基于数据挖掘和 Apriori 算法的高校就业分析模型

张艺竞[1] 胡筱雨[1]
ZHANG Yijing HU Xiaoyu

摘 要 为了解决学生就业问题，通过对高校学生就业现状进行研究，文章提出一种基于数据挖掘技术与 Apriori 算法的高校就业分析模型。其研究创新点在于采用数据挖掘技术对高校学生就业相关数据进行挖掘，从而准确分析就业需求。同时引入改进 Aprioril 来构建就业分析模型，实现学生岗位的精准匹配。在基于兴趣的岗位匹配中，改进 Apriori 算法岗位匹配准确度为 0.962 3，优于同类模型。而在训练误差分析中，当数据分别为 20 万、60 万、100 万条时，改进 Apriori 算法误差分别为 1.025 6、0.832 4、0.625 4，均表现最优。可见，研究模型在高校就业分析中具有出色的应用效果，研究内容将为高校信息化管理以及学生就业指导提供技术支持。

关键词 数据挖掘；Apriori 算法；高校；就业；岗位匹配

doi：10.3969/j.issn.1672-9528.2025.08.033

0 引言

根据中国教育数据统计，2023 年中国高校毕业生人数规模达到了 1 158 万，相较于往年增加 82 万。大量高校毕业生进入社会并未解决部分企业用工荒问题，反而出现大量高校毕业生“毕业即失业”的现象 [1]。特别是疫情后， 社会发展节奏稍有平缓，加上高校每年毕业人数激增，导致学生就业问题严峻。近年来，随着信息化技术的不断发展，基于数据挖掘与数据信息匹配的相关技术正为高校学生个性化就业提供相关支持。如陈刚 [2] 为解决高校人事档案信息管理不足问题，其基于人工智能技术提出一种数据信息挖掘方法，通过机器学习建立数据训练模型，从而准确挖掘重要人才信息。实现结果显示，该技术挖掘效率更高，显著优于人工数据管理，且与同类技术相比技术表现优异。由上述研究可以看出，数据挖掘技术以及机器学习在就业数据挖掘以及分析过程具有优异的性能表现。面对高校学生工作就业困难问题，研究基于 Apriori 算法与数据挖掘技术提出一种智能化的高校就业分析模型，通过对高校学生就业信息的挖掘，为学生匹配适合的岗位，从而解决学生就业问题。研究内容将为高校信息化建设以及学生就业管理提供技术支持。

1 基于 Apriori 算法的高校就业分析模型构建

近年来，高校学生面临的就业问题越发严峻，如何有效地找到合适的工作已成为高校学生面临的首要问题 [3]。对此，研究基于 Apriori 算法提出一种高校就业分析模型，通过对学生兴趣特征等数据的挖掘分析从而为其匹配适合的工作，以解决学生就业问题。整个高校大学生就业分析系统如图 1 所示。

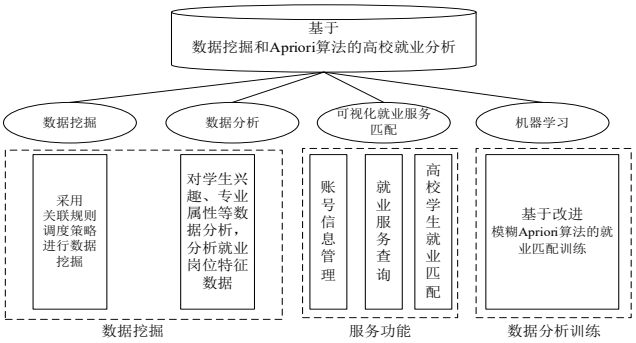

图 1 高校大学生就业分析系统

根据图 1 中的高校就业分析系统，该系统由数据挖掘、数据分析、可视化就业服务匹配以及机器学习 4 个部分构成。首先，就业分析系统将收集的学生兴趣特征、专业属性特征等信息通过大数据技术进行融合，得到用户与项目模糊聚类自适应学习权重，用公式表示为 [4]：

$$\begin{cases} W_k(\boldsymbol{U}) = \alpha\left(\frac{1}{m}\sum_{i=1}^{m} r_{ij}\cdot\delta\right)\ j \in \mathrm{User}_i \\ W_k(\boldsymbol{V}) = \alpha\left(\frac{1}{n}\sum_{i=1}^{m} r_{ij}\cdot\delta\right)\ j \in \mathrm{User}_i \end{cases} \quad (1)$$

