DOI:10.11721/cqnuj20140424

基于 Weka 的就业数据分析和模式挖掘*
——以重庆市信管专业为例

朱 德 利
(重庆师范大学 计算机与信息科学学院, 重庆 401331)

摘要:在 Weka 平台下使用了时间序列预测算法、朴素贝叶斯聚类和 Apriori 关联规则算法分析真实就业数据,得到如下结论:该专业各年的男女比例总体较为平衡,很大程度上是因为该专业在重庆市的 5 所高校中既招收文科生,又招收理科生;女性党员学生且没有工作经验与"考取研究生"这一状态关联度较高;企业是信管专业的最主要的就业领域;获奖次数多的学生选择传统就业领域的概率比较大;该专业中、东部的就业信息向重庆市传递还不够充分。文章还分析了产生这些结论的 2 个原因:一是图书情报部门已经不能形成一个信息管理与信息系统专业独特的就业类型,但社会对该专业产生了新的需求;二是实际就业环境与专业定位设计存在差距。

关键词:信息管理与信息系统;就业;数据挖掘
中图分类号:TP311.13　　**文献标志码**:A　　**文章编号**:1672-6693(2014)04-0120-06

信息管理与信息系统专业作为一门将信息技术同管理学科相结合而产生的专业,在国家机关、企事业单位中起着举足轻重的作用。但是,在学生就业过程中往往因为各种因素导致就业困难。本文试图对近几年重庆市该专业学生就业真实情况的数据进行分析和挖掘,展现本专业就业全景,分析专业的就业瓶颈,为该专业的专业设计、教育教学、职业规划等提供数据支撑。

1 数据来源

在本研究中,笔者主要收集了重庆市 5 所高校(重庆师范大学、西南大学、重庆大学、重庆工商大学、重庆理工大学)2004—2011 届信息管理与信息系统专业毕业生的就业信息,这些就业信息来自各个学校的就业指导中心,数据真实可靠,同时该数据来自不同的学校,数据具有广泛性和实用性。收集的数据信息中,主要由姓名、性别、在校任职、获奖次数、单位性质、专业排名、单位名称等组成。表 1 展现的是部分数据的包含部分属性值的数据信息。

表 1　部分原始数据表

姓名	性别	党团	工作经历	获奖/次	单位性质	成绩排名	…	单位名称	毕业时间/年
魏恒	男	党员	有	3	升学	4	…	武汉大学研究生	2004
张东强	男	团员	无	4	国企	19	…	重庆公交公司	2005
雷鑫	男	团员	有	1	民企	9	…	重庆渝强集团	2006
梁艳妮	女	团员	无	0	民企	5	…	吉林延边江都实木家具有限公司	2007
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
韦韬	男	团员	有	1	三资	3	…	广州三菱商事	2009
张帆	女	群众	有	0	民企	10	…	四川恒升钢构工程有限公司	2010
刘刚	男	团员	无	1	三资	4	…	成都中特科技有限公司	2011
贾娟	女	党员	无	3	国企	12	…	重庆齿轮箱有限责任公司	2011

* **收稿日期**:2013-05-11　**修回日期**:2013-07-22　**网络出版时间**:2014-7-3 23:03
资助项目:重庆市软科学项目(No. cstc2011cx-rkxA0341)
作者简介:朱德利,男,讲师,研究方向为智能算法在数据处理和图像处理中的应用,E-mail:Zdlxml@126.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20140703.2303.024.html

2 数据预处理

各个学校在建立数据库时选择的字段不尽相同，有的有序号、身份证号码、联系电话、住址等相关信息，有的这些信息不全面，而且排列顺序也不尽一致，还有就是字段名称、长度不统一，在对这些数据进行分析和挖掘之前需要对其进行预处理[1]。本研究使用微软 SQL Server 2005 integration services(SSIS 2005)作为预处理工具，建立一个就业数据仓库，然后对所有数据进行统一清理，清除数据噪声和与挖掘主题明显无关的数据，如身份证号、学号及其一些明显出错的数据。本次数据挖掘使用的挖掘工具是 Weka，在用 Weka 进行数据挖掘时，为了防止出现乱码，数据最好是英文或者阿拉伯数字。如对以上表格进行数据处理后变为如表 2 所示。其中 *Sex*1 表示性别，*Work-E* 表示是否有工作经验，*W-number* 表示获学金次数，*U-character* 表示单位性质，*P-ranking* 表示成绩排名，*Time*1 表示毕业时间。

