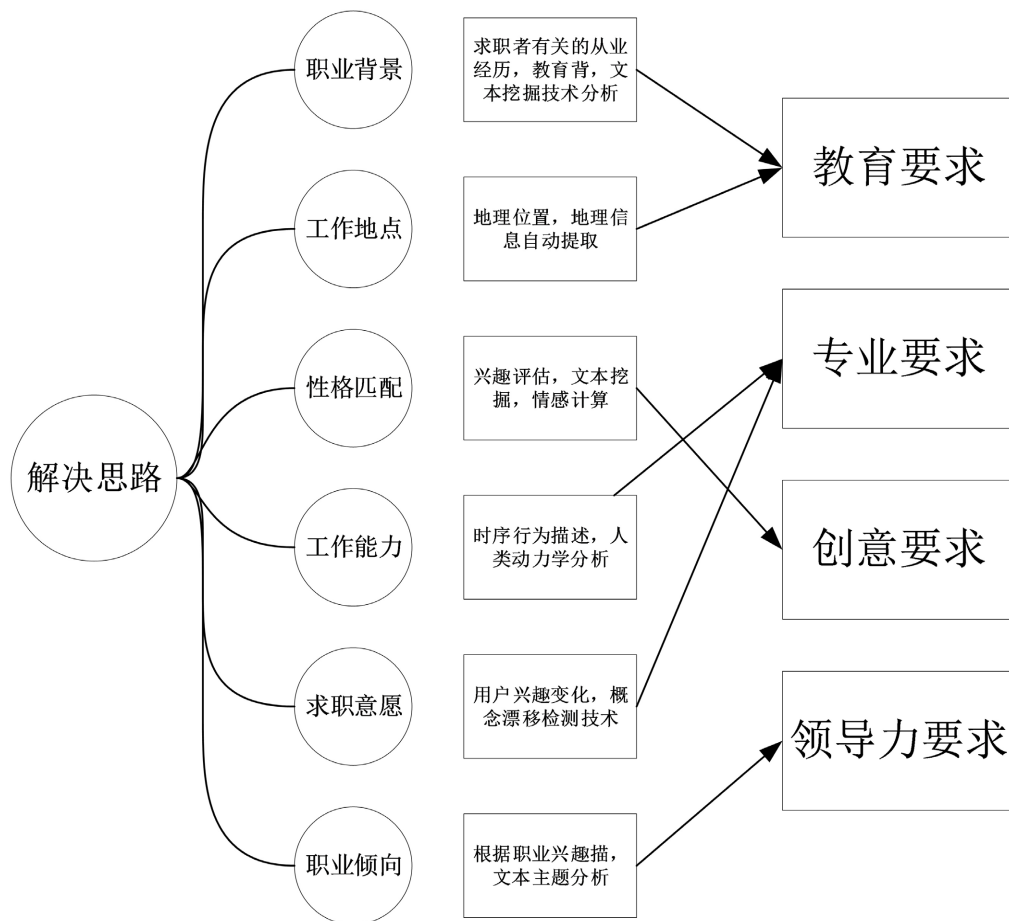
首先需要通过网络爬虫爬取智联招聘、51job 等招聘网站上，大数据相关职位的招聘信息，提取出其中的关键数据，包括但不限于职位名称、职位待遇、职位描述、公司介绍、公司规模、公司性质等信息。通过对这些信息的挖掘分析，能够比较精确、清晰地指导应聘者所在行业的薪酬水准、目标企业可能的薪酬水平及其对企业、行业的投资选择。同时运用爬虫技术对互联网上的求职者个人信息加以收集，然后运用大数据分析平台，对互联网上收集下来的资料加以分类挖掘，从而挖掘出与职业、薪酬、学历、薪酬水平等各种因素间的相互关联，从而产生更高价值资讯。解决思路如图 1 解决思路图所示。

本系统在需求分析阶段，已经分为了 4 大部分，分别是数据采集、预处理、模型生成、模型评测。所以一共分为 4 大模块，如图 2 解决过程图所示，数据采集模块、预处理模块、模型生成模块和模型评测模块，这 4 大模块依次按照顺序运行，才能保证系统有效地运行。只有数据采集模块过后，方可有数据，然后数据分析模块才能对数据进行分析，获得数据分析后的结果，最后通过数据展示模块把数据展示出来和对用户进行个性推荐[1]。

4.3. 整体方案

4.3.1. 数据采集

本系统采用分布式爬虫系统。



分布式的主从系统设置了一台 Master 客户端和多台 Slave 客户端, Master 端负责管理 Redis 数据库并分配加载工作, 由 Slave 的 Scrapy 爬虫采集网站并分析获取内容, 然后再把分析的结果存放一台 MongoDB 数据库中[2]。