式中：$W_k(\boldsymbol{U})$ 与 $W_k(\boldsymbol{V})$ 分别表示用户与项目模糊聚类自适应权重；$\boldsymbol{U}$ 与 $\boldsymbol{V}$ 表示模糊聚类特征向量；α 表示标准就业匹配

1. 吉林农业科技学院 吉林吉林 132000
[基金项目] 吉林省高教学会高教科研课题（JGJX2023D599）

参数；m 表示大学生数量；n 表示就业项目数；r_{ij} 表示两两学生间的关联系数；δ 表示就业分析特征参数；User 表示大学生关联指向权重。

通过系统数据分析，为了进一步确定用户与兴趣特征相关性，引入差异度策略挖掘用户潜在就业特征信息，基于兴趣的就业匹配就需要满足最大联合特征分布要求，用公式表示为：

$$\max_{x_{i,j,m,n}} \sum_{\alpha\in r_i}\sum_{\alpha\in r_j}\sum_{\alpha\in r_m}\sum_{\alpha\in r_n} x_{i,j,m,n}\boldsymbol{V}_n\alpha \tag{2}$$

式中：r_i 表示就业信息关联系数；r_j 表示大学生关联系数；r_m 表示项目关联系数；r_n 表示就业项目关联系数；$\boldsymbol{V}_n$ 表示就业项目对应模糊特征向量；$x_{i,j,m,n}$ 表示学生兴趣、专业属性等个性化特征序列。

在对学生就业兴趣偏好信息分析中，将利用模糊 Apriori 算法来对学生群体规模进行划分，从而更好地适应对学生就业数据的分析。则 Apriori 算法的学习函数计算公式为：

$$W(k)=W_k(\boldsymbol{U})\left[1-W_k(\boldsymbol{V})\right]^{k-1} \tag{3}$$

式中：$W(k)$ 表示学习函数；k 表示兴趣参数。

接下来计算学生兴趣与个性化就业需求间的模糊度量化函数，用公式表示为：

$$E(k)=1/(1-1/n)^{m-1} \tag{4}$$

式中：$E(k)$ 表示模糊度函数，用于学生兴趣与项目匹配分析。

计算 Apriori 算法学习的时隙参数如公式所示：

$$T_{\text{lary}}=E(k)n_i=L/(1-1/n)^{m-1} \tag{5}$$

式中：n_i 表示兴趣就业指数；T_{lary} 表示平均时隙参数。

接下来，研究利用 Apriori 算法对就业匹配的兴趣特征点进行匹配，获得基于用户兴趣的学生就业匹配模型如公式所示 [5]：

$$\begin{aligned}E^{cv}(c_1,c_2)=&\mu\cdot\text{Length}(C)+v\cdot\text{Area}(\text{inside}(C))+\\&\lambda_1\int_{\text{inside}(C)}|I-c_1|^2+\lambda_2\int_{\text{outside}(C)}|I-c_2|^2\end{aligned} \tag{6}$$

式中：μ、v、λ_1、λ_2 均表示就业语义相关参数，取值为常数；c_1 与 c_2 表示两组数据相似属性偏好；Length(C) 表示就业岗位分布长度属性；Area(inside(C)) 表示区域分布参数；I 表示项目集合；inside(C) 表示区域分布外部尺寸参数；outside(C) 表示区域分布内部尺寸参数。

