

# 基于动态聚类的网上学员细分实证研究

毛 布<sup>1</sup>, 田 林<sup>2</sup>, 谢 汶<sup>3</sup>

(1. 四川自贡广播电视大学, 四川 自贡 643000 2 云南楚雄师范学院, 云南 楚雄 675000 3 四川大学, 成都 610000)

**摘 要:**自主式学习在远程教育中占有重要的地位, 如何更高效、快捷的建立学员所需要的知识资源库是当前我们研究的一个热门领域。本文以自贡电大 2009 级近百名本科学员基本资料及学习记录为采样数据, 利用动态聚类的方法进行了有效的学员细分及数据分析, 并在此基础上提出了相应的建立适合远程教育的资源库的策略。

**关键词:** 动态聚类; 网上自主式学习; 学员细分; 数据分析

**中图分类号:** TP311

**文献标识码:** A

## 引 言

近年来, 远程教育已经逐渐成熟, 它已从传统的以教师为主的指导学习方式过渡到以学员为主的自主学习方式。在自主学习中借助网络媒体的学习尤为占其主导地位。

为了进一步提高对学员的服务质量, 我们采样了自贡电大近一百名学员的上网自学行为记录、学员类型及学员偏好等; 并以此为数据基础, 做了详细的技术分析后, 制定出了如何建立适合远程教育的学习资源库的策略。

面对学员的学习行为记录及偏好等, 我们用传统的统计分析方法已经无法从中获取更多有价值的信息, 而数据挖掘的提出则为深入进行学员分析提供了有效的途径<sup>[1-2]</sup>。

本文在学员行为记录为主的数据分析基础上, 深入研究影响学生学习的各种可能因素, 构建了学员多维细分模型, 并利用动态聚类技术对学员进行细分<sup>[3-4]</sup>。

## 1 网上学员多维细分模型

### 1.1 学员多维细分指标体系

综合分析相关研究成果<sup>[5]</sup>, 提出“网上学员多维细分指标体系”, 如图 1 所示, 该体系以最大满足学员所需知识为最终目标, 具体包括 4 个一级指标:

学员性质 为了有针对性的建立适合学员的远程教

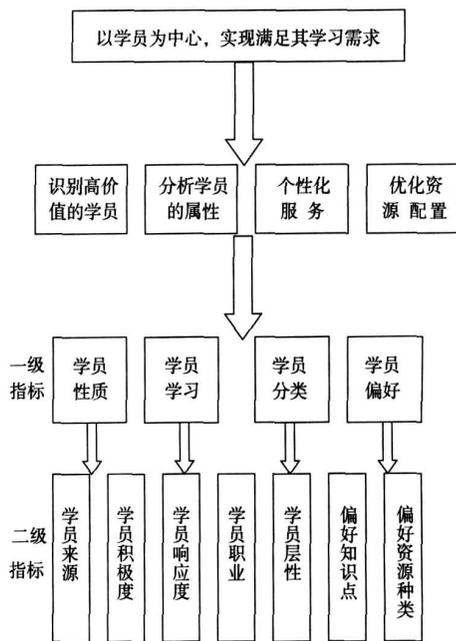


图 1 网上学员多维细分指标体系

育资源, 必须对学员性质进行分析汇总。我们把学员性质再细分为学员来源。

学员学习 学员的学习很大程度上反映了学习资源的应用情况。为此, 我们从学员的学习积极程度及学员对学校发出的学习响应程度上进行了数据分析。

学员分类 远程教育类的学员大多为的在职人员, 他们来自不同的岗位。我们从学员职业及学员层次上

进行了细分。

学员偏好 每个学员都有自己不同的喜好, 这很大程度上反映了他们所需知识的范围, 同时决定了哪些资源才适合他们。

### 1.2 指标计算

本文提出的模型中一级指标包含 9 个二级指标<sup>[6-9]</sup>。各个维度的含义、计算机方式和维度取值范围如下:

#### 1) 学员来源

学员来源是指招生时生源的来源, 根据样本数据和二八法则确定阈值, 划分维度区间为 {自费, 公派}, 确定阈值为 {2 1}。

#### 2) 学员学习积极程度

学员学习积极程度可以通过计算各学员的上网行为记录来衡量, 其计算方式采用学员上网行为记录数据为研究对象:

积极程度 = 学员日均访问远程教育网站时间 ÷日均上网时间

参照上网记录标准确定阈值 {1.5 1, 0.5 0.1}, 划分维度区间为 {很积极, 比较积极, 不太积极, 很不积极}。

#### 3) 学员职业

远程教育类学员大都为在职工作人员, 所以评估各学员从事的行业背景对反馈他们的学习需求有一定的作用。根据对自贡电大采样学员的评估, 划分维度区间为 {公务员、事业单位、企业、各体、其它}; 设定阈值为 {5 4 3 2 1}

#### 4) 学员层次

学员层次主要是指学员入学前的学历, 根据学员入学前的情况, 划分维度区间为 {大专学历、高中或中专等相当学历}; 设定阈值为 {2 1}。

#### 5) 学员响应度

学员响应度是指学员对学校发出网上学习指令的响应速度:

学员响应度 = 访问学习网站次数 ÷发出学习指令次数

参照行业标准确定阈值为 {1, 0.5}, 划分维度区间为 {高, 低}。

#### 6) 偏好知识点

偏好知识点是指学员在上网行为中喜欢访问的知识内容, 该维度很大程度上能反映出学员学习行为的心理状况。

某学习资源偏好度 = 某学习资源访问时间 ÷上网学习总时间

根据样本数据, 划分维度区间为 {高、中、低}, 设定

阈值为 {1, 0.5 0.1}。

#### 7) 偏好资源种类

学员在网上进行远程教育自主学习时, 他们往往会更为关注自己感兴趣的学习资源种类。我们把网上学习资源种类分为 {视频、图表、文字}, 设定阈值为 {3 2 1}。从数据上看, 以视频为主的资源更让学员亲睐。

#### 8) 发帖率: 周发帖数 /周上网数

9) 交流度为教师和学生之间有交互恒量指示: 交流信息数 /上网学习次数

## 2 网上学员细分的实证分析

### 2.1 数据准备

#### 2.1.1 样本数据

聚类采用的维度数量越多, 对于数据的精度要求也就越高, 因此我们选取了自贡电大网上学员学习记录、学员基本资料为数据来源<sup>[10]</sup>。总共收集了 102 名学员的基本信息和 3212 条学习记录, 时间跨度为 2008 年 4 月 17 日至 2010 年 1 月 3 日。在此基础上进一步对数据进行采样, 选取其中代表性较强的数据, 得出表 1。

表 1 分类 1 和分类 2 的维度数值统计表

维度	分类 2(样本量: 学号: XXX)			分类 1(样本量: 学号: XXX)		
	均值	标准差	说明	均值	标准差	说明
积极性	0.775	0.002	比较积极	0.453	0.003	不太积极
层次度	1.72	0.003	较高	0.86	0.003	较低
响应度	0.75	0.015	较高	0.45	0.012	低
偏好知识点	0.67	0.01	较高	0.83	0.01	高
偏好种类	2.3	0.26	较高	2.7	0.24	高
发帖率	0.23	0.001	中	0.21	0.001	较低
交流度	0.85	0.002	中	0.76	0.003	较低

#### 2.1.2 数据的重构和整合

根据前文中列出的二级指标计算样本学员的指标值。由于样本数据集的数据完整性限制, 我们就有选择地删减指标数量。

#### 2.1.3 聚类过程的数据处理

比例缩放和加权的概念在聚类过程中都起着重要的作用。考虑到不同变量是以不同单位或者在不同范围中测量的事实, 比例缩放可用于调整变量的值。加权方式则提供了对变量的一种相对调整, 使得其中的一些变量变得比另一些变量更重要<sup>[11]</sup>。

### 2.2 动态聚类过程

通过 SQL Server2005 中的数据挖掘组件对学员数据进行挖掘<sup>[12]</sup>, 过程主要有如下几个步骤。

1) 使用不同的 k 值 (即学员分类数) 进行自动聚类探测, 比较不同 k 值下的聚类结果的质量, 初步确定最优 k 值: 聚类区别于分类的一大特征就是分类数的不确定性, 在聚类开始之前, 很难预先制定一个理由去选择一个特定的 k 值, 为了避免最后聚类得到的类别过于繁