表 2　处理后部分数据表

*Sex*1	*Work-E*	*W-number*	*U-character*	*P-ranking*	…	*T*/年
Male	Yes	3	Private E	4	…	2004
Male	No	4	State E	19	…	2005
Male	Yes	1	Private E	9	…	2006
Female	No	0	Private E	5	…	2007
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Male	Yes	1	Foreign E	3	…	2009
Female	Yes	0	Private E	10	…	2010
Male	No	1	Foreign E	4	…	2011
Female	No	3	State E	12	…	2011

3 数据挖掘过程及部分可视化结果分析

3.1 数据挖掘过程

对数据进行整理后，将数据文档输出为 CSV 格式，再在 Weka 中将数据转存为 ARFF 格式的文件，即进入 Weka 数据挖掘流程[2](图 1)。

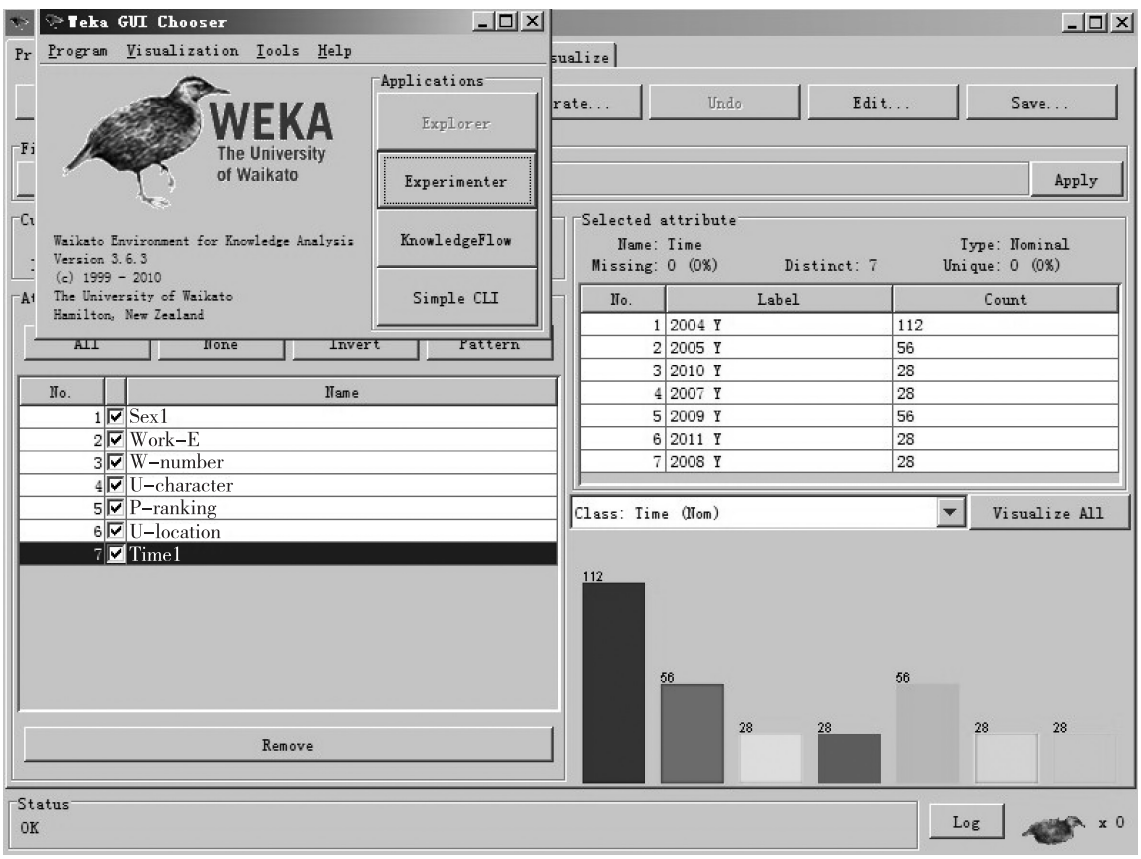


图 1　Weka 工具进行就业数据挖掘过程

本研究使用了时间序列预测算法、决策树分类、朴素贝叶斯聚类和改进的 Apriori 关联规则算法对数据进行挖掘处理，挖掘过程从以下 5 个问题选择变量：

(1)各年份毕业生男女比例变化趋势及其预测;

(2)性别和工作选择关联;

