职达

第九届中国大学生服务外包创新创业大赛

队伍成员：

曾永家右　黄泰北　赵辉

龙兰心　罗雪蕾

基于大数据的人岗匹配系统“职达”

个性化推荐算法

职达信息技术有限公司

达芬奇队

指导教师：徐进 尹帮旭

目录

一、概述

推荐的本质就是匹配，对于求职者进行岗位推荐，亦可以看作是人岗匹配，为此，我们采用基于内容的推荐算法（Content-based Recommendations）和神经网络相结合的方法，首先进行数据建模，确定岗位及求职者的描述维度，然后进行指标的量化，最后确定各指标的初始权重。

当此系统运行后，我们按照初始权重计算匹配度，并根据匹配度给出用户最佳岗位推荐。待用户浏览完岗位信息后，我们会邀请用户对我们的推荐准确率进行打分（分数为 1~10）。然后，我们根据用户的打分结果，以系统给出的匹配度与用户打分结果（转换为 0~1）的差的绝对值最小为目标，训练一个深度神经网络，以此确定各指标的最优权重。随着用户反馈结果的不断增加，各指标权重也在不断的优化，最终实现精准推荐

二、基于内容的推荐算法

2.1 指标筛选

通过对各类别简历信息和经爬虫获取的十余万条岗位信息进行观察分析比较，对匹配算法资料查询，以及根据对人岗匹配的信息特点进行重要性分析，最终将匹配指标根据人才素质、公司属性、职位属性、其它要素四个基本点进行分类，来进行指标信息的量化提取，纳入算法模型。

2.1.1 指标选取原则

科学完备的指标体系是匹配度统计结果质量的基础。由于人岗匹配度计算是一个多功能、多层次、多目标的一个评价对象，影响其评价的因素很多，所以首先应确定选取指标时的原则。

(1)科学性。指标要建立在科学分析的基础上，能够客观地反映人岗画像的本质特征及复杂性，每个指标必须概念清晰，各指标之间既要有内在联系，又要避免混淆；

(2)全面性。人岗匹配是一个涉及到人才基本素质、公司属性、职位具体属性等多个方面的复杂的综合系统，所以要求指标系统要具有综合性，能反映人才、公司、职位等对象的主要属性及相互关系，既能反映当前的局部特征，又能反映全局的长远的综合的特征，既有主观指标，也要有客观的指标。

(3)规范性。指标的选取应避免摄影不常用、难以统计的指标，做到指标的规范化，并使数据资料等容易获取，计算方法简单易行，且便于横向、纵向比较。

(4)开放性。人岗指标选择要允许随着求职者和公司要求的增多产生的一些新的元素指标随时进入指标体系，以保证指标体系的及时更新，达到科学、合理地反映当前背景下的匹配度状况。

2.2 指标选择

根据指标选取的原则，首先对人岗匹配度这一系统进行目标分析，划分目标层次，对总目标进行逐项逐级分解，直到目标能用具体、直观的数据指标表示为止，从而形成一个层次化、具有阶梯结构的科学指标体系。

2.2.1 初级指标体系的建立

将选取的指标整理分类，得到初级的评价指标体系如下表 2-1 所示：

一级指标	二级指标	三级指标
匹配度	人才素质	学历水平
		工作经验
		技能水平
	公司属性	公司规模
		公司性质
	职位属性	薪资
		工作地点
		工作性质
	其它要素	发布日期
		招聘人数

表 2-1 初级指标体系

2.3 指标遴选

2.3.1 可行性分析

对于指标的可行性要求，主要包括两个方面：

(1)指标数据便于收集与整理，指标体系中不应包含一些难以量化的指标，其数据来源不应难以统计、时效性差；

(2)与现行统计方法相衔接，尽可能的采用相对成熟与相对公认的指标，以保证评价结果的准确性与时效性。

2.3.2 重要性分析

对于指标的重要性分析，主要是基于求职时的指标特点，有一些因素具有可商榷性，有一些指标确属于硬性指标，若硬性指标不被匹配，人岗匹配度将直接为 0，不再纳入匹配范围。根据这种特征，系统将指标分为硬性的筛选指标和可商榷的匹配度计算指标。