首先创建一个 spider 爬虫任务，这时候会有一个入口 URL，spider 从这个入口开始爬取。

调用 Downloader 组件去执行 http 请求下载整个页面；spider 解析页面中的内容，将需要的内容放入 item 里面，同时将页面中的子 URL 放入 Scheduler 组件；Pipeline 负责将 item 中的数据就行持久化存储；URL 放入 Scheduler 组件，Scheduler 组件会对 URL 进行去重，避免重复爬取；spider 爬完当前页面后就继续从 Scheduler 拿 URL，如果有 URL 则继续爬取，没有则说明所有页面都爬完了，spider 任务结束。

4.3.2. 数据预处理

1) 自然语言处理过程：

对企业公司的招聘公告进行关键词的提取，提取出公司名称、招聘岗位、岗位要求以及薪资待遇；对于用户输入的个人信息进行关键词的提取，提取出用户名字、年龄、学历、项目经历以及掌握技术。

2) 验证数据的完整性与实效性：

验证所提取数据的完整性，是否缺少某项数据，如果有缺失，对企业公司则删除这一条数据，对用户则进行提醒，建议其填写完整；由于本系统是根据近期招聘领域的情况而对用户推荐最优选择，所以数据需要有较好的实时性，因此在使用数据库中的数据进行模型训练前，对于超过一定时间的数据，需进行删除，保证推荐模型是最优的。

3) 数据合并处理：

将相关的数据项进行合并处理，比如将用户输入信息合并。

4.3.3. 模型生成

1) 岗位画像生成：

利用聚类算法，将同一岗位下，不同公司的要求描述进行聚类处理，处理得到形如“熟练掌握”、“了解”、“育要求”、“薪资待遇”等数据项，并将其可视化。

2) 个人画像生成：

利用个人信息提取出的关键词，生成个人画像。

3) 推荐算法

利用基于内容的推荐算法，对比用户信息与企业要求，在各个对应的数据项，比如“教育水平”与“教育要求”，“专业水平”与“专业要求”等，对每一对数据进行相似度计算并进行评分，得到一个综合分数，将靠前的几个招聘公告提供给用户。总技术架构如图 3 技术架构图所示。

4.4. 核心算法

1) TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency)是信息检索与挖掘过程中使用的加权方法[3]。TF 是术语频率，IDF 是逆文档频率。

2) TextRank

TextRank 方法，是一个使用关键字提取和文字概括的，基于图的排序算法。它是由 Google 的 PageRank 算法改进的。它可以通过使用文档中单词之间的共现信息(语义)来提取关键字。它能够获取给定文本的关键词和语句，并利用自动文摘方式获得文中的重要语句。TextRank 算法的基础思路是把文本看作词汇网络，网络中的链接描述了词汇间的语义关联。

3) LDA (线性判别分析)

线性判别分析法(Linear Discriminant Analysis, LDA)，是基于 fisher 线性微分归纳法发展的，该方法通过统计、模式识别和机器学习方法，试图发现经过线性组合的二个对象或事物之间的共同特性，并表征或区别它们。所得到的组合也可以用于线性分类器，或者更普遍的是，可以用来减少后续分类的维数。

