Job Recommendation Challenge

一、研究背景及分析流程

1. 研究背景

人岗匹配是商务智能在人力资源管理领域中的一个重要应用，通常指通过识别不同工作所需要的人才能力，将合适的人才匹配到合适的工作岗位之上。在以往的工作中，不同学者也对这个问题进行过诸多探索，早期的研究较多地从定性视角出发(Holland, 1973; Amabile, 1983; Cable & Derue, 2002)，如通过问卷衡量用户的人格特征等，但基于定性方法的能力衡量和工作匹配相对主观，可能会导致不必要的偏见。近年来，逐渐有学者将人岗匹配问题视为推荐问题，试图通过机器学习、协同过滤等商务智能技术，促进求职者和工作之间更好的匹配。如 Diaby 等(2013)、Lu 等(2013)、Paparrizos 等(2011)分别从工作内容、用户特征和用户历史工作轮换的角度出发，设计了适用于不同业务情境下用户-工作匹配的岗位推荐系统。此外，一些学者也致力于设计和改进现有的岗位预测模型，实现更精准的匹配，如 Zhu 等(2018)提出了一套基于 CNN 的人岗匹配网络；Zhou 等(2019)利用协同过滤方法，构建了考虑用户偏好的用户-工作匹配模型；Chen 等(2020)提出了一类考虑雇主行为特征的双边匹配方法。本研究旨在综合考虑用户信息、用户历史工作和岗位信息三方面特征，设计面向双边需求的用户-工作匹配度预测模型，以实现用户和工作的精准匹配。

2. 分析流程

本研究给出了一套包含用户-工作匹配预测和工作推荐的分析框架，主要包括以下三部分工作。

首先，基于 CareerBuilder.com 的原始数据集，进行一系列数据清洗、数据负采样、文本表示及特征构建，筛选出能够实现本研究需求的用户-工作匹配数据集，涵盖用户信息（基本信息和历史工作）、工作信息、用户申请情况等。

其次，利用经过上述预处理的数据集训练用户-工作匹配预测模型，标签定义为用户是否会申请相应工作，本研究分别考虑了基于 TF-IDF 文本表示策略的传统机器学习模型和基于 Word2Vec 词嵌入策略的 CNN 网络(PJFNN-modified)。

最后，分别在两个任务场景(Person-job Fit 和 Job Recommendation)下探究了不同模型的表现。Person-job Fit 是在给定用户和工作岗位的情境下，判断用户和工作是否匹配，本质上是一个二分类的问题；Job Recommendation 则是给指定用户推荐合适的工作。

二、数据预处理与特征构建

1. 数据说明

本研究采用的实验数据来自于 Kaggle 竞赛平台所披露的 CareerBuilder.com 的公开数据集。CareerBuilder.com 为美国最大的求职网站之一，他们链接了数十万寻求优秀人才的雇主和数百万找合适机会的求职者，致力于降低人才市场供需双方的信息不对称，实现求职者与岗位之间更精准的匹配。

数据集主要包含四张基本表，分别为用户信息表 users.tsv，历史工作表 user_history.tsv，工作发布表 jobs.tsv 和用户申请表 apps.tsv，数据集的时间窗口为 2012 年 6 月 5 日-2012 年 6 月 18 日。数据集中包含的字段以及含义如下。

(1) 用户信息表

表 1 用户信息表

字段	字段说明	数据量	类型	示例
User ID	用户 ID	43334	int64	14350
Split	训练集/测试集	43334	int64	Train
City	城市	43334	string	Columbus
State	州	43276	string	OH
Country	国家	43334	string	US
ZipCode	邮政编码	43142	string	43230
DegreeType	学历类型	43334	string	Bachelor's
Major	专业	32433	string	Business
WorkHistoryCount	历史工作数	29703	int64	6
TotalYearsExperience	工作年限	41733	float64	9.0
CurrentlyEmployed	是否在职	40653	string	0
ManageOthers	是否担任管理者	43334	string	1
ManageHowMany	管理幅度	43334	int64	5