根据上述分析，便可以得到算法学习模型：

$$C=\text{Min}\left\{\max(C_i)\right\} \tag{7}$$

式中：C_i 表示模糊就业关联系数。

匹配满意度计算为：

$$\sum_{j=1}^{n}Z_j=1,\forall i\in(1,n),\forall j\in(1,n_i) \tag{8}$$

式中：Z_j 表示就业满意度水平。

接下来，需要通过学习训练使 Apriori 算法获得最优解。因此，研究中将高校大学生就业关联信息分布集定义为 $S=\overline{X_1},\overline{X_2},\cdots,\overline{X_k},\cdots$，而根据就业分析数据给学生匹配有效的工作岗位，则匹配的岗位特征集定义为 $T_R=T_1,T_2,\cdots,T_K,\cdots$。则 Apriori 最优训练特征向量用公式表示为：

$$\overline{w}_k^i=\overline{w}_{k-1}^i\frac{l(z_j/\overline{x}_k^i)(\overline{x}_k^i/x_{k-1}^i)l}{q(\overline{x}_k^i/\overline{x}_{k-1}^i)} \tag{9}$$

式中：$\overline{w}_k^i$表示兴趣偏好权重；q 表示就业意向指数；$\overline{x}_k^i$表示就业偏好兴趣度。

则对算法训练过程进行优化，得到优化后的匹配迭代：

$$x_i(k+1)=x_i(k)+\alpha\left(\frac{x_j(k)-x_i(k)}{\|x_j(k)-x_i(k)\|}\right) \tag{10}$$

式中：$x_j(k)$ 表示项目兴趣特征序列；$x_i(k)$ 表示学生兴趣特征序列。

2 基于数据挖掘的就业分析系统建模

在大学生就业分析系统中，数据挖掘是系统服务层的基础，需要通过有效的数据挖掘与分析获得学生就业属性特征。因此，研究采用基于关联规则调度策略进行相关数据特征挖掘。通过数据挖掘得到的学生融合信息为：

$$p(x)=\frac{x_m}{\sum_{i=1}^{n}I_i\cdot u_m} \tag{11}$$

式中：x_m 表示大学生就业意向特征序列，其中既有兴趣特征，也包含专业特长特征以及相关经历特征；u_m 表示学生就业兴趣指数；I_i 表示学生就业项目发展集合；i 表示就业信息量。接下来需要计算用户兴趣特征分布，以更好分类数据信息，计算公式为 [6]：

$$P(k)=p(x)/\sum_{k=1}^{n}I_i(l(k)\cdot q(k)) \tag{12}$$

式中：l 表示就业项目指数；$P(k)$ 表示兴趣特征分布。

接下来挖掘就业大学生用户间的相似度参数：

$$\sin(u_a,u_b)=\sum_{i\in I_a\wedge i\in I_b}p(x)P(k)\cdot r/\sqrt{\sum_{i\in I_a}(r_{u_a,i}-\overline{r}_{u_a})^2\cdot\sum_{i\in I_b}(r_{u_b,i}-\overline{r}_{u_b})^2} \tag{13}$$

式中：r 表示就业匹配关联性系数；I_a 表示用户 u_a 评价集；I_b 表示用户 u_b 的评价集；$r_{u_a,i}$表示用户 u_a 的就业信息关联性系数；$\overline{r}_{u_a}$表示用户 u_a 平均匹配关联性系数；$r_{u_b,i}$表示用户 u_b 的就业信息关联性系数；$\overline{r}_{u_b}$表示用户 u_b 平均匹配关联性系数。

通过上述研究便可以挖掘到大学生就业关键数据信息，但仅通过融合的用户特征并不利于系统训练，为大学生匹配适合的工作 [7]。对此，需要对学生就业项目特征进行进一步提取分析，则就业相关性特征提取为：

$$y=F(x)=(f_1(x),f_2(x),\cdots,f_m(x))m \tag{14}$$

式中：$F(x)$ 表示岗位分配节点集；$f_m(x)$ 表示通过对学生就业分析，向 m 学生分配就业。

考虑不同领域学生对就业需求不同，因此需要对不同专业需求学生进行就业与兴趣相关特征的匹配，得到适应特征空间分布为：

$$f(n_i)=\{f_k \mid f_k=1,k=1,2,\cdots,m\} \tag{15}$$

式中：f_k 表示兴趣参数特征集合；$f(n_i)$ 表示适应特征空间分布。

考虑到研究中采用模糊 Apriori 算法作为就业分析匹配模型，为了方便模型学习训练，在数据中引入差异度因素，通过不同组间用户就业匹配程度，得到模糊函数的兴趣匹配优化值为：