根据对人才信息以及岗位招聘信息的比较以及各平台上所收集的反馈信息，我们将以下几项定为筛选指标：

(1)学历水平:专科、本科、硕士、博士

岗位的最低要求作为筛选值，若人才信息达到岗位的要求，则为符合要求，若求职者未达到岗位要求，则无法匹配。

(2) 公司性质：国企、私企、外企

在求职者有公司性质要求的情况下，作为筛选项，仅将满足条件的岗位加入匹配队列。

(3) 工作地点

根据就业的特殊性，岗位工作地点与求职者意向的工作地点是匹配的必要项。

(4) 工作性质

系统将工作性质分为全职、实习、兼职三类，由求职者选择自己的工作性质，来进行筛选。

经分析和筛选后的指标体系如表 2-2 所示：

一级指标	二级指标	三级指标
匹配度	人才素质	工作经验
		技能水平
	公司属性	公司规模
	职位属性	薪资
	其它要素	发布日期
		招聘人数

表 2-2 人岗画像匹配指标体系

2.4 指标量化

(1) 工作经验:总的工作年限-岗位要求工作年限

若所得结果大于等于 0，则指标值为 1；如计算结果小于 0，则指标取值为 0。

(2) 技能水平：求职者所掌握的技能点数占岗位所要求的技能点数的百分比。

(3) 公司规模：系统以公司人数 x 来对公司规模进行定义，针对互联网企业，x>=300, 大型企业，量化取值 1；100<=x<300，中型企业，量化取值 0.8；10<=x<100，小型企业，量化取值 0.5；x<10，微型企业，量化取值 0.2。

(4) 薪资

根据薪资有多种写法的多种表达方式，取求职者期望薪资为 x，职位提供薪资为 a~c，职位提供的薪资平均值为 b。

薪资量化值分以下三种情况：薪资面议，量化取值为 1;x>c ，量化取值为(x-b)/x；x<a，量化取值为 b/x。

(5)发布日期,根据岗位发布的日期距求职者查询的日期间隔长短来对这一指标进行量化。近一周：1；近两周：0.8；近一个月：0.5；两个月以上：0.2。

(6)招聘人数

本着同一岗位招聘人数越多，求职者在该岗位就业成功率越大的原则，系统利用以下原则对该指标进行量化。招聘人数 x>10，取值为 1；5<x<=10，取值为 0.8；5>=x，取值为 0.5。

2.5 指标初始权重的确定

建立模型的首要工作室确定各指标的权重。由于在该人岗画像匹配指标体系中，各三级指标的关联性都较小，且难以进行直接比较，故系统采用层次分析法的思想来进行指标赋权。层次分析法的一个重要特点就是用两两重要性程度之比的形式表示出两个方案的相应重要性程度等级。如对某一准则，对其下的各方案进行两两对比，并按其重要性程度评定等级。基于对本系统指标的分析，系统指标体系赋权如下：

一级指标	二级指标a_i	三级指标b_{ij}
匹配度	人才素质(0.4)	工作经验(0.3)
		技能水平(0.7)
	公司属性(0.15)	公司规模(1)
	职位属性(0.3)	薪资(1)
	其它要素(0.15)	发布日期(0.6)
		招聘人数(0.4)

表 2-3 人岗画像匹配指标体系权重设置

2.6 匹配度计算公式

根据求得权重可得以匹配度值为：

$$Z_n = \sum_{i=1,j=1}^{n} a_i b_{ij} x_{ij}$$

其中，Z_n为人岗匹配度的值。

三、神经网络

3.1 人工神经网络简介

人工神经网络(Artificial Neural Network, 简称 ANN)是由大量的、同时也是很简单的处理单元广泛连接而形成的网络系统。它最早开始于 1943 年 Meculloch 和 Pitts 提出的神经元的数学模型。它反映了人脑功能的许多基本特征, 是一个并行处理的非线性系统, 但它并不是人脑神经系统的真实写照, 而只是对人脑的行为作某些简化、抽象和模拟。