(2) 历史工作表

表 2 历史工作表

字段	字段说明	数据量	类型	示例
User ID	用户 ID	193853	int64	14350
Split	训练集/测试集	193853	string	Train
Sequence	历史工作排序	193853	int64	2
JobTitle	工作名称	180658	string	Counselor

(3) 工作发布表

表 3 工作发布表

字段	字段说明	数据量	类型	示例
Job ID	工作 ID	115691	int64	14350
Split	训练集/测试集	115691	int64	Train
City	城市	115691	string	Columbus

State	州	115691	string	OH
Country	国家	115691	string	US
ZipCode	邮政编码	71509	string	43230
Title	工作名称	115691	string	Administrative Assistant
Description	工作描述	115691	string	...
Requirements	工作需求	115691	string	...
StartDate	发布时间	115691	datetime	2012-05-13 01:16:58.923
EndDate	截止时间	115691	datetime	2012-06-12 23:59:59

(4) 用户申请表

表 4 用户申请表

字段	字段说明	数据量	类型	示例
User ID	用户 ID	120457	int64	14350
Job ID	工作 ID	120457	int64	14350
Split	训练集/测试集	120457	string	Train
ApplicationDate	申请时间	120457	datetime	2012-05-23 09:52:03.327

2. 数据清洗

考虑到原始数据集存在诸多冗余字段、缺失数据、无效字段等，本研究考虑对 4 个数据集进行了如下清洗工作：

- 基于 Country 字段筛选用户信息表和工作发布表，保留美国地区的求职者和工作发布信息 (data.Country == 'US')；
- 用户信息表中去除没有 work_history 或 application_record 的用户；
- 用户信息表中去除申请数量过多（超过 5 个）的用户；
- 去除各表中的冗余字段，用户信息表中的"Country", "ZipCode", "Major", "GraduationDate"，历史工作表中的"Sequence"，工作发布表中的"Country", "ZipCode", " StartDate ", " EndDate "，用户申请表中的"ApplicationDate"等；
- 去除各表剩余字段缺失值所在元组和重复元组；
- 将用户信息表中的字符型字段转换为数值型(binary)字段，即对字段 CurrentlyEmployed: Yes=1, No=0；对字段 ManagedOthers: Yes=1, No=0；对字段 DegreeType: "None=0, High School=1, Vocational=2, Associate's=3, Bachelor's=4, Master's=5, PhD=6；
- 对各字段数据进行标准化。

3. 数据负采样

用户-工作匹配预测问题本质上可以认为是一类二分类问题，即认为如果用户会选择申请相应工作，则用户-工作匹配度(label)取值为 1，而如果用户不会选择申请相应工作，则用户-工作匹配度(label)取值为 0。

考虑到在用户申请历史记录中，只会记录用户申请的工作，而不会记录用户不愿意申请的工作，即在数据集中只存在正样本(positive sample)而不存在负样本(negative sample)。参考 Zhu 等人(2018)，本研究对原始数据集进行了负采样，1:1 地为用户申请表中的每一条申请记录从工作发布表中随机采样一条当前用户未申请过的工作作为“未申请记录”，负样本除工作 ID 和样本标签(label)外与正样本完全相同。经过负采样的用户申请表数据量为原表的 2 倍（label 为 0 和 1 的数据各 120457 条）。

4. 地区特征构建

考虑到不同层次地区(City, State, Country)之间存在嵌套关系，且每一层次的地区取值采用简单的整数编码或 one-hot 编码均存在一定的不合理，本研究考虑在用户-工作匹配过程中构建相同地区(取值为 1)和不同地区(取值为 0)的二元离散变量 City 和 State（清洗后数据集中 Country 均为 US），用以表示求职者和工作是否处在相同的城市、是否处在相同的洲。

5. 文本表示

针对工作发布表中的 Title, Description, Requirements、历史工作表中的 JobTitle 等字段，需要进行一定的特征提取、文本表示处理，才能输入分类预测模型。本研究针对不同的分类预测模型分别考虑了两类处理方法，在基准模型中，本研究考虑了基于 TF-IDF 的文本表示策略，在 CNN 模型中，考虑了基于 Word2Vec 框架的文本表示策略。