多,导致不可解释性和不符合业务需要,需要预先使用不同的  $k$  值进行自动聚类探测,对于每一次的聚类结果,通过评价簇 (cluster) 的好坏、可解释度、业务需求的契合度等因素,初步确定  $k=7$  (如图 2)。

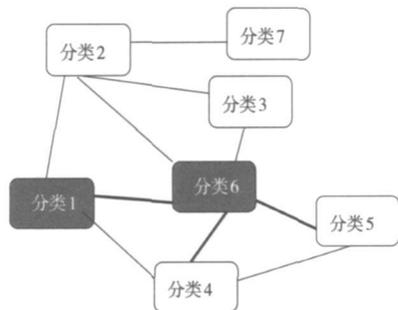


图 2 显示聚类之间连接的关系图

2)分析分类关系图,评价簇的好坏,提出改进方案:簇好坏的评价主要有两个指标:“簇内形似”和“簇间差异”<sup>[10]</sup>。从几何学的角度看,就是簇内各个记录之间彼此相近,而各个簇之间分开得很远。图 2 是  $k=7$  时所获得的分类关系视图,表现了各簇的强度 (簇内形似) 以

及簇与簇之间的连接强度 (簇间差异)。图中连线的深浅程度标识簇间差异,各节点的深浅代表差异的大小。

在每次聚类过程中,时常会产生一个或者若干个强簇——即样本数量相当大的簇,其中的记录非常相似。

3)根据改进方案,适当调整  $k$  值,得到最终学员群:除去强簇和表现良好的强力牵引,在剥离的样本中,存在较原先表现更为良好的簇;但同时,这部分样本数量较少、个体差异较大,导致某些离群值对聚类的影响也会放大。因此,在新一轮聚类中,我们将原先剥离 3 个簇按  $k=4$  进行单独聚类 (见图 2)。最终,总共得到 5 个簇,即 5 个学员类。

### 2.3 聚类结果与分析

根据聚类结果,在对每个指标值进行统计的基础上,按照 1.2 节中的取值范围说明,可以得到 5 个学员类的特征描述 (见表 2)。

针对每类学员的特征描述,对类内样本单独做深入数据分析之后,提出以下具有针对性的建立学习资源 (见表 3),其中学员类按照重要程度排序。

表 2 聚类结果

分类 (样本数)	结果	学员分析				行为分析	
		积极性	层次感	响应度	偏好种类	发帖率	交流度
分类 1	指标值 (特征描述)	0.453	0.86	0.45	2.7	0.21	0.76
分类 2	指标值 (特征描述)	0.775	1.72	0.75	2.3	0.23	0.85
分类 3	指标值 (特征描述)	0.65	1.54	0.73	2.6	0.28	0.91
分类 4	指标值 (特征描述)	0.24	0.56	0.35	2.5	0.09	0.54
分类 5	指标值 (特征描述)	0.56	1.25	0.79	2.9	0.20	0.80

表 3 学员分析与建设学习资源策略

学员类	学员分析	建设学习资源策略
分类 4	属睡眠学员,学习积极差、响应程度低	应反馈于教师,做适当辅助性教学
分类 5	属不活跃,但响应程度高学员	需教师学习指引,提高其学习积极性
分类 3	属具有学习潜力类学员,发帖及交流度强	增加此类学员关注程度,正确引导、开发
分类 1	学习类型偏好性重,则重于视频类资源	有针对性建设某类学习资源
分类 2	学习行为记录时间长,积极度高。	可对这类的行为资源进行个别分析,加大针对性资源建设

### 3 结束语

本文提出了一个具有网上学习特色的多目标多维度的学员细分模型,并利用动态聚类方法进行了实证分析,最终挖掘出具有现实意义的学员群,用于帮助远程教育学校建设更为适合学员的学习资源<sup>[13]</sup>。当然,在指标相关性研究及通过建立学员类型预测模型来识别数据不足的新学员等方面,还有待今后进一步深入研究。