(3)获奖次数和单位性质的关联分析;

(4)毕业生就业单位性质类别分析;

(5)毕业生就业地区类别分析。

由于Weka数据挖掘在展现方面缺乏直观性,本文就基于Weka挖掘的结果进行了再处理[3],使得结论更易于理解。

3.2 数据挖掘可视化结果及其分析

挖掘过程根据本研究确定的5个问题分别得出了挖掘数据和可视化结果。以下是这些挖掘结果的详细阐释。

(1)各年份毕业生男女比例变化趋势。根据本研究获取的数据,重庆市5所高校的信息管理与信息系统专业各年毕业生男女比较直方图如图2所示。

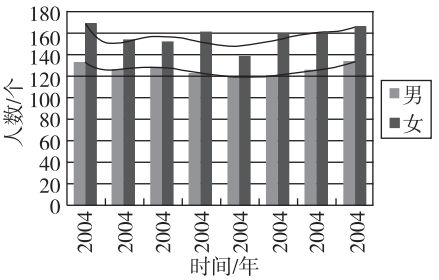


图2 2004—2011年男女毕业生比较直方图

通过对男女各年的变化进行拟合,可以得到男女毕业生数的方程分别为:

男毕业生:$y=0.025\ 7x6-0.742\ 1x5+8.439\ 6x4-47.496x3+136.76x2-188.87x+223.87$。

女毕业生:$y=0.054\ 2x6-1.568\ 3x5+17.88x4-101.4x3+297.45x2-422.65x+378.5$。

由此拟合方程式可以发现,该专业各年的男女比例总体较为平衡,这很大程度上是因为该专业在重庆市的5所高校中既招收文科生又招收理科生,同时该专业所在的学院也有较为平衡的文理趋向。但从整体上看,女毕业生人数还是多于男毕业生人数[4]。对重庆5所高校信息管理与信息系统专业招生过程中的文理特点的调查结果如表3所示。

(2)性别和工作选择关联分析。本研究使用Weka数据挖掘工具所带的Apriori算法研究信息管理与信息系统专业就业数据中的性别与工作选择的关联规则。

通过对表1的数据处理,再用Apriori算法分析,将最小支持度下界设为0.3,上界设为0.8,置信度作为关联分析的度量,设定其最小值为0.85,其他参数保持为默认值[5],可以得到频繁项集6个,靠前的3个强关联为:

1)$Sex1$='male',$group1$='league member',$Work\text{-}E$=yes,319 ==>$job\text{-}property$='other enterprises',301,conf:(0.89);

2)$Sex1$='male',$group1$='masses',$Work\text{-}E$=no,116 ==>$job\text{-}property$='waiting job',101,conf:(0.87);

3)$Sex1$='female',$group1$='party member',$Work\text{-}E$=no,36 ==>$job\text{-}property$='graduate',31,conf:(0.86)。

job-property表示工作属性,group1表示党群关系。关联1)的含义可以解释为:在所有数据中有339个样本属性为"性别为男,同时党群关系为团员,有工作经验",符合这些条件的样本中有301个工作属性为"其他企业"。置信度为0.89。这个规则说明有工作经验的男毕业生选择其他企业的居多,根据高校的实际情况,党群关系为团员的同学占大多数,有工作经验也是一个较为泛化的统计,故此关联规则实际意义不大。

表3 重庆5所高校信管专业的文理趋向分析

学校	信管专业招生特点		
	招生的文理要求	信管专业隶属的学院	学院的文理趋向
重庆师范大学	理科	计算机与信息科学学院	偏理科
西南大学	文理兼收	计算机与信息科学学院	偏理科
重庆大学	文理兼收	经济与工商管理学院	偏文科
重庆工商大学	文理兼收	管理学院	偏文科
重庆理工大学	文理兼收	计算机科学与工程学院	偏理科

关联2)的解释是在所有数据中有116个毕业生属性为"性别为男,同时党群关系为群众,没有工作经验",符合这些条件的样本中有101个工作属性为"待就业"。置信度为0.87。这说明"性别为男,党群关系为群众,没有工作经验"与"待就业"的就业状态关联度较高。

关联 3)可以解释为在所有数据中有 36 个毕业生符合“女生，同时也是党员，但是没有工作经验”的条件，他们中有 31 个的工作状态为“考研”，置信度为 0.86。这说明“性别为女，党群关系为党员，没有工作经验”与“考研”的就业状态关联度较高。不过这一挖掘过程中的有效样本还显不足。

