Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2019, 8(2), 62-67
Published Online May 2019 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2019.82008

Hans汉斯

Visual Analysis and Prediction of Graduate Demand Based on Apriori Algorithm

Qingzhen Wang, Xiao Ju

Department of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou Henan
Email: weixingji123@163.com

Received: Apr. 8th, 2019; accepted: Apr. 29th, 2019; published: May 6th, 2019

Abstract

In order to quickly establish a good supply and demand relationship between enterprises and college students, in the general environment of increasing graduates' employment difficulties, we are using a network platform to collect information on the basic supply and demand of graduates and the data collected for visual analysis and prediction research. Firstly, the data is extracted from the network platform, and Excel is used to sort out the graph. Then the task of data mining is determined; the data of chart is preprocessed; and the data is analyzed based on Apriori algorithm. Finally, the analysis and prediction are completed by Java program.

Keywords

Graduate, Apriori Algorithm, Visualized Analysis

基于Apriori算法的毕业生需求状况的可视化分析和预测

王清珍，巨　筱

郑州科技学院，信息工程学院，河南　郑州
Email: weixingji123@163.com

收稿日期：2019年4月8日；录用日期：2019年4月29日；发布日期：2019年5月6日

摘　要

在毕业生增多、就业困难的大环境下，为了快捷建立企业与大学生良好供求关系，现采用网络平台搜集毕业生基本供求信息和企业的需求信息数据，并对收集的数据进行可视化分析与预测研究。首先从网络平台中提取数据，运用Excel归类整理，使图表呈现。然后确定数据挖掘的任务，对图表数据进行预处理，基于Apriori算法分析数据，最后通过Java程序完成分析和预测。

文章引用: 王清珍, 巨筱. 基于 Apriori 算法的毕业生需求状况的可视化分析和预测[J]. 人工智能与机器人研究, 2019, 8(2): 62-67. DOI: 10.12677/airr.2019.82008

关键词

毕业生，Apriori算法，可视化分析

1. 引言

互联网的快速发展与传播，使人们体会到了数据的庞大和资源的丰富。面对如此巨大的数据资源，如何从众多数据中发现隐藏的，对我们研究分析有用的数据信息，成为了一大难题。合理的分析毕业生资源总量、层次、专业结构，运用可视化技术对海量异构研究数据进行科学文献分析，如科学知识图谱、学科知识地图、科学知识网络(或学科知识网络、领域知识网络)等[1]，得出科学的预测模型，为高校教育的工作者要开拓思路，转变高校毕业生就业观念，提高就业竞争力。我们的研究就是运用可视化从数据集中将需要的数据呈现在人们眼前。毕业生需求状况可视化分析与预测研究具有重要意义。

2. 研究概述

本设计系统分为收集数据、分析呈现数据和可视化分析预测三部分。

2.1. 收集数据

通过 Excel 收集毕业生信息和企业需求信息。

学生编号	年龄	身高	专业成绩	薪资	外语	健康状况	工作时长
1	20	166	90	3000	0	H	8
2	22	168	58	6000	4	S	12
3	23	178	68	10000	6	H	8
4	25	182	94	12000	0	H	8
5	19	192	75	11000	0	H	10
6	22	176	72	8000	6	H	10
7	24	186	61	4000	4	S	12
8	24	162	44	7000	4	S	8
9	23	187	58	6000	6	H	10
10	21	185	83	7000	0	H	12

Figure 1. Information thumbnails for graduates
图 1. 毕业生信息缩略图

企业编号	年龄	身高	专业成绩	薪资	外语	健康状况	工作时长
T100	20-25	160-170	0	3000	0	H	8
T200	18-25	160-171	0	5000	0	H	8
T300	22-25	170-177	10	8000	0	H	8
T400	20-22	170-175	25	10000	0	H	8
T500	23-24	180-190	50	12000	0	H	12
T600	24-25	160-170	65	13000	0	H	12
T700	24-26	160-170	80	5000	4	H	12
T800	19-20	180-190	70	6000	0	H	8
T900	23-24	165-175	20	7000	0	H	8
T1000	24-26	160-190	65	5000	8	S	12

Figure 2. Thumbnail of enterprise demand information
图 2. 企业需求信息缩略图

2.2. 分析呈现数据

本文使用的数据来源是名为“student”的 excel 文件中的“源数据” [2]。图 1、图 2 为缩略图，企业需求数据一共 100 行，九个属性，毕业生信息共 100 行，12 个属性。