神经网络的学习算法有多种, 根据所研究问题的性质和神经网络的有关理论, 本算法采用 BP 神经网络的结构形式。

3.2 BP 神经网络

3.2.1 BP 算法简介

BP 网络是在 1985 年由 RumeChart 等人提出的反向传播算法的基础上发展起来的, 是一种多层次反馈型网络, 所使用的是有导师学习算法, 网络结构如图 3-1 所示

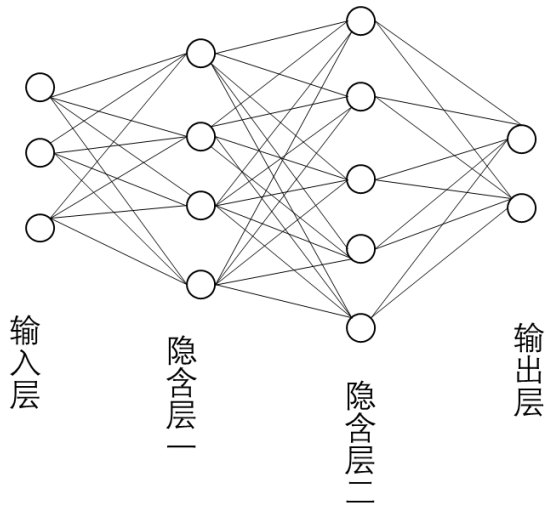

图 3-1

BP 算法由信号的正向传播和误差的反向传播两个过程组成。

正向传播时，输入样本从输入层进入网络，经隐层逐层传递至输出层，如果输出层的实际输出与期望输出(导师信号)不同，则转至误差反向传播；如果输出层的实际输出与期望输出(导师信号)相同，结束学习算法。

反向传播时，将输出误差(期望输出与实际输出之差)按原通路反传计算，通过隐层反向，直至输入层，在反传过程中将误差分摊给各层的各个单元，获得各层各单元的误差信号，并将其作为修正各单元权值的根据。这一计算过程使用梯度下降法完成，在不停地调整各层神经元的权值和阈值后，使误差信号减小到最低限度。