(1) TF-IDF

TF-IDF 模型是一类在信息检索和文本挖掘等领域广泛应用的文本表示技术，词频 TF(Term Frequency)为文本内词汇出现的频率，逆文本频率 IDF(inverse document frequency)为一个词语普遍关键性的度量。本研究利用 sklearn 库中提供的 TfidfVectorizer 类，分别对工作发布表和历史工作表中的各类文本构建 VSM(Vector Space Model)矩阵，向量维度分别设定为 100 和 50。

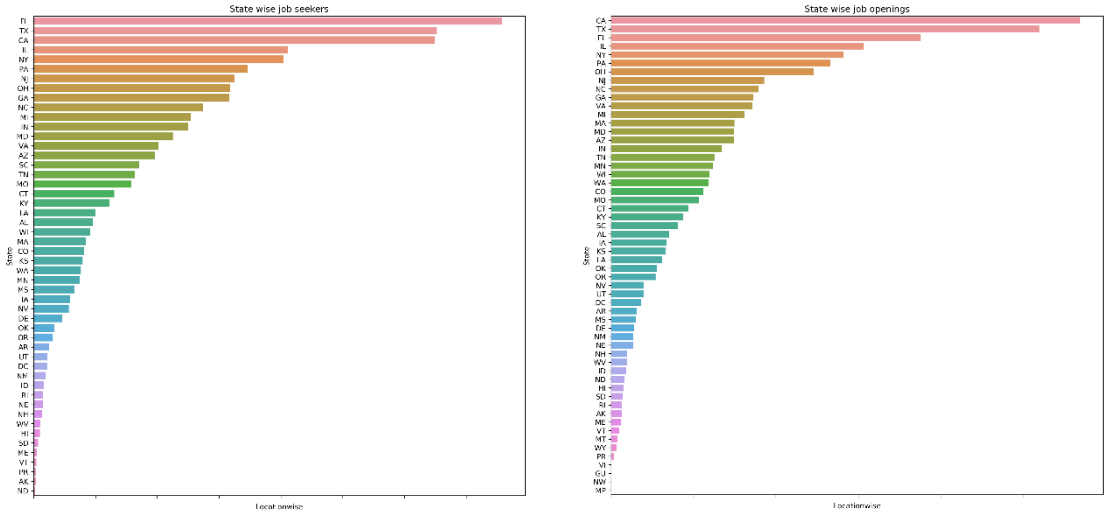
(2) Word2Vec

考虑到 TF-IDF 仅仅通过词频衡量文本中词语的重要性，词汇之间相互独立，并没有考虑到文本的语义特征，无法反映文档的序列信息，本研究进一步考虑了采用 Word2Vec 词嵌入进行语义表示。Word2Vec 词向量模型是 Mikolov 等(2013)提出的用于文本词向量学习的三层神经网络模型，通过无监督的方式训练语言模型(language model)，将词汇量化为低维空间中的稠密实值向量，实现文本词汇的特征表达。