首先对源数据进行预处理，清除掉一些无关事项，在本次数据中包含属性，身高、健康状况。

数据预处理的第二步：使用 dm = xlsread('shen')；导入“student”.xls 文件，在 Matlab 中对一些连续数据离散化。

如下：

1、[18,20]1，[21,23]=2，[24,26]=3

2、[0,33]=1，[34,66]=2，[67,100]=3

3、[0,3000]=1，[3000,4500]=2，[4500,6000]=3

4、[0]=1，[4]=2，[8]=3

5、[0,8]=1，[8,10]=2，[10,12]=3

生成以下图表：

企业 ID	
企业属性	1：世界五百强，2：全国五百强，3：普通国企，4：普通私企
企业地址(总部)	1：一线城市，2：二线城市，3：海外
企业要求专业成绩	1：前 5%，2：前 20%，3：前 50%，4：无要求
企业要求工作时长	1：8 h，2：12 h
企业提供薪资水平	1：0~3 k，2：3~5 k，3：5~7 k，4：7~10 k，5：10 k+
企业要求年龄	1：20~22，2：22~26，3：26+
毕业生外语水平	1：六级，2：四级，3：两者都不是

毕业生 ID	
毕业生高校属性	1：985 高校，2：211 非 985 高校，3：普通一本高校，4：二本高校
毕业生专业属性	1：理工专业，2：经济管理专业，3：文史专业
毕业生专业成绩水平	1：前 5%，2：前 20%，3：前 50%，4：50%以后
毕业生外语水平	1：六级，2：四级，3：两者都不是
毕业生家庭经济情况	1：优越，2：中等、一般，3：较差
毕业生生源地情况	1：城市，2：农村
毕业生期望薪资水平	1：0~3 k，2：3~5 k，3：5~7 k，4：7~10 k，5：10 k+
毕业生期望工作城市	1：一线城市，2：非一线城市
企业要求年龄	1：20~22，2：22~26，3：26+
企业要求工作时长	1：8 h，2：12 h

然后对不同数据的相同数据进行整合：

由于不同属性之间的属性值存在相同情况，所以利用下面语句对一共 2 个条件属性中的 100 个数据进行如下赋值，使每条属性唯一确定。从而得到个 200 唯一数据，只不过 200 个里面有且只能出现 100 个。

程序如下：

```
m1=[8,8,10,12,12,8,8,8,12,10];k=1;w=m1(k);dm3=dm2;
for i=1:2
    dm3(i)=dm2(i)+w;
    if rem(i,303)==0
        k=k+1;
        w=w+m1(k);
    end
end
```

从而得到 dm(3)矩阵。而且决策属性分为 1：专业成绩；2：工作时长

2.3. 可视化分析预测

Apriori 算法[3]是常用的用于挖掘出数据关联规则的算法，它用来找出数据值中频繁出现的数据集合，找出这些集合的模式有助于我们做一些决策。常用的频繁项集的评估标准有支持度，置信度和提升度三个。

支持度就是几个关联的数据在数据集中出现的次数占总数据集的比重。或者说几个数据关联出现的概率。如果我们有两个想分析关联性的数据 X 和 Y，则对应的支持度为：

$$\text{Support}(X,Y)=P(X,Y)=\text{number}(XY)/\text{num}(\text{AllSamples}) \tag{1}$$

以此类推，如果我们有三个想分析关联性的数据 X，Y 和 Z，则对应的支持度为:

$$\text{Support}(X,Y,Z)=P(X,Y,Z)=\text{number}(XYZ)/\text{num}(\text{AllSamples}) \tag{2}$$

一般来说，支持度高的数据不一定构成频繁项集，但是支持度太低的数据肯定不构成频繁项集。

置信度体现了一个数据出现后，另一个数据出现的概率，或者说数据的条件概率。如果我们有两个想分析关联性的数据 X 和 Y，X 对 Y 的置信度为：

$$\text{Confidence}(X\Leftarrow Y)=P(Z|Y)=P(XY)/P(Y) \tag{3}$$

也可以以此类推到多个数据的关联置信度，比如对于三个数据 X，Y，Z,则 X 对于 Y 和 Z 的置信度为：

$$\text{Confidence}(X\Leftarrow YZ)=P(X|YZ)=P(XYZ)/P(YZ) \tag{4}$$

在毕业生需求数据中，通信工程专业对应电子类专业的置信度为 40%，支持度为 1%。则意味着在电子行业数据中，总共有 1%的用人单位既需要通信工程专业的人才又需要电子类专业的人才；同时需要电子类专业人才的用人单位中有 40%的用人单位需要通信工程专业学生。

