文章编号：2096-1472(2017)-05-10-04

基于k-means和关联度分析的网络招聘信息数据挖掘

章 胤，赵文慧，包恒玥，李亚健，周克强
（燕山大学，河北 秦皇岛 066004）

摘 要：本文基于K-means算法对网络招聘数据进行聚类分析，并运用关联规则对大数据和IT行业进行关联预测。从分析结果可知，学历和经验直接影响薪资水平，且金融银行职业类型的平均薪资水平在所得分类中最高，同时也得到大数据和IT行业对学历要求较高，其占总体职业类型比例有增加趋势。

关键词：网络招聘；数据挖掘；聚类算法；关联度分析
中图分类号：TP311 **文献标识码**：A

Data Mining of Online Recruitment Information Based on K-Means and Correlation Analysis

ZHANG Yin,ZHAO Wenhui,BAO Hengyue,LI Yajian,ZHOU Keqiang
(*Yanshan University*,*Qinhuangdao* 066004,*China*)

Abstract: Based on the k-means algorithm,the paper conducts cluster analysis of the online recruitment data,and adopts the association rules to predict big data and the IT industry.The analysis results show that education background and work experience directly influences the salary,the average salary in the financial and banking industry is higher than all the other professions,the educational requirements are relatively high in the big data and IT industry,and there is a growing tendency of its proportion in all the professions.

Keywords:online recruitment;data mining;clustering algorithm;correlation analysis

1 引言(Introduction)

在“大数据时代”的背景下，网络招聘数据呈爆炸式增长，但是对于网络招聘数据隐含的信息挖掘方面，国内并没有太多的学者进行深度研究，其中原因有网络招聘虚假信息较多、信息杂乱、时效性弱[1]。本文在充分考虑诸多原因的前提下，重点着手于对招聘信息隐含信息的挖掘，并根据分析结果进行短期前景预测。与此同时归纳挖掘方法，在日后进行类似方向研究的过程中，只需要抓取相应的数据，应用相应算法即可得到预期结果，大大节省了同类问题的时间成本。

2 关联规则技术概述[2](Correlation analysis)

设$I=\{i_1,i_2,\cdots,i_m\}$是D中全体项组成的集合。设任务相关的数据D是数据库事务的集合，其中每个事务T是项的集合，使得$T\subseteq I$对应每一个交易有唯一的标识，记作TID。设X是一个I中项的集合，如果$X\subseteq T$，那么称交易T包含项集X。若X、Y为项集，$X\subset I,Y\subset I$,并且$X\cap Y=\Phi$,则形如$X\Rightarrow Y$的表达式称为关联规则。

支持度是指所有交易中X与Y同时出现的频率记为：

$$\begin{aligned}&\text{support}(X\Rightarrow Y)=P(X\cup Y)\\&=|\{T:X\cup Y\subseteq T,T\in D\}|/|D|\times 100\%\end{aligned}\quad(1)$$

支持度是对关联规则在整个事务数据库中的统计重要性的衡量，它说明了该规则在所有事务中有着多大的代表性，支持度越大，规则就越重要。

置信度是指所有出现了X的活动中出现Y的频率记为：

$$\begin{aligned}&\text{support}(X\Rightarrow Y)=P(X\cup Y)\\&=|\{T:X\cup Y\subseteq T,T\in D\}|/|D|\times 100\%\end{aligned}\quad(2)$$

置信度是一个对关联规则准确性的衡量指标，有些关联规则置信度高，但是支持度却很低，这表示该规则实用的机

基金项目：河北省自然科学基金项目(A2015203121)，燕山大学国家级大学生创新项目(201610216027).

www.rjgcz.com

会很小，一般而言并不重要。

本文使用SPSS Modeler软件作为关联度分析相关工作的首选工具。

3 数据挖掘过程(Data mining process)

3.1 数据抓取

本文在国内三大招聘网站(智联招聘、前程无忧网、中华英才网)上抓取了一线城市北京、上海、广州(简称，“北上广”)，二线城市大连、长沙、福州(简称，“大长福”)六个城市的招聘信息，时间跨度为三个季度共九个月，这样选取数据的好处在于分析的时候一方面可以根据一二线城市进行相关度分析，又可以从南到北的地域上职业类型差距进行相关度分析。相对全国数据来说，具有很好的代表性。