权值和阈值不断调整的过程, 就是网络的学习与训练过程, 经过信号正向传播与误差反向传播,

权值和阈值的调整反复进行，一直进行到预先设定的学习训练次数，或输出误差减小到允许的程度。

3.2.2 BP 算法流程图

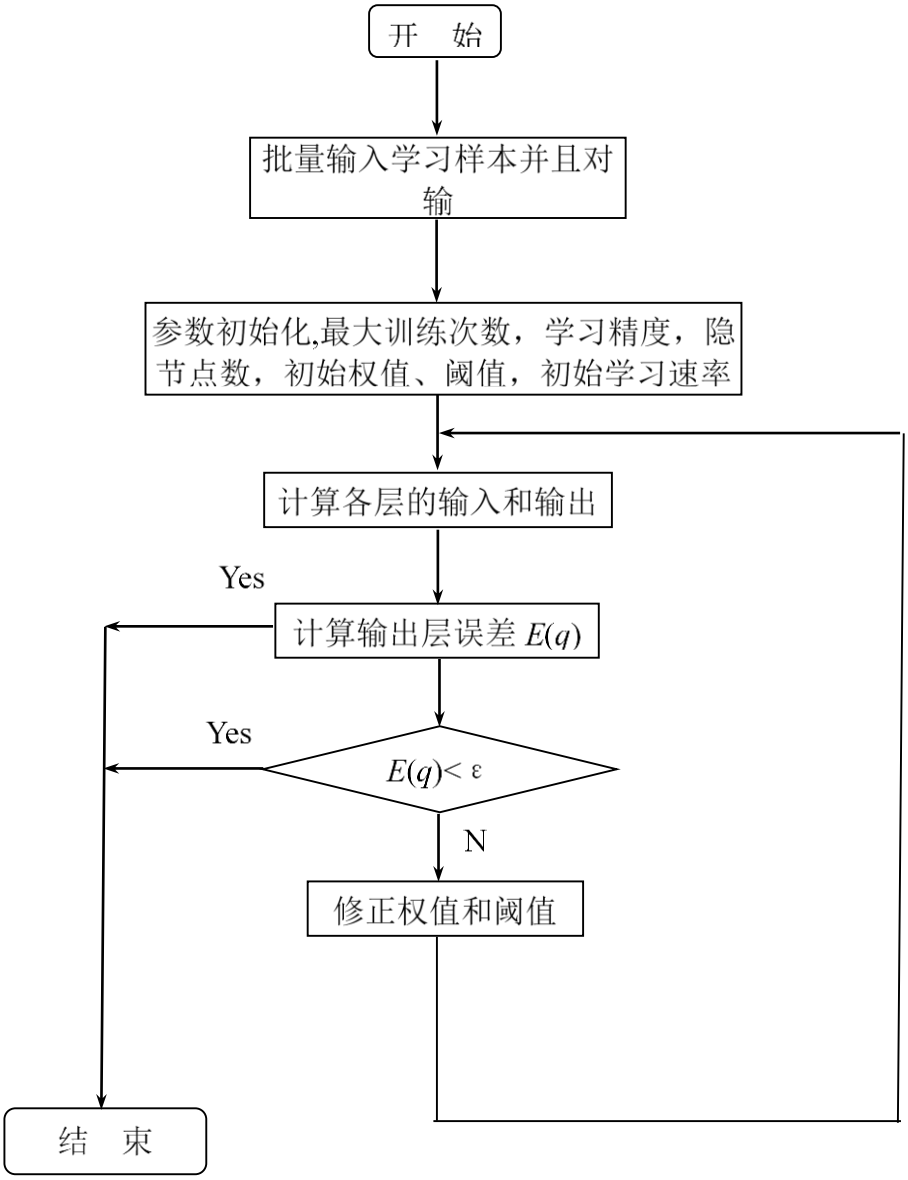

图 3-2

3.2.3 BP 算法的数学过程

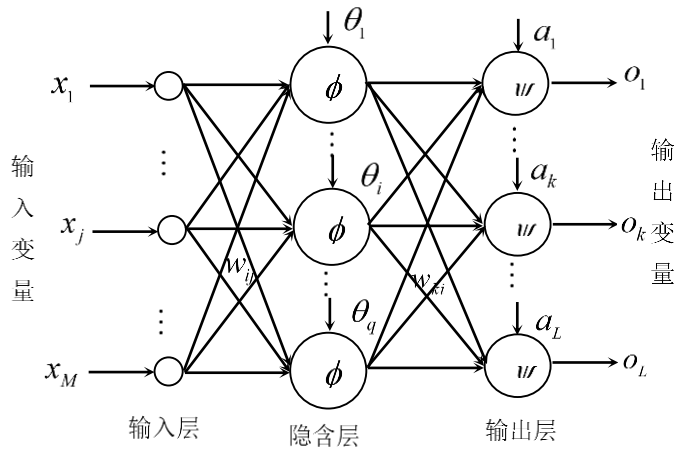

图 3-3

输出层中 M=14，隐含层中 q=9，学习率 η=0.01，L=1， ε =1e-10，最大训练次数为 5000 次.这个神经网络就只有一个输出神经元，所以每组样本值输出一个值。

一般取初始权值在（-1，1）之间的随机数，初始阈值取（0,1）之间的随机数。

w_{ij} 表示隐含层第 i 个节点到输入层第 j 个节点之间的权值；

θ_i 表示隐含层第 i 个节点的阈值；

$\phi(x)$ 表示隐含层的激励函数； $\phi(x)$ =1/(1+e^{-x})

w_{ki} 表示输出层第 k 个节点到隐含层第 i 个节点之间的权值， i=1,…,q；

a_k 表示输出层第 k 个节点的阈值,k=1,…,L；

$\psi(x)$ 表示输出层的激励函数， $\psi(x)$ =x；

o_k 表示输出层第 k 个节点的输出，k=1。

（1）信号的前向传播过程

隐含层第 i 个节点的输入 net_i：

$$net_i = \sum_{j=1}^{M} w_{ij} x_j + \theta_i \tag{3-1}$$

隐含层第 i 个节点的输出 y_i：

$$y_i = \phi(net_i) = \phi(\sum_{j=1}^{M} w_{ij} x_j + \theta_i) \tag{3-2}$$

输出层第 k 个节点的输入 net_k, k=1：

$$net_k = \sum_{i=1}^{q} w_{ki} y_i + a_k = \sum_{i=1}^{q} w_{ki} \phi(\sum_{j=1}^{M} w_{ij} x_j + \theta_i) + a_k \tag{3-3}$$