普遍认为，男毕业生在就业中比较占优势，同时工作经验对大学的成绩也是由正面影响的[6]，故 2)和 3)体现的规则和这些认识有一定的差距，是本研究中发现的较有价值的规则。值得管理学生工作和就业工作的部门研究。

(3)获奖次数和单位性质的关联分析。同样，把表 1 的数据处理为适合获奖次数与单位性质分析的表，把获奖次数作为文本字段处理[7]，再载入到 Weka 用 Apriori 算法分析其关联规则，设置最小支持度下界设为 0.3，上界为 0.7；置信度最小值为 0.7，其他参数保持为默认值，可以得到频繁项集 8 个，靠前的 3 个强关联为：

1). prize＝'0'，*Work-E*＝yes，*group*1＝'league member'，229 ＝＝＞*job-property*＝'other enterprises'，185，conf：(0.81)；

2). prize＝'3'，*Work-E*＝yes，*group*1＝'party member'，62 ＝＝＞*job-property*＝'state-owned enterprises'，46，conf：(0.75)；

3). prize＝'1'，*Work-E*＝no，*group*1＝'league member'，98 ＝＝＞*job-property*＝'flexible job'，71，conf：(0.72)。

规则 1)可以解释为，获奖次数为“0”，有工作经验且为团员的人，有 81％选择的工作属性为“其他企业”；规则 2)则表明获得奖励次数为“3”，且为党员的有工作经验的人有 75％选择国企；而规则 3)则表明获得一次奖励的没有工作经验的团员选择灵活就业的有 72％。这三个关联规则表明，获奖次数多的学生选择传统就业领域的概率比较大。

(4)毕业生就业单位性质分析。图 3 是重庆市信管专业毕业生的流向统计，按单位性质流向划分。由图 3 所知，信管专业的毕业生在其他企业的就业人数约占总人数的一半，待就业的毕业生占 10％～30％左右，也就是说，还有很大一部分信管专业的毕业生和全国其他大学毕业生一样，处于未就业状态。另外还可以看到，信管专业的毕业生灵活就业比例较大。还有很多毕业生选择考研/升学继续深造，以更高的学历和更多的知识来获得更好的工作。公务员和事业单位的工作人员也是信管专业毕业生的选择之一[8]。另外，自主创业、教育单位和其他企业都涉及信管专业的毕业生。从这个分析数据可以发现，信管专业的就业面较广，但企业是最主要的就业领域。

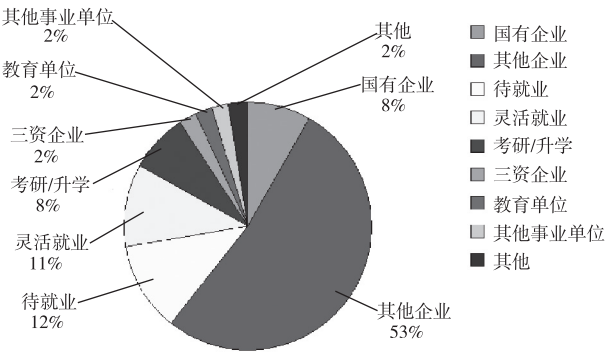
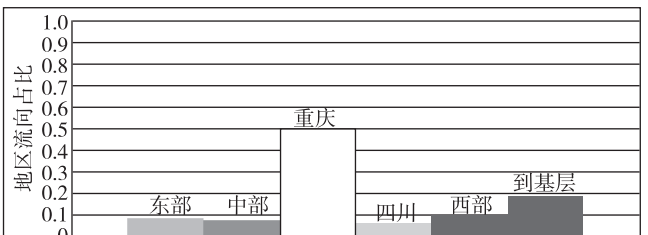


图 3　毕业生就业单位性质流向图

(5)毕业生就业地区分析。图 4 是按地方区域流向划分的毕业生地区流向图。在图中以百分比表征了重庆市信息管理与信息系统专业签约就业的地区差异。

通过图 4 可以看到，在众多的重庆市信管专业毕业生中，大都选择在重庆就业，一方面，重庆本地生源占有很大比例，另一方面，重庆近几年经济高速发展，很多外地来重庆求学的毕业生也愿意留在重庆发展。中、东部等经济相对发达的地区就业人数反而比较少，体现出该专业中、东部的就业信息向重庆市传递还不够充分。