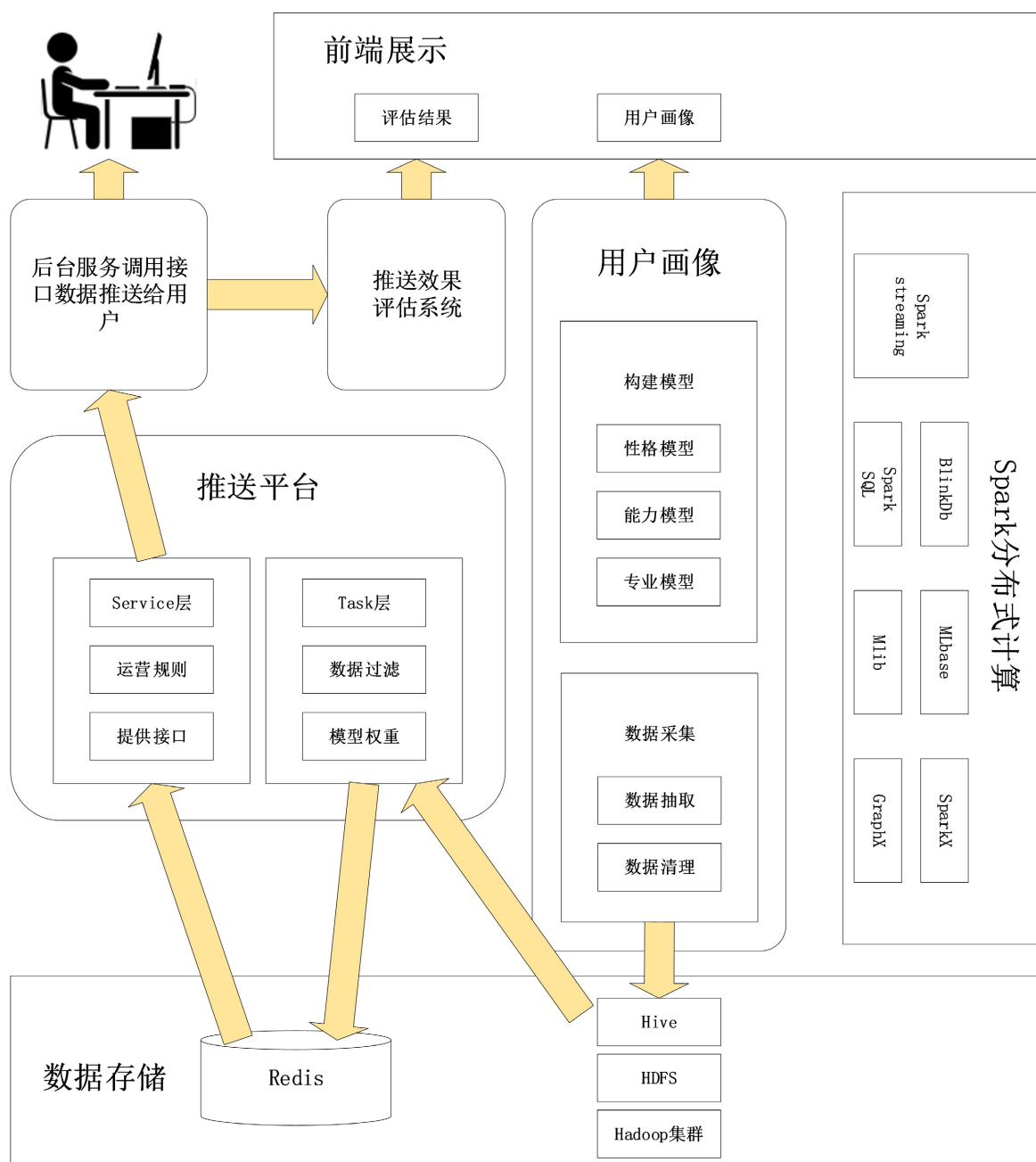


Figure 3. Technical architecture diagram

图 3. 技术架构图

4) KNN

KNN 算法存储样本数据，其中包含大量冗余数据，这些冗余数据增加了存储开销和计算成本。减少练习数据的做法是：在原有数据中剔除若干与分类不有关的数据，或把剩余的数据用作新的练习数据；或在原有训练样本集中选择几个较有意义的样品，成为新的练习样品；或经过聚类后，使聚合得到的中心点成为新的练习样品。

5) 贝叶斯网络

贝叶斯网络(Bayesian Network), 又称信念网络(Belief Network)或是有向无环图模型(Directed Acyclic Graphical Model), 是一种概率图型模型。一个贝叶斯网络是一个有向无环图(Directed Acyclic Graph, DAG), 由代表变量结点及连接这些结点有向边构成。结点代表随机变量, 结点间的有向边代表了结点间的互相关系(由父结点指向其子结点), 用条件概率进行表达关系强度, 没有父结点的用先验概率进行信息表达。

$P(A)$ 是 A 的先验几率或边界几率。之所有称作“先验”是因为它不顾及所有 B 方面的原因。 $P(B)$ 则是 B 的先验几率或边界几率。贝叶斯定理也能够描述成: 后验事件 = (相似性度 * 先验机率)/标准化常量。

比例的 $P(B|A)/P(B)$ 也有时称之为标准相似率, 而贝叶斯定理也可以表达为: 后验概率 = 标准相似率 * 先验概率。

设 $\{A_i\}$ 为事件集里的部分集, 对任意的 A_i , 则贝叶斯定理可以用下图图 4 贝叶斯定理说明公式说明:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

Figure 4. Bayes' theorem illustrates the formula

图 4. 贝叶斯定理说明公式

6) 神经网络算法

BP 神经网络是一种多层次的前馈神经网络, 其基本逻辑思想如图 5 神经网络算法的逻辑思想所示, 其特征为: 信息为正向传输, 而错误则为反面传输[4]。BP 神经网络的发展过程, 大致包括了二个阶段。第一步是将信息从进入层通过隐蔽层, 并终于到达出口层的方向传送。第二步则是信息错误的正反向传输。从出口层到隐蔽层, 终于到达入口层, 然后顺序调整从隐蔽层至出口层的重量与偏置, 并最后分开调整从入口层至隐蔽层的重量与偏置。