三、数据可视化

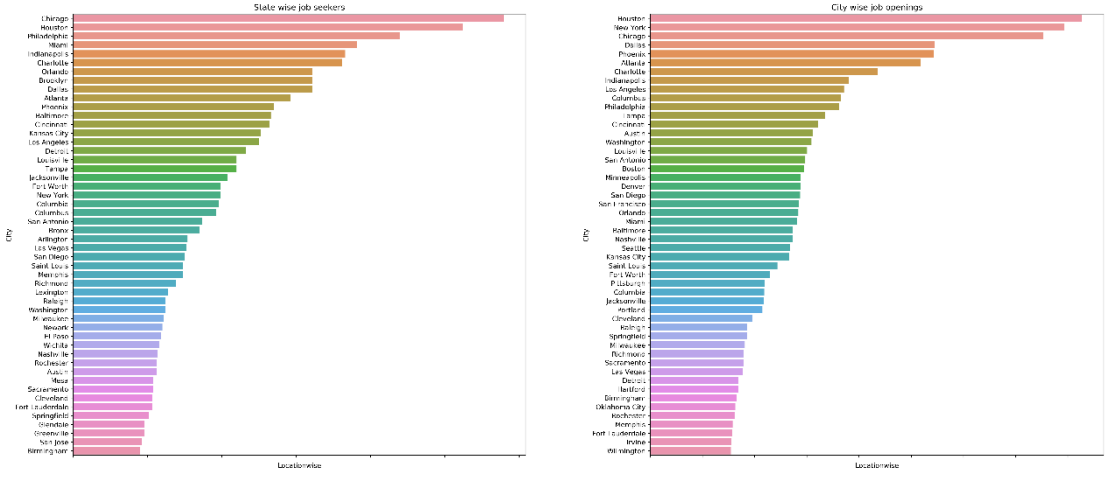
1. 基于地区特征的探索性分析

考虑到用户信息和工作信息中均有关于国家、州和城市等地区信息，可以对其进行一定探索性分析。如图 1、2 给出了不同用户、不同工作的洲及城市分布图像，在洲分布方面，可以看到，大多数的求职者和用人单位分布在加利福尼亚州、德克萨斯州和弗罗里达州等，虽然大致分布相似但仍存在一定的不均衡、不匹配问题；在城市分布方面，呈现出一定的长尾分布，且存在更明显的不均衡现象，如纽约虽然具有比较高的岗位需求，但用户需求事实上并不突出。



