基于大数据技术的求职用户画像系统研究与设计

□李锦锐　章家宝　彭　梅

【内容摘要】随着大数据时代的开启，利用大数据技术进行数据收集和分析，并根据需求建立模型，从而进行商业的数据分析与运营获取更多商用价值，于是“用户画像”应运而生。本文介绍了用户画像的理论研究，利用大数据中的爬虫技术对网站上的求职信息进行收集，再利用大数据平台，对网上收集下来的数据进行分析挖掘形成高价值信息。

【关键词】大数据；求职用户画像；爬虫技术

【基金项目】本文为 2018 年度广州工商学院国家级大学生创新创业训练项目“基于大数据技术的岗位画像和求职者画像设计”（编号：201813714001）研究成果。

【作者单位】李锦锐，章家宝，彭梅；广州工商学院

大数据和“云计算”像是一枚硬币的正反面一样慢慢勾勒出当今世界的财富价值风向。大数据的出现得益于互联网行业的快速发展、计算机硬件和软件能力的不断提升。随着大数据技术逐渐成熟，大数据技术不但可以使人们很容易地进行数据的获取，还可以根据应用需求采用数据分析方法，为企业创建更多的商用价值。针对大数据的应用本文接下来谈谈大数据如何在人力资源管理中的在线招聘领域，利用大数据技术等前沿技术对企业招聘人才的需求进行智能分析，一方面能让企业更好地了解到目前求职市场的供需情况，另一方面可更好地帮助求职者理性择业。

一、求职用户画像在招聘行业现状

“用户画像”又称用户角色，就是利用大数据技术收集用户数据，然后勾勒用户需求、用户偏好的数据分析方法。在战略层面来看用户画像可以帮助企业进行市场洞察、预估市场规模，从而辅助制定阶段性目标，知道重大决策，提升 ROI；更有助于避免同质化，进行个性化营销。从产品层面来看用户画像可以围绕产品进行人群细分，确定产品的核心人群，从而有助于确定产品定位，优化产品的功能点。从数据分析层面来看用户画像有助于建立数据资产，挖掘数据的价值，使数据分析更为精确，甚至进行数据交易，促进数据流通。而求职用户画像早期相对简单，相当于用户信息档案，传统的网站招聘就是让求职用户通过上传简历至招聘网站中，企业用户的人力资源部门在海量的个人简历中搜寻自己所需要的人才，劳动力需求方与供给方通过互联网实现对接。但随着新技术发展使得求职用户的数据高度结构化比如职业类型、从事行业简历呈现方式等，但现实生活中随着互联网技术兴起，新兴行业越来越多，从业者所在行业或职业类型可能会更加精细，招聘网站即使拥有多个数据属性也无法清晰定义用户能力与经验。

在国内随着大数据从理论层面渐渐投入应用层面，大数据挖掘及分析技术逐渐成熟，通过构建求职用户数据分析模型，整合求职用户网络行为习惯及社交网络的数据，能形成其更加精准立体的求职用户画像。“求职用户画像”是大数据应用的基础，是招聘企业基于求职用户的构成、社会属性、地域分布等信息，提炼高精度的求职用户特征标识，通过强大的大数据分析处理和机器学习等能力，构建精准、多维的标签化求职用户模型。

二、构建求职用户画像场

（一）求职用户画像构建流程。主要通过：数据采集、数据清洗、数据标准化、用户建模、标签挖掘、标签验证、数据可视化等 7 个步骤来实现。

1. 数据采集。数据采集的数据来源有很多，常用的有以下几类：产品的数据后台、问卷调查、网络访问记录等以及用户行为、用户访谈等。

2. 数据清洗。采集回来的数据存在很多“脏数据”包括数据重复、空缺、错误、不一致等问题，为了保证数据的准确性，避免对标签挖掘及决策的影响，需要对收集来的原始数据进行清洗。

线问诊、转诊或者会诊时，将相关数据封装为区块，经过加密和安全认证后上链。由于医疗数据量庞大，可以将数据主索引封装上链。再通过索引查找数据的详细内容。进一步提高了传输速率和安全性。

五、安全隐私保障

医疗健康信息是个人的隐私信息，在医疗健康行业中使用区块链技术可以满足数据的安全隐私的保障。例如，数据加密、数字签名、环签名和混淆者模式等。

在医疗健康业务中，建议上链内容仅包含索引信息而不是具体的患者数据，避免隐私数据泄漏。同时各诊疗系统信息中间层实现访问认证，降低非法获取患者隐私数据的机率。

【参考文献】

[1] 杨保华，陈昌. 区块链原理、设计与应用 [M]. 北京：机械工业出版社，2017

[2]（美）Roger，Wattenhofer，陈晋川等. 区块链核心算法解析 [M]. 北京：电子工业出版社，2017

3. 数据标准化。求职用户画像的建立需要有跨媒介整合多源数据的能力，建立统一标准才能完整标识实体的用户画像。

4. 用户建模。对求职用户资料进行分析和加工，提炼关键要素，使用数据分析上的算法构建可视化模型。

5. 标签挖掘。通过部署环境平台来进行标签的加工和计算，比如通过大数据中的 Hadoop 平台进行数据的加工和并行计算。

6. 标签验证。通过真实的 case 验证标签挖掘结果的正确性，达到保证标签对应的处理结果跟预期大体相符。

7. 数据可视化。使用多维分析工具使视觉呈现群体或个人的求职用户画像包括柱状图、饼图、折线图，圆环图、表格等。

（二）用户画像数据信息标识。大数据时代，大数据挖掘的数据信息是非常多的，需要通过数据分析技术提炼求职用户特征标识的数据，求职用户画像主要数据信息标识有如图 1 所示。