东部	0.081 199 005
中部	0.072 690 781
重庆	0.496 094 943
四川	0.062 088 224
西部	0.096 819 233
到基层	0.191 107 814

图 4　毕业生地区流向比例图

4 问题及原因分析

4.1 数据挖掘暴露出的问题

在本研究中对数据的分析结果支持大部分普遍认同的观点。比如，工作经历与进入民企、三资企业就业的关联度有 82％，表明实践能力较强的，有一定社会工作经验的毕业生较容易被民企、三资企业录取。国家机关、事业单位以及国有企业更加注重学生的综合素质[9]，因为

挖掘结果显示获奖次数为3次的毕业生与在国有单位就业的关联度达到75%。

数据挖掘的重要价值在于发现存在于数据中而又和传统认识有所差距的模式。本次研究发现的一个和传统认识切合度不高的一个重要问题是:信息管理与信息系统专业的就业单位几乎很难聚类。笔者选取了经典的K-means聚类算法对就业单位进行聚类分析,多次调整参数后的聚类效果都不令人满意,这表明本专业的就业单位类型比较分散,没有较为固定的就业单位。在通常的认识中,一个专业应该有一个较为固定的就业方向,但是本研究收集到的数据不支持这一观点。

4.2 原因分析

出现以上问题的原因,笔者认为有以下两个方面:

(1)传统的作为信息管理与信息系统专业主要就业阵地的图书情报部门对于本科生的需求量越来越少,已经不能形成一个特定的就业类型。而与此同时在国家信息化建设的宏观环境下,信息管理和服务的观念、内容及形式发生了深刻变化,信息管理人才的培养和就业有更多的新的市场需求[10]。

(2)实际就业状况和专业设计的就业方向有较大的差距。在重庆市多个高校的信息管理与信息系统的教学计划中,几乎都对该专业以"方向选修课"的形式分了方向,归纳起来,大概有政务信息管理、管理学、计算机应用、图书馆学等4个方向[9],按照各个学校的教学大纲说明,理想中的专业方向和就业领域关系如图5所示。但是在实际就业环境中,只有少部分毕业生在就业的时候是按照这个设计找到的"对口单位",每个方向的毕业在就业领域上都是相互交叉的现象非常普遍。图6表明了这种现象。

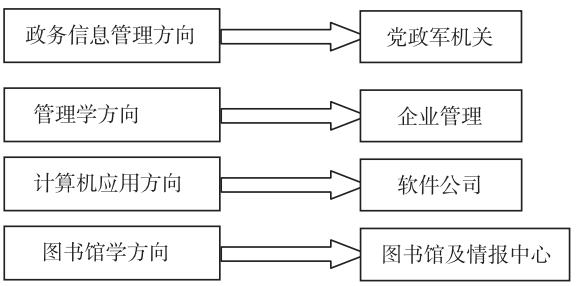


图5 理想中的专业方向和就业领域关系图

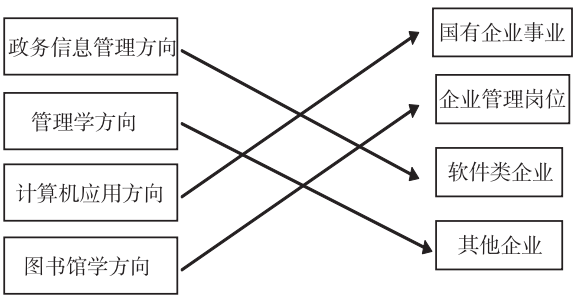


图6 实际就业环境的专业方向和就业领域相互交叉图

5 结语

经过对信息管理与信息系统专业学生就业情况的数据挖掘[11-14],笔者发现随着社会信息化程度的不断提高,信息管理与信息系统专业的学生遍及各行各业,就业领域的增大,为信息管理与信息系统专业学生提供了更多的就业机会。但是面对就业机会的增加,本专业学生应该不断提高自己的综合素质和各方面能力,以适应社会的需求。另外,学校也要给该专业的学生拓宽就业渠道,社会应该尽快提高对该专业的认识。发挥社会、学校、老师、学生的综合效应方能从根本上提高信息管理与信息系统专业的就业质量。

参考文献:

[1] 范明,孟小峰.数据挖掘概念与技术[M].北京:机械工业出版社,2006.
Fang M, Meng X F. The concept of data mining[M]. Beijing: Mechanical Industry Press, 2006.

[2] Weka. Wiki home[EB/OL]. (2013-6-20). http:// weka.wikispaces.com/.