3.2 数据预处理

由于招聘信息的纷呈复杂，所以我们需要进行数据的预处理，目的是得到具有较为规范格式的实验数据。

首先进行招聘信息的去重去空处理，然后将处理好的招聘数据按照时间序列进行统一安放在Excel表格文档中。其次我们需要建立一个包含职业类型特征词的词袋，词袋是作为文本信息聚类的标准，所以词袋的准确程度直接决定了文本聚类效果的好坏，我们将招聘网站上职业分类运用jieba中文分词工具进行分词处理，经过剔除无效数据和重复数据后，得到具有较好区分度的词袋文档。

3.3 文本向量化

由于计算机并不能够直接读取文本信息，所以我们需要将文本数据向量化，表达成计算机能够直接处理的数字形式。我们选用的文本向量化的模型是向量空间模型(VSM)[3]，文本向量化的准备工作是将词袋特征词进行权重赋值，权重赋值的方法是根据TF-IDF算法[4]的理论应用得到，具体使用公式和方法如下：

$$\text{逆向文本频率}(\text{IDF})=\log\left(\frac{\text{数据总体的文本数}}{\text{包含该词的文本数}+1}\right) \tag{3}$$

$$\text{权重}=\frac{\text{该特征词的IDF值}}{\text{全部特征词IDF值总和}} \tag{4}$$

然后用特征词权重做作为文本向量的维度赋值，从而得到各个实验数据的空间向量。

3.4 聚类分析

在完成文本信息向量化处理后，使用统计分析软件SAS中IML模块进行矩阵化运算[5]，通过对K-means聚类算法的研究和应用[6,7]，得到关于职业类型的聚类结果，我们对经过聚类分析的数据进行挖掘，便可以得到网络招聘信息中关于职业类型、薪资、地域、学历和工作经验的知识模式。

4 结果分析和预测(Interpretation of result and prediction)

4.1 对整体数据进行分析和挖掘

(1)通过北上广职业分布和大长福职业分布词云图(图1和图2)对比可以得出：

相同点：销售行业在两类城市占比最高，依此可以推断出在全国范围内，销售行业仍然占据主流招聘人群。

不同点：一线城市中除销售相关职业外，“PHP”“.NET”“JAVA”“UI设计师”“iOS”“C++”等词出现频数较大，直接说明了大数据和IT相关职业在一线城市的兴起和热门，而二线城市各个行业需求较销售都比较小，借此推断高端技术的需求量和城市类型有一定关系。

图1 北上广职业分布词云图

Fig.1 The first position word cloud

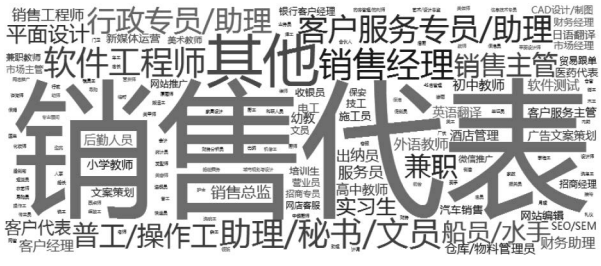

图2 大长福职业分布词云图

Fig.2 The second position's word cloud

(2)对比两类城市的学历、经验分别与平均薪资的分布图(图3和图4)可以看出，无论是哪类城市，随着学历的提高，薪资水平提高明显；随着工作经验的丰富，薪资水平同样提高