1. 求职用户层次。中高端人才：寻求更广阔的发展空间，有明确的薪资要求；白领：追求高效率以及更好的用户体验，多样性需求；蓝领：互联网化程度不高、流动性大，关注信息是否透明、更新是否及时；大学生：短期兼职、实习、应届求职。希望获得简历书写、测评等指导。

2. 地域分布。基于互联网的求职者分布区域，可以显示不同年龄段求职用户对全国一线、二线、三线城市求职地域需求，同时便于企业需求方进行精确的地域招聘及投放招聘广告。

3. 用户学历。用户学历即学历、专业、毕业学校、继续教育等标签。

4. 性别。性别即男女性别标签，通过投放电子简历判读不同性别对不同职业的偏好信息判断。

5. 薪资需求。薪资需求即职位偏好、公司偏好等信息。

6. 行为特点。行为特点即点击求职应用次数标签，可以结合用户的浏览或者下单行为，结合用户的活跃留存数据获得。

7. 爱好兴趣。爱好兴趣即休闲娱乐、学习的偏好等标签。

8. 网络社交。网络社交即注册登录的社交网络平台、公司交流网络平台等标签。

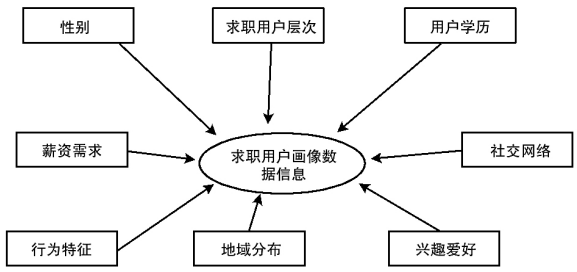

图 1　求职用户画像数据信息图

三、用户画像设计技术实现及流程

用户画像设计就是利用有效时间期限内的求职用户行为和内容建立一个临时用户画像，并且使该临时求职用户画像从求职用户画像当中继承与有效时间期限内的求职用户行为和内容相匹配的描述性标签属性，而当有效时间期限内的求职用户行为和内容与求职用户画像的描述性标签属性不匹配时，则在临时用户画像中新建描述性标签属性。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，可以更加精准、清晰地指导求职者所在行业的待遇水平、自身可能的待遇以及对公司、行业的选择。通过爬虫技术对网站上的求职信息进行收集，再利用大数据平台，对网上收集下来的数据进行分析挖掘，挖掘出岗位、工资、学历、待遇等不同因素之间的关系，形成高价值信息。具体流程如下。

（一）网络爬虫首先根据选定的网站列表进行爬取。为提高爬取效率，整个模块应支持爬虫的水平扩展，并且可基于开源系统实现。场景举例：系统管理员可以对网站列表进行增加、删除、修改操作，可以设置开始爬取的时间，爬取的频率，设置完毕后，网络爬虫根据指定的条件进行爬取。爬虫支持深度优先或广度优先策略，要求提供自研算法。

（二）对爬取的数据进行结构化处理和分析。对于爬去的数据支持丰富的解析能力，要求提供优质挖掘算法或解析规则。在对爬取的数据进行结构化处理的基础上，要求分析：一是岗位工资的影响因素；二是岗位能力需求图谱；三是招聘企业因素等三大主题，为下个环节的智能推荐做准备。

（三）系统根据求职者画像作出智能分析并进行匹配。当求职者输入学历、专业、学校、求职地、工作年限、技能、岗位名称等基本信息后，系统将智能分析出该职位的待遇水平、求职者的待遇区间、可能去的公司、公司性质和规模、行业、匹配概率等信息。要求提供：求职者画像及岗位个性化推荐算法。

四、求职用户画像注意事项

（一）结合业务。在设计求职用户画像时候一定要与具体的业务场景或所属的行业相结合，这样可以避免太过抽象，同场景下标签名称可能表示不同意思，例如性别分为真实生理性别以及网络虚拟性别，需区别对待。

（二）控制粒数。画像的粒度不是越细越好，划分的标签不是越多越好，划分越多，对应的覆盖人群就就越少，表征能力弱就有可能是伪特征。

（三）动态变化。不能盲目使用用户画像，用户画像多为静态特征，用户的特征随着时间动态变化，也可能随着场景空间的不同而不同。当然有动态的用户画像信息，比如用户访问的路径、访问时间长短等。

五、结语

在大数据、云计算及人工智能时代，利用大数据技术对数据进行挖掘和分析并通过智能筛选清洗出有商用价格的数据在现代社会变得尤其重要，大数据势能也将从平台级企业向更多细分垂直领域释放。

【参考文献】

[1] 一步一步教你看懂大数据时代下的“用户画像”[EB/OL]. 中国大数据产业观察网，2016－4－16

[2] 郝胜宇，陈静仁. 大数据时代用户画像助力企业实现精准化营销[J]. 中国集体经济，2016，4：61～62