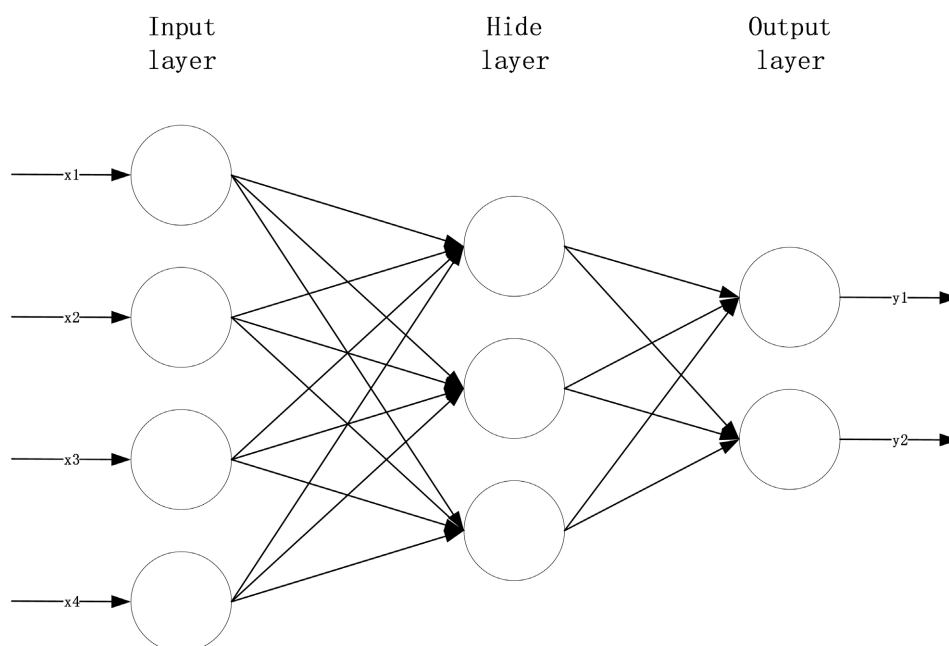


Figure 5. Logical idea of neural network algorithm

图 5. 神经网络算法的逻辑思想

5. 创意亮点

5.1. 求职工作流程完整

在求职者来说, 填写简历, 浏览岗位要求, 与岗位进行匹配, 投递简历的以一系列流程十分完整, 简化了求职者步入公司的过程。能够让求职者感受应聘的相应流程, 做好心理准备。

5.2. 匹配模型灵活

我们提供了灰色关联评估模型, 测算模型, 冰山理论模型, 洋葱理论模型等等, 多种模型的终点不一样, 也为公司选择不同类型的员工提供灵活的选择。求职者在查看多种模型生成的画像之后也能够对自己有一个更清晰的认识, 方便后续的应聘。

5.3. 信息数据化可视化

通过我们平台后台的计算, 将求职者的求职信息简历信息和公司递交的招聘信息进行特征提取等, 使信息数据化, 方便比较匹配。使用生物信息学技术来实现对复杂数据的获取、分类与可视化, 并在大数据分析中获取科学价值。

5.4. 提供智能化的建议

我们的后台进行对比个人能力与岗位要求的对比, 找出优点以及劣势, 给企业录取特殊技能人才提供方便, 给求职人员提出应聘该岗位的建议。

使用大数据分析, 通过将传统系统迁移到新供应商体系、并购以及将核心系统升级, 来改善顾客体验。通过自助服务的灵活性分析数据, 可以迅速掌握市场领先态势。并通过大数据分析来更好地认识顾客, 通过降低顾客流失来减少收入损失的风险。

6. 结论

文章总结并阐述了消费者画像的概念, 剖析了标签系统构建的基本过程以及统计分析方式。本文的工作主要围绕着基于大数据分析的职业画像与求职者画像研究这个系统平台而展开的, 这个系统平台针对各种用户的求职信息, 把这些数据从网上爬去下来上传到 Hadoop 集群上进行 Spark 分析, 使海量数据标签化, 帮助企业更精准解决问题。

基金项目

汉江师范学院 2022 年大学生创新创业省级项目: 基于大数据技术的岗位画像与求职者画像设计。

参考文献

- [1] 武继飞. 基于大数据技术的岗位画像和求职者画像设计[D]: [硕士学位论文]. 金华: 浙江师范大学, 2019.
- [2] 张民忠. Apache Mina 中 TCP 服务器集群功能扩展研究[D]: [硕士学位论文]. 武汉: 武汉理工大学, 2018.
- [3] 余江维, 余泉, 张太珍, 等. 基于 TF-IDF 相对熵的中医证候量化研究[J]. 世界科学技术-中医药现代化, 2015, 17(10): 1986-1991. <https://doi.org/10.11842/wst.2015.10.005>
- [4] 沈泽君. 粒计算思维在神经网络上的应用[D]: [硕士学位论文]. 漳州: 闽南师范大学, 2020.