输出层第 k 个节点的输出 o_k , k=1：

$$o_k = \psi(net_k) = \psi(\sum_{i=1}^{q} w_{ki} y_i + a_k) = \psi\left(\sum_{i=1}^{q} w_{ki} \phi(\sum_{j=1}^{M} w_{ij} x_j + \theta_i) + a_k\right) \tag{3-4}$$

（2）误差的反向传播过程

误差的反向传播，即首先由输出层开始逐层计算各层神经元的输出误差，然后根据误差梯度下降法来调节各层的权值和阈值，使修改后的网络的最终输出能接近期望值。 T_k 为预期输出

对于每一个样本 p 的二次型误差准则函数为 E_p：

$$E_p = \frac{1}{2} \sum_{k=1}^{L} (T_k - o_k)^2 \tag{3-5}$$

系统对 P 个训练样本的总误差准则函数为：

$$E = \frac{1}{2} \sum_{p=1}^{P} \sum_{k=1}^{L} (T_k^p - o_k^p)^2 \tag{3-6}$$

根据误差梯度下降法依次修正输出层权值的修正量 Δw_{ki}，输出层阈值的修正量 Δa_k，隐含层权值的修正量 Δw_{ij}，隐含层阈值的修正量 $\Delta\theta_i$ 。

$$\Delta w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}}; \quad \Delta a_k = -\eta \frac{\partial E}{\partial a_k}; \quad \Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}; \quad \Delta\theta_i = -\eta \frac{\partial E}{\partial \theta_i} \tag{3-7}$$

输出层权值调整公式：

$$\Delta w_{ki} = -\eta \frac{\partial E}{\partial w_{ki}} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial w_{ki}} = -\eta \frac{\partial E}{\partial o_k} \frac{\partial o_k}{\partial net_k} \frac{\partial net_k}{\partial w_{ki}} \tag{3-8}$$

输出层阈值调整公式：

$$\Delta a_k = -\eta \frac{\partial E}{\partial a_k} = -\eta \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial a_k} = -\eta \frac{\partial E}{\partial o_k} \frac{\partial o_k}{\partial net_k} \frac{\partial net_k}{\partial a_k} \tag{3-9}$$

隐含层权值调整公式：

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = -\eta \frac{\partial E}{\partial net_i} \frac{\partial net_i}{\partial w_{ij}} = -\eta \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial net_i} \frac{\partial net_i}{\partial w_{ij}} \tag{3-10}$$

隐含层阈值调整公式：

$$\Delta \theta_i = -\eta \frac{\partial E}{\partial \theta_i} = -\eta \frac{\partial E}{\partial net_i} \frac{\partial net_i}{\partial \theta_i} = -\eta \frac{\partial E}{\partial y_i} \frac{\partial y_i}{\partial net_i} \frac{\partial net_i}{\partial \theta_i} \tag{3-11}$$

又因为：

$$\frac{\partial E}{\partial o_k} = -\sum_{p=1}^{P} \sum_{k=1}^{L} (T_k^{\ p} - o_k^{\ p}) \tag{3-12}$$

$$\frac{\partial net_k}{\partial w_{ki}} = y_i \quad , \quad \frac{\partial net_k}{\partial a_k} = 1 \quad , \quad \frac{\partial net_i}{\partial w_{ij}} = x_j \quad , \quad \frac{\partial net_i}{\partial \theta_i} = 1 \tag{3-13}$$

$$\frac{\partial E}{\partial y_i} = -\sum_{p=1}^{P} \sum_{k=1}^{L} (T_k^p - o_k^p) \cdot \psi'(net_k) \cdot w_{ki} \tag{3-14}$$