(a)用户分布　　(b)工作分布

图 1　不同用户和不同工作的洲分布



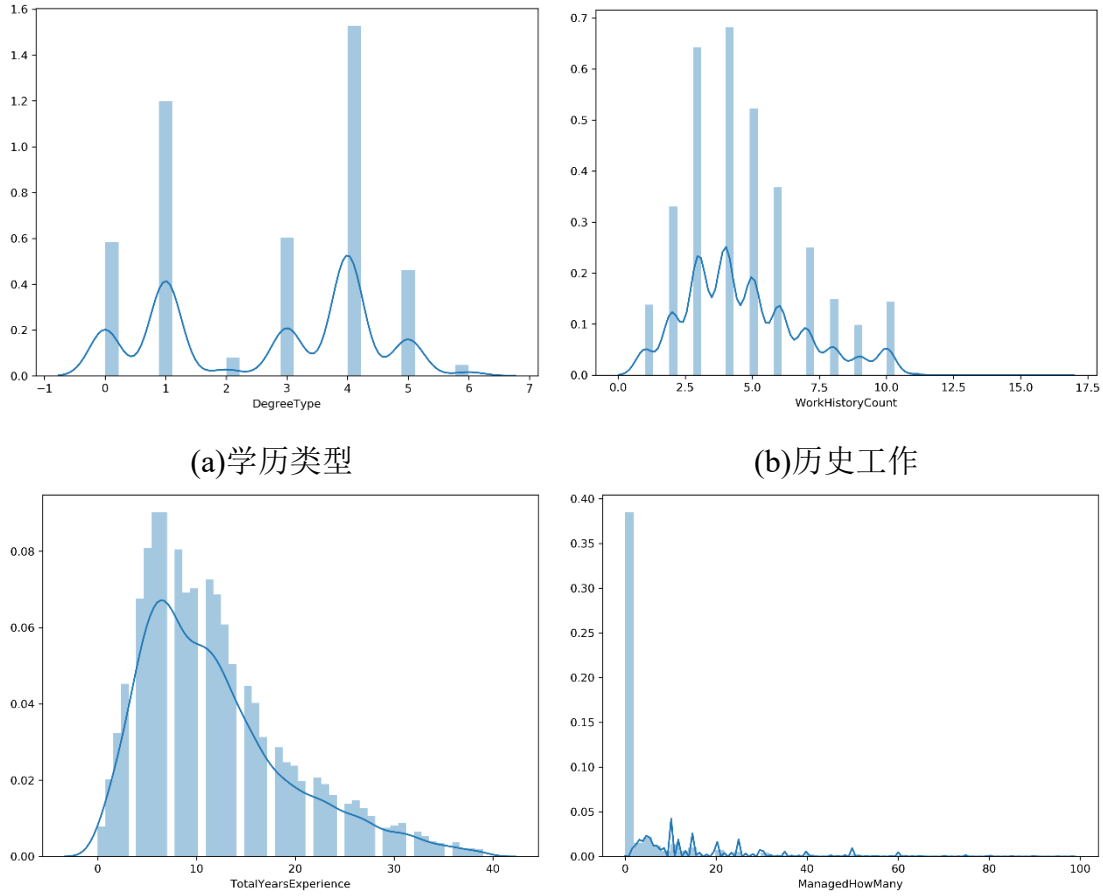
(a)用户分布　　(b)工作分布

图 2　不同用户和不同工作的城市分布

2. 用户信息可视化分析

下面的四张图描述了用户数据集中一些用户特征的基本分布。在学历方面，用户多数为本科学历(Bachelor's)和高中学历(High School)；在历史工作数量方面，

大多用户集中在 2 年到 7 年；在历史工作年限方面，用户则呈现出近似右偏的偏态分布，峰值约在 5-6 年；在管理幅度方面，绝大多数用户并没有管理经验，其管理幅度表现为 0。



(a)学历类型　　(b)历史工作

(c)工作年限　　(d)管理幅度

图 3 用户基本特征的分布

3. 文本可视化分析

下面的两张图分别给出了工作信息表和用户历史工作表中文本的可视化词云图。可以看到，工作信息表中的文本提供许多有关服务岗位、销售岗位的信息，并且通常需要员工具备一定的工作经验(experience)和与岗位相匹配的工作技能(skill, ability)；用户历史工作分布则相对更广泛和分散，虽然也以服务岗位、销售岗位等居多，但可以看出与工作信息表中的工作需求存在一定的差异，针对不同用户、不同工作的人岗匹配和工作推荐仍具有一定的现实意义。

(a)工作信息　　(b)历史工作

图 4 文本可视化词云图

四、PJFNN 与 PJFNN-modified

Zhu 等(2018)设计了一个端到端的 Person-Job Fit Neural Network，名为 PJFNN。其主要由 Text CNN 组成，结构如图 1 所示。首先用两个卷积层和池化层分别处理简历和岗位描述的文本信息，得到特征表示后进行 Mean/Max pooling 操作。最后将简历和岗位描述的特征向量输入到多层感知机中预测匹配结果。