www.rjgcz.com

明显。

所以我们有理由得出：工作经验和学历都与薪资水平呈正相关系，即随着学历的提高和经验的丰富，薪资水平提高明显。

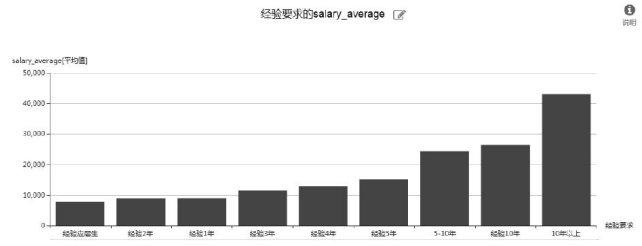

图3 总体经验要求的平均薪资分布图

Fig.3 The experience and average salaries' distribution diagram

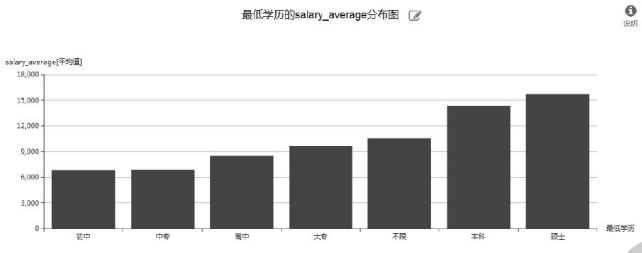

图4 总体最低学历的平均薪资分布图

Fig.4 The education and average salaries' distribution diagram

(3)通过聚类结果，得到关于职业类型和对应的平均薪资待遇的表格(表1)。

表1 各职业类型平均薪资

Tab.1 Different professional types of average salary

职业类型	传媒艺术	制造生产	销售市场	医师服务业	互联网通讯	房产建筑	法律教育	财务人力	金融银行
平均薪资	6584.7元	7949.4元	7267.8元	5534.4元	8919.5元	7611.8元	7487.0元	4302.5元	11862元

通过对上述平均薪资状况的对比得知：金融银行业相应职位平均薪资最高，其次是要求专业技术较高的互联网通讯行业平均薪资条件优秀，其余产业薪资水平接近，但是鉴于职位供应地区是一二线城市。所以数据显示薪资水平与实际相比较为合理。

4.2 大数据和IT行业的预测

鉴于近年来大数据和IT行业的崛起，高精尖产业和职业的兴起已经成为了当今社会的主流走向，所以对这类行业进行分析更有价值。

(1)通过对这类专业招聘信息和总体招聘信息对比，得出大数据相关职业所占百分比条形图(图5)。通过对图表信息的解读，我们可以得到三点结论：

大数据相关职业招聘比例在2016年9月开始快速增长，到2017年2月开始放缓增长速度，但是仍然以一种放缓的趋势增长。

根据上图所示，大数据相关职业招聘比例最大的三个月份是2016年8月、2017年3月和2017年4月，结合实际分析，可以得出每年的这两个季度是大学生毕业求职的高峰期，也是传统意义上的秋招和春招，侧面说明了大数据行业对于学历要求和能力要求较高。所以集中大学生毕业求职期间发布招聘信息。

对比2016年8月和2017年4月数据可以得出，大数据行业正在逐渐增长所占比重。因此可以认为短期内，大数据相关行业所占比重在未来短期内会持续上升。

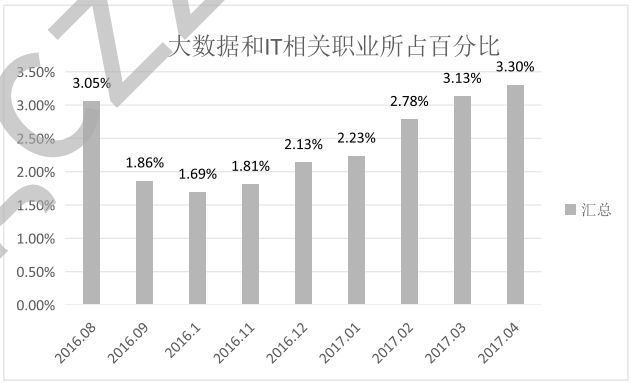

图5 大数据和IT相关职业占总体职业的比例

Fig.5 The proportion of the big data and IT in all positions

(2)我们针对从整体数据中筛选出的大数据和IT行业的招聘信息数据的分析，得到关于大数据和IT行业的学历与经验要求柱形图(图6和图7)，从图像中我们可以得到如下结论：

根据图6得出，大数据和IT行业需求学历更高，半数以上的最低学历要求是本科，专科及以上学历占据绝大部分比例。说明这个行业对人才的要求更高，所以相对应聘者来说除专业适合以外，学历要求门槛也较高。