### 参考文献:

- [1] 张国政. 客户关系管理中基于数据挖掘的客户细分研究 [J]. 商业研究, 2006(13): 153-155
- [2] Rygelski c Wang J-C Yen D C. Data mining techniques

for customer relationship management [J]. Technology in Society 2002, 24(24): 483-502

- [3] 李斌, 郭剑毅. 聚类分析在客户关系管理中的研究与应用 [J]. 计算机工程与设计, 2005, 26(2): 540-542
- [4] 赵宏霞, 杨皎平, 陈宗娇. 面向客户需求细分的数据挖掘研究 [J]. 科技管理研究, 2005(11): 207-210
- [5] 王国胤, 何晓. 一种不确定性条件下的自主式知识学习模型 [J]. 软件学报, 2003(6): 57-59
- [6] 钟志贤, 黄林凯. 对教学信息系统开发与应用的几点反思 [J]. 中国远程教育, 2010(1): 62-67
- [7] 李卫东, 宋威, 杨炳儒. 利用数据挖掘方法分析客户生涯价值 [J]. 计算机工程与应用, 2005, 41(6): 17-18, 36
- [8] 郭蕴华, 陈定方. 基于模糊聚类分析的客户分类算

- 法研究 [J]. 计算机应用研究, 2005, 22(4): 52-53
- [9] 宋艳, 梁静国. 基于模糊聚类的客户分类应用研究 [J]. 物流科技, 2005, 28(1): 26-28
- [10] [http://www.zgrtu.com/adm\\_inuser/htglxt.php](http://www.zgrtu.com/adm_inuser/htglxt.php) 2010-3-16
- [11] 别荣芳, 尹静, 邓六爱, 译. 数据挖掘技术 [M]. 北京: 机械工业出版社, 2006
- [12] 邝祝芳, 焦贤龙, 高升, 译. 数据挖掘原理与应用—SQL Server2005数据库 [M]. 北京: 清华大学出版社, 2007.
- [13] 王慧敏, 陈泽宇, 王敏娟, 等. 移动学习情境中教育智能应用探究 [J]. 中国远程教育, 2010(1): 68-71.

## Empirical Research on Student of Internet Based on Clustering

MAO Bu<sup>1</sup>, TIAN Lin<sup>2</sup>, XIE Wen<sup>3</sup>

(1. Sichuan Broadcast and Television College, Zigong 643000, China

2. Chuxiong Teachers College, Chuxiong 675000, China; 3. Sichuan University, Chengdu 61000, China)

**Abstract** Online unattended learning is of key importance in distance education, therefore making the establishment of a knowledge resource pool for the learners in a more efficient and faster manner a hot field for our research. This essay tries to venture an effective analysis with data and study with the basic information and study segmentation of almost a hundred students of the year 2009 as sampling data and based on that proposes solutions to create resource pool in keeping with distance education.

**Keywords** dynamic clustering; study on Internet; student segmentation; data study

(上接第 681 页)

- [5] 陆庆, 周世杰, 秦志光, 等. 对等网络流量检测技术 [J]. 电子科技大学学报, 2007, 36(6): 133-137.
- [6] 李明伟, 张大方. 基于有效载荷分析的 BT 流量识别技术 [J]. 计算机应用, 2007, 27(9): 230-232.
- [7] 侯自强. P2P: 让互联网无处不在. <http://www.dchhljnpit.net.cn>, 2005.
- [8] 陆晓雯. P2P 流量识别与控制系统的的设计研究 [D]. 南京理工大学, 2008.

## Technology of BT Traffic Identification Based on Small Data Package

ZENG Yan, LI Chunwei

(School of Computer Science, Sichuan University of Science & Engineering, Zigong 643000, China)

**Abstract** Due to the heavy burden of bandwidth resources caused by the BT traffic, traffic identification and control is very important. Through the BitTorrent protocol analysis and comparing the actual transmission data of BitTorrent, Thunderbolt etc, BT traffic was identified by finding its characteristics of transmission. The experiment shows that BT traffic was characterized by data packages payload size of 5 bytes on the data transmission process, and it has a certain value to identify BT traffic.

**Keywords** BitTorrent; traffic identification; payload