[3] 李玲俐.数据挖掘中分类算法综述[J].重庆师范大学学报:自然科学版,2011,28(4):44-47.
Li L L. Summary of data mining classification algorithms [J]. Journal of Chongqing Normal University: Natural Science, 2011, 28(4): 44-47.

[4] 秦必瑜.信息管理与信息系统专业人才培养与教学改革探讨[C]//北京市高等教育学会2007年学术年会论文集.北京:北京市高等教育学会,2007.
Qin B Y. Reform of information management and information system specialty[C]//Beijing institute of higher education 2007 conference proceedings. Beijing: Beijing Institute of Higher Education, 2007.

[5] 吕赛鹍,李志平.关联分析方法在高校教学评价中的应用[J].现代教育技术,2009,19:34-36.
Lü S D, Li Z P. Correlation analysis method in teaching evaluation Universities[J]. Modern Educational Technology, 2009, 19: 34-36.

[6] 李志坚,莫建麟. 一种改进的基于概念格的数据挖掘算法

[J]. 重庆师范大学学报:自然科学版,2013,30(2):92-95.
Li Z J, Mo J L. An improved concept Lattice-based data mining algorithm[J]. Journal of Chongqing Normal University: Natural Science, 2013, 30(2): 92-95.

[7] 刘美玲,李熹,李永胜. 数据挖掘技术在高校教学与管理中的应用[J]. 计算机工程与设计,2010,31(5):1130-1133.
Liu M L, Li X, Li Y S. Application of data mining technology in teaching and management in universities[J]. Computer Engineering and Design, 2010, 31 (5): 1130-1133.

[8] 岳昌君,巩建闽,黄潞. 高校毕业生就业特点及其变化趋势[J]. 教育发展研究,2008(7):26-30.
Yue C J, Gong J M, Huan L. Employment of college graduates characteristics and trends[J]. Education Development Research, 2008(7): 26-30.

[9] 张洁. 信息管理与信息系统专业课程体系设置研究[D]. 河北:河北工业大学,2008.
Zhang H. Curriculum system of information management and information system[D]. Hebei: Hebei University of Technology, 2008.

[10] 查先进. 信息管理与信息系统专业人才培养方向和课程体系探索--基于科技信息专业背景的实证[D]. 武汉:武汉大学信息管理学院,2003.
Zha X J. Orientation and course system of information management and information system[D]. Wuhan: College of Information Management, Wuhan University, 2003.

[11] 王海荣. 数据挖掘在学生成绩分析中的应用[J]. 电子设计工程,2013(4):54-56.
Wang H R. Application of data mining in analysis of students performance[J]. SAMSON, 2013(4): 54-56.

[12] 王加年. 基于数据挖掘的高校自动化办公系统建设[J]. 电子设计工程,2013(24):49-53.
Wang J N. Construction of university office automation iystem based on data mining[J]. SAMSON, 2013(24): 49-53.

[13] 杨金花. 基于 Web 挖掘的层次凝聚类算法研究[J]. 电子设计工程,2012(12):30-32.
Yang J H. Algorithm-based Web mining the level of cohesion class[J]. SAMSON, 2012(12): 30-32.

[14] 朱金坛. 数据挖掘 Apriori 算法的改进[J]. 电子设计工程,2013(15):37-40.
Zhu J T. Improve of data mining Apriori algorithm[J]. SAMSON, 2013(15): 37-40.

Employment Data Analysis and Pattern Mining Based on Weka

——Take Specialty of Information Management and Information System in Chongqing for Example

ZHU Deli

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: We use WEKA as a tool for data analysis and pattern mining; the data come from five universities in Chongqing. Time series forecasting algorithm, decision tree classifier, and naïve Bayesian clustering and improved a prior association rules algorithm are used for our research. We analysis some of the visualization of mining results, get some patterns in the data of specialty of information management and information system in Chongqing. The results of our research include: the proportion of male and female is balanced overall, largely because both liberalists students and science students are recruited; female students with no work experience are high correlation with the state "admitted students"; companies are the leading area of employment of this specialty; the students get more awards, they choose more traditional job areas; it is not sufficient for job information of eastern China transferring to Chongqing. This paper also analyzes two reasons: first, libraries have been unable to form a unique job type, but new needs are generated at the same time; second, there are gap between job environment and the design of specialty position.

Key words: information management and information systems; employment; data mining

(责任编辑 游中胜)