根据图像可以看出，经验要求大部分集中于经验三年以下和不限经验。参考社会大背景下，大数据和IT行业较传统工商业起步晚，说明现有供给量已经无法满足日益增长的需

www.rjgcz.com

求量，基于此原因出现了经验要求低的现状。

结合以上两点，我们大胆预测市场急需关于大数据和IT行业的人才，正处于求大于供的阶段，而且学历越高、经验越多的人才薪资待遇越好。

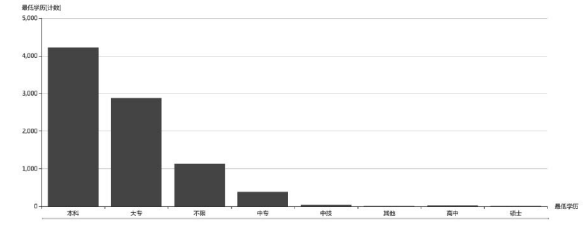

图6 大数据和IT行业最低学历所占比例

Fig.6 The big data and IT's education distribution

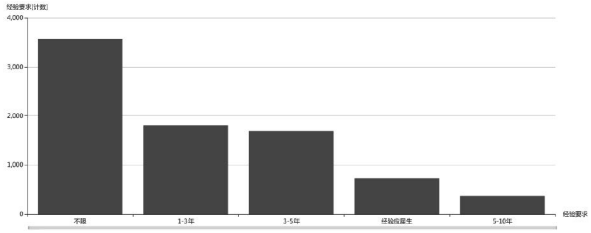

图7 大数据和IT行业经验要求所占比例

Fig.7 The big data and IT's experience distribution

(3)通过对编码后各指标之间的关联规则进行挖掘，结果显示图如图8所示。

其中编码对应文本信息为：

c1=北京；c2=上海；c3=广州；c4=大连；c5=长沙；c6=福州。

s1=0-5k；s2=5-10k；s3=10-15k；s4=15-20k；s5=20-25k；s6=25-30k；s7=30k以上。

e1=高中以下；e2=大专以下；e3=大专；e4=本科；e5=硕士；e6=不限；e7=其他。

w1=一年以下；w2=1-3年；w3=4-5年；w4=5年以上；w5=其他。

后项	前项	支持度百分比	置信度百分比
salary = s1	city = c5 jingyan = w1 education = e3	4.611	95.541
salary = s1	city = c5 jingyan = w1	6.109	91.827
education = e4	salary = s5 jingyan = w3 city = c1	4.523	85.714
education = e4	salary = s5 jingyan = w3	4.64	84.81
education = e4	salary = s6	2.173	83.784
education = e4	salary = s6 city = c1	2.144	83.562
education = e4	salary = s5 city = c1	7.548	80.545

图8 关联分析部分结果

Fig.8 The part of correlation analysis

通过图8的结果表示，在所有的大数据和IT相关职位中，存在的关联规则如下：

(1)如果一个企业在长沙，且要求学历是大专，工作经验一年以下，那么这家企业95.54%的概率提供平均薪资为0—5k。

(2)如果一个在北京的企业提供平均薪资为20k—25k，且要求工作经验是4—5年的人才，则有85.7%的概率需要最低学历为本科。

5 结论(Conclusion)

数据挖掘是一项综合技术，熟悉运用数据挖掘技术可以得到许多有价值的信息，通过对网络招聘信息的分析与挖掘可以得到学历和经验直接影响薪资水平，各类型职业的平均薪资状况，而且有针对性的对大数据和IT行业的行业现状分析，得出了相关行业有较大缺口，正处于求大于供的阶段。通过对大数据和IT相关信息的挖掘，得出在置信度很高的关联规则下的解读信息。本文针对大数据和IT行业着重进行分析和挖掘，并没有得到所有职业类型数据信息，这是本文存在的不足之处，同样也是我们下一步的研究方向。

参考文献(References)

[1] 赵鹏.企业网络招聘的现状与对策研究[J].人力资源,2016,11:145.

[2] Benites,et al.Evaluation of Hierarchical Interestingness Measures for Mining Pairwise Generalized Association Rules[J].IEEE TRANSACTIONSON KNOWLEDGE AND DATA ENGI NEERING,2014,26(12):3014-3015.

[3] Erkens,et al.Improving collaborative learning in the classroom:Text mining based grouping and representing[J]. International Journal of Computer Supported Collaborative learning,2016,11(4):389-391.

[4] 田瑞.针对特定主题的短文本向量化[J].软件,2012,33(11):202-203.

[5] 张晓冉.统计分析及其SAS实现[M].北京:清华大学出版社,2011.

