doi: 10.7690/bgzdh.2023.05.011

基于决策树算法的 IT 专业就业模型

李　川，刘洲洲
（西安航空学院计算机学院，西安 710077）

摘要：针对当前高校毕业生就业困难、人才培养方案不能适用社会需求等问题，对基于决策树算法的 IT 专业就业岗位和知识需求聚类模型进行分析。基于大数据技术，以 IT 专业学生就业及社会招聘信息为例，通过爬虫技术进行数据挖掘，并对数据进行爬取、清洗、存储等操作，获取包括岗位、工作地点、薪资、学历、工作经验、知识技能等属性的数据集合。根据特征集合及数据集合，采用 ID3 算法，以最大信息增益为目标，构建基于决策树算法的 IT 专业就业岗位和知识需求聚类模型。试验结果表明，该模型可促进 IT 人才资源科学管理与决策、解决高校 IT 专业毕业生就业难、提高高校 IT 人才培养水平。

关键词：IT；决策树；ID3；爬虫；数据清洗
中图分类号：TP312　　**文献标志码：**A

Employment Model of IT Majors Based on Decision Tree Algorithm

Li Chuan, Liu Zhouzhou
(*College of Computing*, *Xi'an Aeronautical Institute*, *Xi'an* 710077, *China*)

Abstract: In view of the current employment difficulties of college graduates and the fact that the talent training program can not meet the social needs, this paper analyzes the clustering model of IT professional employment posts and knowledge needs based on decision tree algorithm. Based on big data technology, taking the employment and social recruitment information of IT majors as an example, data mining is carried out through crawler technology, and the data is crawled, cleaned and stored to obtain data sets including job, work place, salary, education, work experience, knowledge and skills. According to the feature set and data set, the ID3 algorithm is used to construct the clustering model of IT professional employment and knowledge demand based on decision tree algorithm with the goal of maximizing information gain. The experimental results show that the model can promote the scientific management and decision-making of IT talent resources, solve the employment difficulties of IT graduates in colleges and universities, and improve the level of IT talent training in colleges and universities.

Keywords: IT; decision tree; ID3; crawler; data cleaning

0　引言

当今社会已经进入大数据时代[1]，大数据技术在不断地改变人们日常生活、工作、学习等，也为个人信用、疫情防控等人类生存与发展提供了方便快捷的服务[2-4]。近年来，高校毕业生的就业形势日益严峻，毕业生与社会需求很难匹配，这已成为一个社会关注的焦点问题[5]。当前，很多招聘求职网站为企业招聘、大学生就业提供了信息服务，但是招聘网站中的信息服务很多都无法精准满足求职人及招聘者所需，主要体现在 2 点：1) 招聘信息量大，求职者很难有效从中获取适合自己信息；2) 没有准确感知并描述劳动力市场的需求，很多高校毕业生的知识技能与需求脱节，达不到用人单位需求，导致就业效率、就业质量低下。

为准确感知并描述社会对人才的需求，笔者以 IT 专业为例，基于大数据技术，通过爬虫技术对 IT 专业海量招聘信息进行挖掘和分析，通过决策树算法对 IT 专业就业岗位和知识需求进行聚类分析，为基于大数据分析 IT 社会需求和人才知识能力的差异性提供了一个框架平台。

1　大数据技术研究

当今社会，大数据技术的应用已经渗透到各行各业，对就业大数据的分析必须研究以下内容。

1.1　大数据采集

大数据采集即获取各种结构化和非结构化海量数据，数据采集的方式可分为 3 种：

1) 数据库采集：可通过分布式数据库采集大数据，常见的有 Sqoop 和 ETL；也可通过 SQLServer、Oracle、MySql 等传统关系型数据库采集[6]。

收稿日期：2023-01-14；**修回日期：**2023-02-18
基金项目：国家自然科学基金(61871313)；西安航空学院 2021 年校级新工科研究与实践项目(21XGK2001)
作者简介：李　川(1980—)，男，陕西人，硕士，副教授，从事软件算法理论、大数据分析研究。E-mail: Lch9964121@163.com。

2) 网络数据采集：可利用 Java 语言实现网页数据采集，也可利用 Python 实现网络爬虫从网页采集数据，爬虫是一个带宽消耗型应用，Python 是一种高效的编程技术[7]。

3) 文件采集：可利用分布式海量日志采集处理系统采集数据，也可通过实时文件采集技术采集数据。

1.2　大数据预处理

通过大数据采集获取的数据称为原始数据，其中含有很多杂质或数据本身存在缺陷；因此，要对原始数据进行预处理，提高数据质量，确保数据的可用性。一般数据预处理的内容包括数据清洗、补全、分割、合并、规格化、一致性检验等操作[8]，其过程如图 1 所示。