$$\text{Opti}=t_{k,j}n_i=\sum_{k=1}^{m}t_{i,k}t_{k,j}n_i \tag{16}$$

式中：Opti 表示模糊度函数优化值；$t_{k,j}$ 表示用户兴趣与项目间的差值；$t_{i,k}$ 表示就业信息与用户信息间的差值。

整个基于数据挖掘与 Apriori 算法的大学生就业分析流程如图 2 所示。

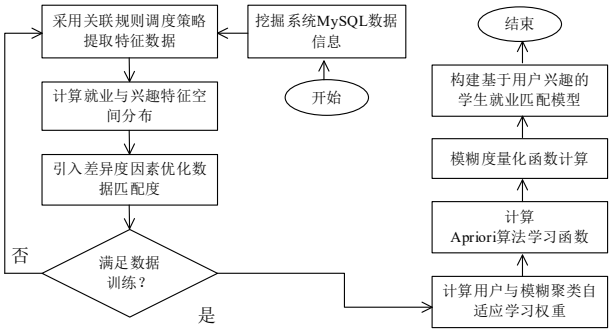

图 2 基于数据挖掘与 Apriori 算法的大学生就业分析流程

3 模型应用效果分析

为了检验研究所提出的技术，研究将开展相应的实验。其中，高校就业分析系统基于成熟的 Django 框架进行开发，系统运行语言采用 Python，其相比传统 JAVA 语言其功能设计更加简单，且能缩短开发周期，系统 Web 前端则采用 Axure PR 软件进行辅助设计，并使用了 Bootstrap 框架。而系统数据可视化部分包括线上数据爬取、数据查询、地区岗位分析等功能。整个系统数据采用 MySQL 进行设计，并在系统中采用改进 Apriori 算法进行岗位数据关联匹配，为高校学生提供重要的就业信息匹配服务。高校就业服务系统可视化界面如图 3 所示。

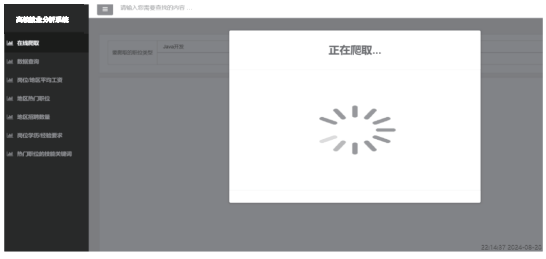

图 3 高校就业分析系统可视化界面

而在具体的实验中，测试系统为 WindowS11，处理器为 INTEL i5 16 核处理器，显卡为 RTX4060Ti，实验仿真平台为 MATLAB 2021。实验中改进 Apriori 算法模型参数设置如表 1 所示。

表 1 实验模型相关参数设置

参数类型	数值
迭代训练次数	100
C_2 与 C 属性偏好参数	0.34 与 0.32
训练误差参数	0.8
就业学生数据特征量	200
实验数据量 / 万条	200

实验中引入频繁模式增长算法（frequent pattern growth, FP-Growth）以及文献 [7] 中的改进词频 - 逆向文件频率算法（term frequency-inverse document frequency, TF-IDF）作为测试基准。首先对不同模型的岗位匹配准确度进行测试，结果如图 4 所示。