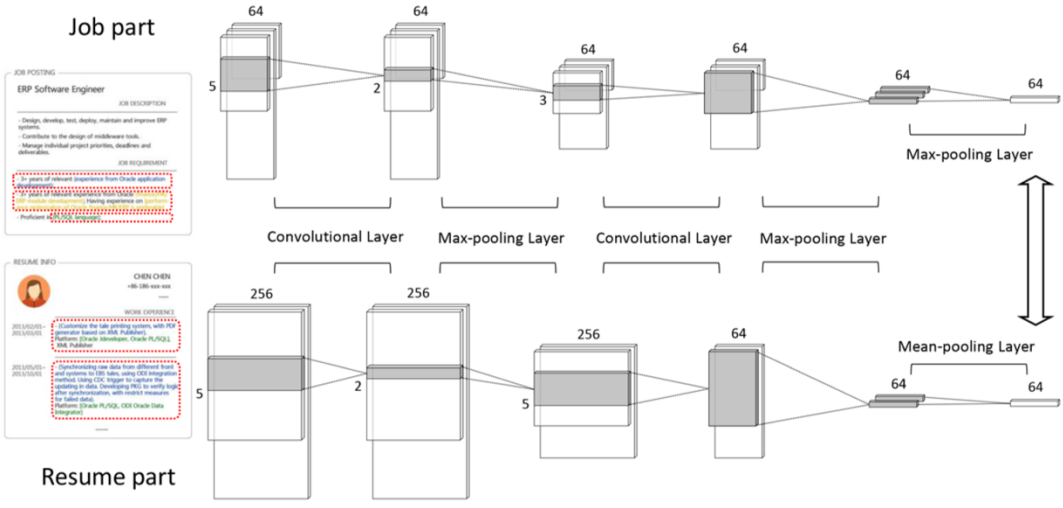


图 5 PJFNN 模型结果

该模型所使用的 Text CNN 模块能够有效的提取文本信息，且参数量较小，计算效率高。本研究所使用的数据不包含用户的简历数据，也没有用户历史工作记录的详细描述，无法直接应用 PJFNN。故本研究将针对简历部分为 Text CNN 修改为全连接的神经网络，该网络的输入是用户的基本信息（学历类型、工作年限等）以及用户历史工作名称的文本表示。将这一模型命名为 PJFNN-modified，后续实验中将仔细探究这一模型的表现。

五、任务一：Person-Job Fit

Person-Job Fit 任务的主要目的是判断用户和工作岗位是否匹配，它可以视为一个典型的二分类问题(0: 不匹配，1: 匹配)。为了探究 PJFNN-modified 在 Person-Job Fit 任务上的表现，本研究选用了以下基准模型：线性回归，Logistic 回归，决策树，朴素贝叶斯，决策树，随机森林，AdaBoost，GBDT，XGBoost。主要的对比指标为常见的二分类评价指标：Accuracy，Precision，Recall，F1-score，Area Under Curve(AUC)。

基准模型的输入是工作岗位描述的 TF-IDF 向量，以及用户的个人属性和历史工作名称的 TF-IDF 向量。PJFNN-modified 中，工作岗位表述中的词汇表示为 200 维的词向量，用户相关的输入与基准模型相同。

表 5 展示了所有模型的评价结果。从结果中可以看出，线性模型表现较差，准确率均略高于 0.5。大部分集成模型（随机森林，GBDT，XGBoost 等）的表现比其他基准模型较优，因为它们可以捕捉特征之间的非线性关系。而 PJFNN-modified 的表现优于大部分的基准模型，能够在准确率上带来约 4%的提升。然而，所有模型的准确率均未超过 0.7，可能原因是与用户相关的信息较少，影响模型结果。同时也在一定程度上反映了 Person-Job Fit 任务的挑战性。

表 5 用户-工作匹配模型评价结果

	Accuracy	**Precision**	**Recall**	**F1-score**	**AUC**
线性回归	0.544	0.546	0.522	0.534	0.549
Logistic 回归	0.534	0.536	0.510	0.523	0.550
朴素贝叶斯	0.514	0.516	0.446	0.479	0.531
决策树	0.605	0.609	0.590	0.599	0.631
随机森林	0.637	0.634	0.647	0.640	0.702
AdaBoost	0.526	0.529	0.465	0.495	0.535
GBDT	0.629	0.630	0.624	0.628	0.663
XGBoost	0.607	0.606	0.615	0.610	0.614
PJFNN-m	**0.660**	**0.654**	**0.679**	**0.667**	**0.715**

六、任务二：Job Recommendation

Job Recommendation 是解决人才错配的关键措施，它的目标是为用户推荐合适的工作。Ramanath 等人(2018)提到 GBDT 模型在 LinkedIn 的工作推荐系统中承担了重要的作用，故本研究仍以（五）中提到的机器学习模型作为基准模型，探究 PJFNN-modified 在 Job Recommendation 任务中的表现。本研究选择了常见的 Hit Rate 作为算法的评价指标，它的具体定义为：

$$HR@N = \frac{n_{test}^{+}}{n_{test}}$$

其中n_{test}^{+}表示模型生成的 Top-N 推荐列表中用户申请的工作数量，n_{test}表示测试集中用户申请的工作数量。此处$N = 1, 5, 10, 20$。