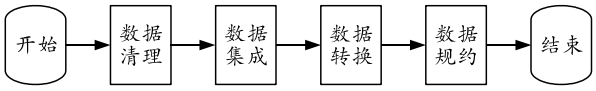

图 1　大数据预处理过程

1) 数据清理：对有缺少重要属性的残缺数据、存在重大错误或严重偏离期望值的噪音数据、存在相似但不一致的数据进行处理。

2) 数据集成：将通过不同方式或来源采集的不同数据进行合并，消除冗余，对结构相容的数据合并，并对有冲突的数据进行检测与处理。

3) 数据转换：将存在不一致的数据进行处理，使数据保存统一结构。

4) 数据规约：对数据做规范化处理，同时压缩数据。

1.3　大数据存储

将采集、预处理后的结构化数据存储到数据库的过程就是大数据处理，常见的大数据存储方式包含基于海量数据实时分析架构 MPP 的大数据存储及基于 Hadoop 的技术扩展和封装的大数据存储[9]。

1.4　大数据分析

大数据分析主要包括以下方面：

1) 可视化分析，使用大数据及图形工具实现折线图、直方图、散点图、饼图、K 线图、BI 的漏斗图、仪表板等，直观的表现海量数据的关联分析。

2) 数据挖掘，根据数据的特征，对特定类型数据的结构和趋势进行分析，创建数据挖掘模型，设计数据分析算法[10]。

3) 预测性分析，通过统计分析、人工智能、预测趋势等多种方法，分析数据中的趋势、模式和关系。

2　IT 职业需求数据爬取与治理

2.1　数据爬取流程

通过网络开放数据资源获取 IT 就业数据，基本流程如图 2 所示。

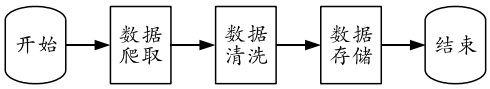

图 2　数据爬取与治理过程

2.2　数据爬取

首先对国内某招聘网站的招聘信息(岗位、工作地点、薪资、学历、工作经验、知识技能)进行爬取，数据爬取流程如图 3 所示。

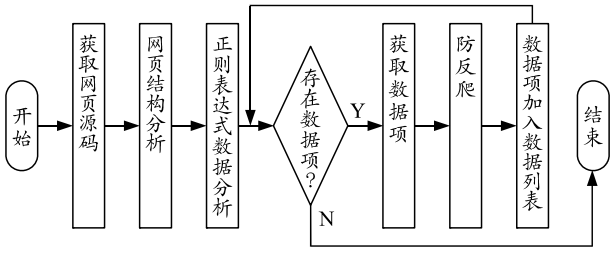

图 3　数据爬取

2.3　数据清洗

完成数据爬取后，收集到大量原始数据，这些数据来源虽然都是通过页面爬取而来，但数据形式多样，格式也不尽相同；因此，需要对数据进行治理清洗，即修正数据中的错误及统一数据的格式[11]，得到数据文件，数据清洗的流程如图 4 所示。