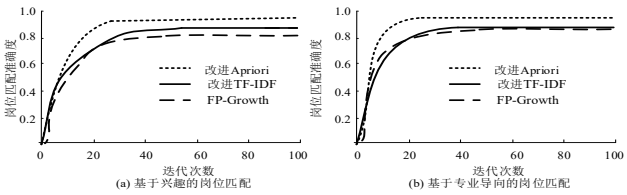

图 4 不同模型岗位匹配准确度比较

如图 4 所示，研究中选取基于兴趣的岗位匹配以及基于学生专业为导向的岗位匹配进行分析，分别如 4（a）与 4（b）。在基于兴趣的岗位匹配中，改进 Apriori 算法相比同类技术其优势明显，如其在迭代 20 次后取得收敛，此刻岗位匹配准确度为 0.962 3，而改进 TF-IDF 表现次之，其在迭代 60 次后取得收敛，此刻岗位匹配准确度为 0.886 8。而 FP-Growth 算法表现一般，其在迭代 40 次取得收敛，此刻岗位匹配准确度为 0.814 5。而在基于专业导向的岗位匹配中，改进 TF-IDF 与 FP-Growth 取得收敛时岗位匹配准确度基本一致，分别为 0.835 2 与 0.862 5，而改进 Apriori 算法岗位匹配准确度最高，为 0.968 3，且最快收敛。接下来，引入平均精确度（mean average precision, mAP）与均方根差误差（root mean square error, RMSE）来比较不同模型性能，如图 5 所示。

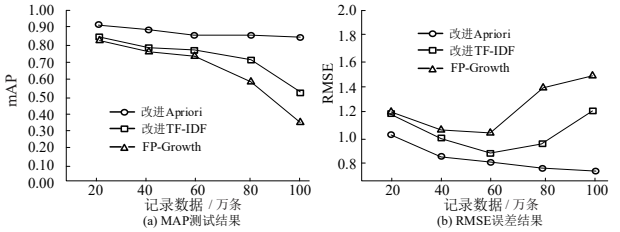

图 5 mAP 与 RMSE 误差测试

图 5（a）为不同模型 mAP 测试结果，从测试结果来看，

在数据规模为 20 万条时，3 种模型训练 mAP 值均比较高，其中改进 Apriori 算法为 0.923 5，而改进 TF-IDF 为 0.856 5，而 FP-Growth 为 0.841 2。随着记录数据规模的不断扩大，3 种模型的 mAP 值均在逐步降低，其中变化幅度最大的是 FP-Growth 算法，当记录数据为 100 万条时，其 mAP 值下降到了 0.324 5。同时改进 TF-IDF 算法的 mAP 值也大幅下降，当记录数据为 100 万条时 mAP 值为 0.512 5。而整体变化最好的是改进 Apriori 算法，记录数据为 100 万条时其 mAP 值下降到了 0.872 2，模型整体性能下降最低。FP-Growth 与改进 TF-IDF 的 mAP 值大幅下降与其数据规模处理能力有关，当记录数据超过 80 万条时，模型数据挖掘能力明显受限。在图 5（b）的 RMSE 误差分析中，同样选取不同规模数据分析模型训练误差情况，其中 FP-Growth 算法与改进 TF-IDF 算法的 RMSE 误差随着数据规模的增大呈现误差减小又随之增大的趋势，如当数据规模为 20 万条时，FP-Growth、改进 TF-IDF 以及改进 Apriori 的 RMSE 误差分别为 1.252 2、1.251 2 与 1.025 6，当数据为 60 万条时，RMSE 误差分别为 1.125 6、0.915 6 以及 0.832 4。当数据为 100 万条时，改进 Apriori 算法的 RMSE 误差为 0.625 4，而 FP-Growth 算法与改进 TF-IDF 算法分别为 1.256 1 与 1.578 2。当数据规模控制在 60 万条时，数据信息的丰富有利于提升模型的训练性能，使得 3 种模型误差均下降，而数据规模超过 80 万条后，FP-Growth 算法与改进 TF-IDF 算法数据处理能力均受到限制，进而影响模型训练效果。接下来比较不同模型的运行耗时与处理能力，如图 6 所示。