各模型的设定与前一部分一致，评价结果如表 6 所示。从结果中可以看出，在 Person-Job Fit 任务中表现较差的模型，在工作推荐任务中大都表现不佳，除了朴素贝叶斯模型在 HR@10，HR@20 指标上有不错的表现。集成模型表现较优，GBDT 模型在 HR@1，HR@5 指标上优于其他基准模型，随机森林模型在 HR@20 指标上优于其他基准模型。最后，PJFNN-modified 在大多数情况下都优于基准模型，主要原因是 PJFNN-modified 在编码岗位特征时考虑了词汇的语义信息。

表 6 工作推荐模型评价结果

	$N = 1$	$N = 5$	$N = 10$	$N = 20$
线性回归	0.008	0.100	0.158	0.288
Logistic 回归	0.008	0.092	0.142	0.300
朴素贝叶斯	0.015	0.085	0.208	0.369
决策树	0.031	0.081	0.162	0.304
随机森林	0.027	0.123	0.227	0.431
AdaBoost	0.019	0.123	0.212	0.327
GBDT	0.038	**0.150**	0.235	0.404
XGBoost	0.027	0.104	0.196	0.377
PJFNN-m	**0.042**	**0.150**	**0.262**	**0.446**

七、结论

随着信息化和人工智能的发展，如何将海量的用户和岗位进行匹配受到越来越多关注。本研究以 Zhu 等人(2018)提出的模型为基础，结合真实数据特点设计了 PJFNN-modified 模型，并进行了充分实验，验证了 PJFNN-modified 的优越性。通过实验，本研究发现简单的机器学习模型表现不佳，但是一系列集成模型，如随机森林、XGBoost 有不错的表现。此外，PJFNN-modified 带来的提升有限，可能的原因是数据中关于用户的信息过少，深度学习模型无法充分发挥易于编码文本信息的优势。因此，在用户信息匮乏时，集成模型在效率和效果上都是不错的选择。

针对本研究所使用数据集用户信息缺失的情况，未来考虑通过工作名称正则化、主修专业正则化等手段来分析用户与工作的匹配情况，以解决不同工作描述中同一对象表述存在差异的问题。而当拥有用户简历信息，以及过往工作经历的详细描述时，有理由相信对文本内容编码能力更强的深度学习模型会取得更显著的提升。

参考文献

[1] John L Holland .Making vocational choices: A theory of careers[J]. Prentice Hall, Upper Saddle River, NJ, 1973

[2] Amabile T M.The Social Psychology of Creativity:A Componential Conceptualization[J].Journal of Personality and Social Psychology, 1983, 45 (2) :357-376.

[3] Cable D M, Derue D S.The Convergent and Discriminant Validity of Subjective Fit Perceptions[J].Journal of Applied Psychology, 2002, 87 (5) :875-883.

[4] Mamadou Diaby, Emmanuel Viennet, and Tristan Launay. Toward the next generation of recruitment tools: an online social network-based job recommender system[C]. InProceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013: 821–828.

[5] Yao Lu, Sandy El Helou, and Denis Gillet. A recommender system for job seeking and recruiting website[C]. InProceedings of the 22nd International Conference on World Wide Web. ACM, 2013: 963–966

[6] Ioannis Paparrizos, B Barla Cambazoglu, and Aristides Gionis. Machine learned job recommendation[C]. InProceedings of the fifth ACM Conference on Recommender Systems. ACM, 2011: 325–328

[7] Zhu C , Zhu H , Xiong H , et al. Person-Job Fit: Adapting the Right Talent for the Right Job with Joint Representation Learning[J]. ACM Transactions on Management Information Systems, 2018, 9(3):12.1-12.17.

[8] Zhou Q , Liao F , Chen C , et al. Job recommendation algorithm for graduates based on personalized preference[J]. CCF Transactions on Pervasive Computing and Interaction, 2019, 1(29).

[9] Chen Y , Mu Y , Wei Q , et al. A two-sided matching and diversity-enhanced method for job recommendation with employer behavioral data[C] 14th International FLINS Conference (FLINS 2020). 2020.

[10] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[J]. Advances in Neural Information Processing Systems,2013,26:3111-3119.

[11] Ramanath R, Inan H, Polatkan G, et al. Towards deep and representation learning for talent search at LinkedIn[C] Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 2253-2261.