[6] Chen,et al.Discriminative Hierarchical K-Means Tree for Large-Scale Image Classification[J].IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS,2015,26(9):2200-2202.

www.rigzz.com

[7] Liu,X&Li,M.A Predictive Fault Diagnose Method of Wind Turbine Based on K-Means Clustering and Neural Networks[J].JOURNAL OF INTERNET TECHNOLOGY,2016,17(7):1521-1528.

作者简介：

章　胤(1978-)，男，硕士，讲师.研究领域：微分方程数值解，数学建模.

赵文慧(1996-)，女，本科生.研究领域：应用统计和大数据分析.

包恒玥(1995-)，女，本科生.研究领域：应用统计和大数据分析.

李亚健(1995-)，男，本科生.研究领域：应用统计和大数据分析.

周克强(1995-)，男，本科生.研究领域：应用统计和大数据分析.

(上接第29页)

表1给出了基于灰度标准差、空间频率信息熵加权、清晰度加权质量评价结果，测试图像为图5四种算法的拼接结果。测试结果中可以看出，本文拼接算法拼接后的图像灰度标准差、空间频率信息熵加权、清晰度加权值相比其他三种算法均有所提高，而灰度标准差值、空间频率值越大说明融合图像的灰度分布越分散，信息熵加权、清晰度加权值越大则表明融合后的图像过渡更加平滑，融合算法方案更优，同时，重影拼接的实时性较好。

表1 测试结果

Tab.1 Result of test

融合方法 \ 评价方法	MSE	SF	MQ_m	VQ_m	融合时间(ms)
加权融合	124.1156	18.3447	0.5534	0.5788	101.34
动态规划	123.2813	20.6384	0.6456	0.6643	235.42
文献[11]	124.1317	18.3564	0.7843	0.7964	131.11
本文算法	124.3582	20.6994	0.8153	0.8278	140.81

5 结论(Conclusion)

本文在研究图像拼接中的融合方法的基础上，提出了基于改进加权融合算法实现图像拼接，从而消除图像拼接中产生的重影。该方法通过求取差异图像确定运动目标区域，并基于其边界求取加权系数，实验表明该融合算法能自适应的消除运动目标的干扰，灰度标准差、信息熵加权、清晰度加权评价指标相比其他融合拼接算法均有所提高，即融合图像的质量与清晰度得到改善。同时，提高了图像去重影拼接的实时性。

参考文献(Reference)

[1] 祝继华,等.基于图像配准的栅格地图拼接方法[J].自动化学报,2015,41(2):285-294.

[2] 桂振文,等.一种融合多传感器信息的移动图像识别方法[J].自动化学报,2015,41(8):1395-1404.

[3] 潘荣江,等.一种基于LCS的物体碎片自动拼接方法[J].计算机学报,2005,28(3):350-356.

[4] 贺行政,康荣学,任贵文.安全生产监控中的红外图像融合技术研究[J].中国安全生产科学技术,2016,12(S1):268-271.

[5] 雷飞,等.基于改进SURF的实时视频拼接方法[J].计算机技术与发展,2015,25(3):32-35.

[6] 漆昇翔,等.面向机载应用的多传感器图像融合技术综述[J].航空电子技术,2016(14):5-11.

[7] Richard Szeliski,Heung-Yeung Shum.Creating full view panoramic image mosaics and environment maps[C].In:Computer Graghics Procceding,Annual Conference Series,ACM SIGGRAPH,Los Angeles,California,1997:251-258.

[8] H Nicolas. New Methods for Dynamic Mosaicking[J].IEEE Transactions on Image Processing,2011,10(8):1239-1251.

[9] 葛仕明,等.基于梯度场的拼接缝消除方法[J].计算机辅助设计与图形学学报,2007,19(2):227-232.

[10] Smith S M,Brady J M.SUSAN-a new approach to low level image processing[J].International Journal of Computer Vision,1997,23(9):45-78.

[11] Yang Y,et al.A research on registration and de-ghosting algorithm in image mosaic[C].Wireless Communications & Signal Processing (WCSP),2013 International Conference on.IEEE,2013:1-5.

[12] 于大海,等.图像灰度标准差对介质阻挡丝状放电分布均匀性的定量评价[J].中国电机工程学报,2013,33(28):193-196.

[13] 姚汉英,等.SAR图像拼接效果的评价方法研究[J].计算机工程,2010,36(15):196-198.

作者简介：

徐　敏(1988-)，女，硕士，助教.研究领域：数字图像处理.

www.rigcz.com