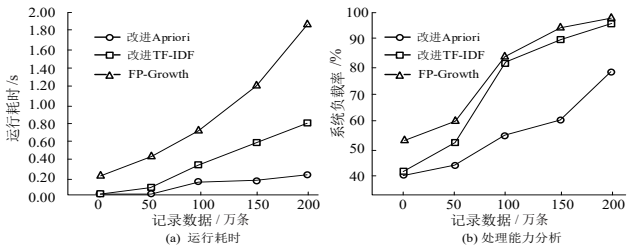

图 6 不同模型运行耗时与处理能力比较

图 6（a）为不同模型运行耗时比较结果，根据测试结果来看，随着数据规模的扩大，三种模型数据处理时间均不断上升，但改进 Apriori 算法整体表现最好。如数据规模为 200 万条时，改进 Apriori 算法运行耗时为 0.235 4 s，改进 TF-IDF 算法为 0.812 5 s，FP-Growth 算法为 1.982 4 s。此外，对不同模型的数据处理能力进行比较，如图 6（b）所示。由图结果来看，改进 TF-IDF 算法与 FP-Growth 算法在数据规模达到 100 万条时其负载均超过了 80%，而改进 Apriori 算法仅为 54%。且当数据规模达到 200 万条时，仅有改进 Apriori 算法未满负载。可见，研究所提出的技术在高校就业分析中具有优异的应用效果，综合表现更为出色。

4 结语

近年来，越来越多的大学生面临就业压力问题。对此，研究基于数据挖掘技术与机器学习技术提出一种智能化的高校就业分析模型，通过对学生就业相关数据的挖掘，引入模糊 Apriori 算法来建模，从而实现对学生就业岗位的精准匹配。在基于专业导向的岗位匹配分析中，改进 TF-IDF、FP-Growth 以及改进 Apriori 算法岗位匹配准确度分别为 0.835 2、0.862 5 与 0.968 3。在 mAP 测试中，改进 Apriori 算法为 0.923 5，而改进 TF-IDF 为 0.856 5，FP-Growth 为 0.841 2。在运行耗时分析中，当数据规模为 200 万条时，改进 Apriori 算法运行耗时为 0.235 4 s，改进 TF-IDF 算法与 FP-Growth 算法分别为 0.812 5 s 与 1.982 4 s。由此可见，研究所提出的模型具有优异的数据处理与分析能力，且在岗位匹配中应用效果优异。不过研究技术并未接入招聘系统，未来接入招聘系统数据，以进一步提高系统应用效果。

参考文献：

[1] 蒋茜茜 , 杨风暴 , 杨童瑶 , 等 . 基于改进 Apriori 算法的高校体测数据关联分析 [J]. 计算机系统应用 ,2022,31(5):345-350.

[2] 陈刚 . 基于 AI 大模型的高校人事档案信息数据挖掘研究 [J]. 江苏科技信息 ,2024,41(2):107-110.

[3] MISHRA R, RATHI S. Enhanced DSSM (deep semantic structure modelling) technique for job recommendation[J]. Journal of king saud university-computer and information sciences, 2022, 34(9): 7790-7802.

[4] 何韦颖 , 钟健 , 陈有拾 . 基于数据挖掘的高校校友管理信息系统研究 [J]. 信息与电脑 ,2024,36(6):86-88.

[5] 张梁 , 杨立波 , 张小勇 , 等 . 基于改进的 Apriori 算法在高校成绩分析中的研究 [J]. 信息记录材料 ,2024,25(3):142-145.

[6] 王昊禾 , 张悦 , 江宇琪 . 基于数据挖掘的高校学生考研成绩预测分析 [J]. 武夷学院学报 ,2024,43(1):93-97.

[7] 李龙 , 金铄 , 黄霞 . 基于改进 TF-IDF 算法的毕业生就业推荐算法研究 [J]. 计算机与数字工程 , 2023, 51(9): 1985-1989.

【作者简介】

张艺竞（1991—），女，吉林省吉林人，硕士研究生，讲师，研究方向：农业数据处理。

胡筱雨（2004—），女，内蒙古呼和浩特人，本科在读，研究方向：农业数据处理。

（收稿日期：2025-03-19 修回日期：2025-07-31）