提升度表示含有 Y 的条件下，同时含有 X 的概率，与 X 总体发生的概率之比，即：

$$\text{Lift}(X\Leftarrow Y)=P(X|Y)/P(X)=\text{Confidence}(X\Leftarrow Y)/P(X) \tag{5}$$

提升度体先了 X 和 Y 之间的关联关系，提升度大于 1 则 X⇐Y 是有效的强关联规则，提升度小于等于 1 则 X⇐Y 是无效的强关联规则。一个特殊的情况，如果 X 和 Y 独立，则有 Lift(X⇐Y)=1，因为此时 P(X|Y)=P(X)。

一般来说，要选择一个数据集合中的频繁数据集，则需要自定义评估标准[4]。最常用的评估标准是用自定义的支持度，或者是自定义支持度和置信度的一个组合。

运行 Apriori 算法实现关联规则，需要借助 JAVA 程序，编写程序后进入关联规则主界面，如图 3 所示。

Figure 3. Main interface of association rules
图 3. 关联规则主界面

3. 结束语

本次分析用的数据主要有两方面：毕业生基本信息和招聘企业基本信息和需求数据，对毕业生信息使用 Apriori 算法时，每个毕业生作为一个项集，这里项集的特征是每个项集包含的 item 个数是一样的，而 item 是所有毕业生属性的所有情况的集合。

利用 Apriori 算法，以 5%为支持度阈值得到毕业生信息的频繁项集及其支持度为：

频繁项集	支持度
985 高校，理工专业，22~26，六级，家境一般，12 h，农村，7~10 k，一线城市	12.5%
985 高校，经济管理专业，22~26，六级，家境一般，8 h，农村，7~10 k，一线城市	11.0%
985 高校，理工专业，22~26，六级，家境优越，8 h，城市，7~10 k，一线城市	8.5%
普通一本高校，文史专业，20-22，四级，家境一般，12 h，农村，3~5 k，一线城市	6.0%
二本高校，理工专业，26+，四级，家境优越，8 h，城市，3~5 k，二线城市	5.5%

再对上述频繁项集进行关联规则处理得到以下置信度 60%以上的规则(已经经过规则合并)：

985 高校，六级，农村　　7~10 k，一线城市
985 高校，六级，城市　　5~7 k，二线城市
普通一本高校，文史专业，四级　　3~5 k，一线城市
二本高校，理工专业　　3~5 k，二线城市

经过分析，可以得到，以下结论：

1) 985 高校毕业生对薪资要求较高，更倾向于一线城市
2) 专业不是影响就业的主要因素

3) 城市生源的毕业生倾向于二线城市，对薪资最低接受限度较低

对企业信息进行类似处理，得到以下置信度60%以上的规则[5] (已经过合并)：

一线城市，世界五百强，10 k+　　985 高校，六级

海外城市，世界五百强，7~10 k　　985 高校，六级

二线城市，普通国企，5~10 k　　211 高校，四级

一线城市，普通私企，5~10 k　　211 高校

经过分析，可以得到以下结论：

1) 五百强企业主要分布在一线及海外城市，且工资普遍较高。最看重毕业生的条件是毕业院校和外语水平，要求偏高。

2) 国企和私企大多分布在二线和一线城市，对毕业生的要求院校依旧偏高，但不高于五百强企业，对外语水平的要求比较宽松。

3) 家庭条件不是影响就业的主要因素。

结合两者分析，可得：

1) 毕业生和企业在高校、外语等方面的供需关系是比较平衡的。

2) 毕业生和企业在薪资方面有一定矛盾。毕业生需求较高，企业提供的较低。

基金项目

2018 年度河南省大中专院校就业创业研究立项课题 JYB2018074。

参考文献

[1] 常大俊. 基于数据仓库和 OLAP 的决策技术研究[D]: [硕士学位论文]. 长春: 长春理工大学, 2009.

[2] Shih, Y.S. (1999) Families of Splitting Criteria for Classification Trees. Statistics and Computing.

[3] 韩天鹏, 宋中山. Apriori 算法的改进[J]. 电脑知识与技术(学术交流), 2007.

[4] 谢俏丽. 基于组合预测模型的湖北省卫生人力资源需求预测研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2016.

[5] 陈必坤, 赵蓉英. 学科知识可视化分析的理伭研究[J]. 情报理论与实践, 2015, 38(1): 23-29.