$$\frac{\partial y_i}{\partial net_i} = \phi'(net_i) \tag{3-15}$$

$$\frac{\partial o_k}{\partial net_k} = \psi'(net_k) \tag{3-16}$$

所以最后得到以下公式：

$$\Delta w_{ki} = \eta \sum_{p=1}^{P} \sum_{k=1}^{L} (T_k^{\ p} - o_k^{\ p}) \cdot \psi'(net_k) \cdot y_i \tag{3-17}$$

$$\Delta a_k = \eta \sum_{p=1}^{P} \sum_{k=1}^{L} (T_k{}^p - o_k{}^p) \cdot \psi'(net_k) \tag{3-18}$$

$$\Delta w_{ij} = \eta \sum_{p=1}^{P} \sum_{k=1}^{L} (T_k^p - o_k^p) \cdot \psi'(net_k) \cdot w_{ki} \cdot \phi'(net_i) \cdot x_j \tag{3-19}$$

$$\Delta \theta_i = \eta \sum_{p=1}^{P} \sum_{k=1}^{L} (T_k^p - o_k^p) \cdot \psi'(net_k) \cdot w_{ki} \cdot \phi'(net_i) \tag{3-20}$$

生成新的权值

$$w_{ki}(k+1) = w_{ki}(k) + \Delta w_{ki}(k)$$

$$w_{ij}(k+1) = w_{ij}(k) + \Delta w_{ij}(k)$$

$$\alpha_k(k+1) = \alpha_k(k) + \Delta \alpha_k(k)$$

$$\theta_i(k+1) = w_i(k) + \Delta \theta_i(k)$$

然后进入下一步的循环，

计算误差，直到误差 ε<1e-10 时为止，然后程序结束

当满足下列条件时候采用动量法调整权值

训练程序设计中采用动量法的判断条件为：

$$mc = \begin{cases} 0 & E(k) > E(k-1) * 1.04 \\ 0.95 & E(k) < E(k-1) \\ mc & \text{其它} \end{cases}$$

，E（k）为第 k 步误差平方和。

权值调整采用如下公式

$$\Delta w_{ki}(k+1)' = (1-mc)\Delta w_{ki}(k+1) + mc\Delta w_{ki}(k)$$

$$\Delta w_{ij}(k+1)' = (1-mc)\eta\Delta w_{ij}(k+1) + mc\Delta w_{ij}(k)$$

$$\Delta\alpha_i(k+1)' = (1-mc)\eta\Delta\alpha_i(k+1) + mc\Delta\alpha_i(k)$$

$$\Delta\theta_i(k+1)' = (1-mc)\eta\Delta\theta_i(k+1) + mc\Delta\theta_i(k)$$

其中 k 为训练次数，mc 为动量因子，一般取 0.9 左右。

3.3 神经网络的搭建

3.3.1 神经网络的架构

我们设置一个输入层、两个隐含层和一个输出层，其中输入层有 6 个神经元，隐含层一有 13 个神经元，隐含层二有 13 个神经元，输出层有 1 个神经元。激活函数我们均选择 Sigmoid 函数。

3.3.2 神经网络的搭建

这里，我们采用 Python 的 Keras 库来搭建 BP 神经网络。

四、参考文献

[1]付继娟.人与岗位匹配的国内外研究综述[J].武汉职业技术学院学报,2004(02):40-43.

[2]史东风. 基于岗位胜任力的石油企业中层管理者人岗匹配模型研究[D].西南石油大学,2011.

[3]张志宇,吕明丽,李从东.基于 BP 神经网络的人岗匹配测评模型的研究[J].天津大学学报(社会科学版),2010,12(05):390-395.

[4]袁珍珍,卢少华.BP 神经网络在人岗匹配度测算中的应用[J].武汉理工大学学报(信息与管理工程版),2010,32(03):515-518.

[5]徐锦阳,张高煜,王曼曦,楼焕钰,薛伟程,毛骧裕.招聘网站职位与简历的双向匹配相似度算法[J].信息技术,2016(08):43-46+51.

[6]吕太之,毕家钦.基于 Hadoop 平台的岗位分析和推荐系统的构建[J/OL].河北软件职业技术学院学报,2017(04)[2018-04-16].https://doi.org/10.13314/j.cnki.jhbsi.20171219.001.