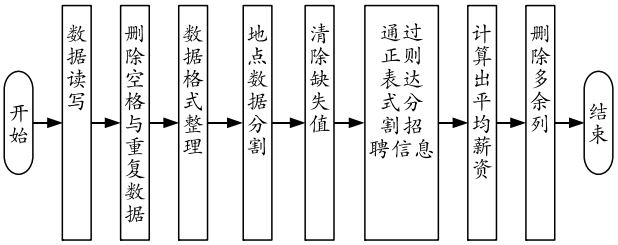

图 4　数据清洗过程

2.4　数据存储

数据存储就是将爬取、清洗后的数据写入数据库，以便对数据进行各种操作和管理，并且规范数据格式、保证数据安全。数据存储的过程如图 5 所示。

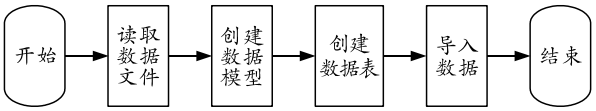

图 5　数据存储过程

3 基于决策树的 IT 岗位与技能聚类模型

3.1 初始化特征集合和数据集合

1) 学历特征集合，如表 1 所示。

表 1 学历特征

序号	学历	序号	学历
1	大专	3	硕士
2	本科	4	博士

2) 岗位特征集合，如表 2 所示。

表 2 岗位特征

序号	岗位	序号	岗位
1	系统分析师	5	测试工程师
2	前端开发工程师	6	数据库管理员
3	后端开发工程师	7	软件运维工程师
4	移动开发工程师		

3) 薪资特征集合，如表 3 所示。

表 3 薪资特征

序号	薪资	序号	薪资
1	20 000	4	8 000
2	15 000	5	5 000
3	10 000	6	4 000

4) 工作经验特征集合，如表 4 所示。

表 4 工作经验特征

序号	工作经验	序号	工作经验
1	半年	3	2 年
2	1 年	4	无

5) 知识技能特征集合，如表 5 所示。

表 5 知识技能特征

序号	知识技能	序号	知识技能
1	Python	5	PHP
2	Java	6	前端框架
3	C#	7	Android
4	C++	8	IOS

6) 工作地点特征集合，如表 6 所示。

表 6 工作地点特征

序号	工作地点	序号	工作地点
1	上海	5	广州
2	北京	6	武汉
3	深圳	7	南京
4	西安	8	杭州

数据集合是由 Python 爬虫爬取并经过数据清洗治理后的包括岗位、工作地点、薪资、学历、工作经验、知识技能等属性的数据集，共有 31 200 条数据。

3.2 基于决策树的 IT 岗位与技能聚类模型构建

决策树是一种树型数据结构，经常用于构建聚类模型，其代表的是对象属性与对象值之间的一种映射关系，树的内部节点表示特征属性上的测试，每个分支表示该特征属性的测试输出，而每个叶节点表示一个类别。决策树是一种机器学习方法，使用决策树进行分类预测，每次预测从决策树根节点出发，根据特征属性的测试值，选择相应的分支，循环该步骤，直到遍历到达叶子节点，叶子节点的值就是分类预测结果[12]。

构造决策树的关键步骤是按属性进行树的分支，每个分支所包含的属性子集应属于同一类别。构造分支时要根据属性类型进行分裂，属性类型如果是离散的，则每个属性值划分一个分支；属性值如果是连续的，一般按照区域划分分支。属性选择度量算法一般采用递归分治法，笔者采用 ID3 算法构建 IT 专业就业分析决策树[13]，过程如下：

1) 计算数据集合信息熵和所有特征的条件熵，选择信息增益最大的特征作为当前决策节点[14]，计算信息熵公式如下：

$$H(X)=-\sum_{i=1}^{n}p(x_i)*\log p(x_i)。$$

式中 $p(x_i)$代表随机事件 X 为 x_i 的概率。

假设将数据集合 X 按属性 A 进行划分，则 A 对 X 划分的条件熵公式为：

$$H(X|\mathrm{A})=\sum_{j=1}^{\mathrm{m}}\left(|X_j|/|X|\right)H(X_j)=$$
$$-\sum_{j=1}^{m}\left(|X_j|/|X|\right)\left(\sum_{i=1}^{n}\left(|X_{ji}|/|X_j|\right)\log_2\left(|X_{ji}|/|X_j|\right)\right)。$$

式中：X_j 表示 X 中特征 A 取第 j 个值的样本子集；X_{ji} 表示 X_j 中属于第 i 类的样本子集。

信息增益=信息熵−条件熵。

具体公式为：

$$\mathrm{gain}(X,\mathrm{A})=H(X)-H(X|\mathrm{A})。$$

根据以上算法公式，计算各属性特征的信息增益，结果如表 7 所示。

表 7 属性特征信息增益

属性特征	信息增益	属性特征	信息增益
Gain(岗位)	0.266	Gain(学历)	0.532
Gain(工作地点)	0.471	Gain(工作经验)	0.338
Gain(薪资)	0.094	Gain(知识技能)	0.194

取信息增益值最大的属性“学历”作为根节点构建决策树的第一个分支。

2) 根据“学历”属性的特征值更新数据集合和特征集合，其中不同分支的数据集合应根据不同分支分别计算。

3) 循环 2)，直到数据集合中仅包含一个属性，其分支节点即为叶子节点。

3.3　实验及结果分析

以某非重点本科高校计算机软件专业硕士毕业生就业情况为例，毕业生就业意向为一线城市，软件后台开发设计能力较强，有一年的开发实习经验，薪资要求中上水平即可，根据基于决策树的 IT 岗位与技能聚类模型构建决策树如图 6 所示。

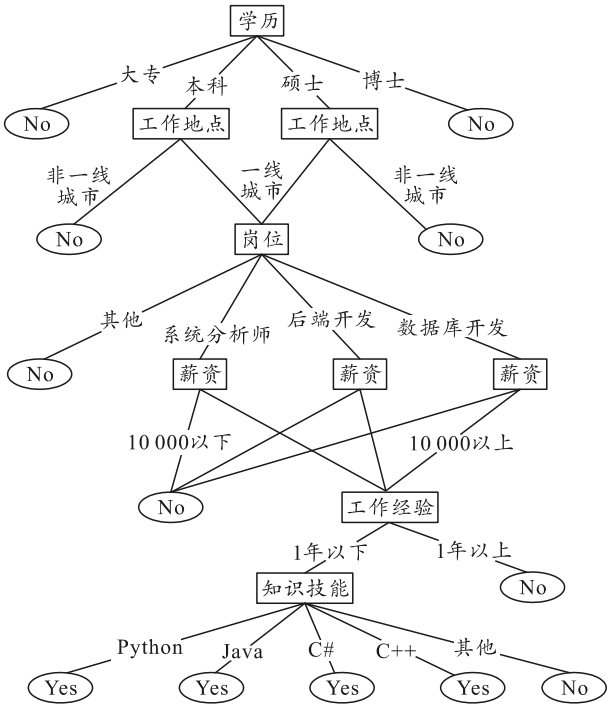

图 6　计算机专业硕士毕业生就业决策树

为验证基于决策树的 IT 岗位与技能聚类模型对高校毕业生就业及人才培养方案设置相关性的预测效果，在 4 所不同层次高校同级毕业的硕士研究生中随机抽取了部分学生应用此模型进行就业分析，得出就业预测率趋势及拟签约率趋势，如图 7、8 所示。

就业预测率算法如下[15]：

$$F=N_S/N_D\times 100\%。$$

式中：F 代表就业预测率；N_S 代表符合属性特征条件的数据总数；N_D 代表爬取治理后有效数据集合的记录数。

拟签约率算法如下：

$$S=N_E/N_F\times 100\%。$$

式中：S 代表拟签约率；N_E 代表已获得录用通知的数据总数；N_F 代表预测成功的数据总数。

通过以上实验结果可知：随着学校水平下降，学生就业竞争力明显下降，特别是拟签约率急剧下降；虽然此结果与社会对不同层级高校学生的认可度不同有很大关系，但也反映了水平相对较低学校人才培养要求与毕业生的真实水平能力不对等的问题，即可能存在水平较低高校照搬了高水平高校的人才培养方案，没有全面考虑学校、学生自身的差异性。

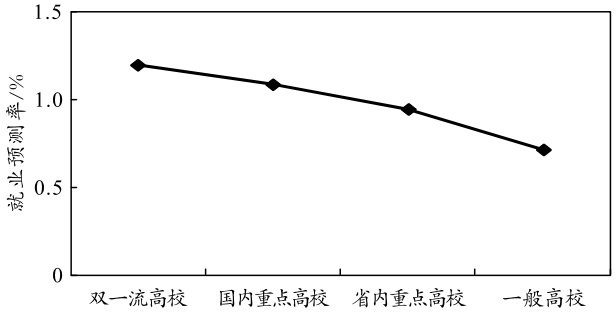

图 7　就业预测率趋势

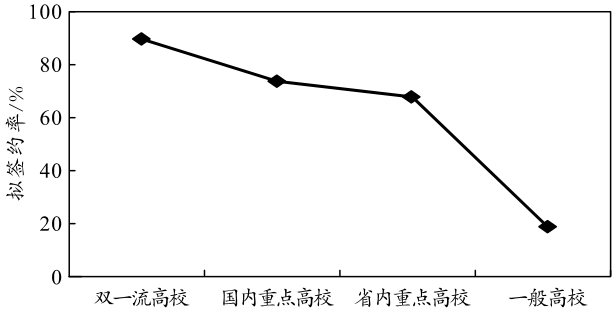

图 8　拟签约率趋势

4　结束语

笔者针对 IT 专业毕业生能力水平不能适用社会需求等问题，通过对大数据技术的研究，采用爬虫技术进行数据挖掘与分析，构建基于决策树算法的 IT 专业就业岗位和知识需求聚类模型，该模型采用 ID3 算法，以最大信息增益为目标，为社会对 IT 人才的需求及 IT 人才能力之间的匹配问题提供了一个解决方案。

将该模型在 4 所不同层次的高校 IT 专业硕士研究生的毕业生中进行实验，结果表明，层次较低高校学生的真实水平能力不能满足社会的需求，说明专业人才培养方案没有按学生能力量身定制，应适当的对培养方案作出修改。本研究对促进 IT 人才资源科学管理与决策、解决高校 IT 专业毕业生就业难、提高高校 IT 人才培养水平都会起到很大的作用；但本模型并未从深层次发掘高校毕业生能力与社会需求之间差距的原因，下一步研究可从企业招聘过程中被淘汰的学生能力方面入手，分析具体的差距因素，这将进一步促进专业人才培养方案的优化。

（下转第 74 页）