移动学习中的数据挖掘研究

□ 刘 钢 王敏娟 张 驰 王慧敏 陈笑怡

【摘 要】随着网络技术与通信技术的迅速发展，移动学习呈现出强大的发展势头。本文以一个实际运行的移动学习网站为研究平台，利用数据挖掘技术对网站客观数据进行分析研究，并就“移动学习使用者特征”、“知识资源个性化推荐技术”和“资源需求趋势预测”这三个比较值得关注的问题进行了实验论证，对移动学习研究者和基于知识服务的网站运营决策者具有一定的启发意义和参考价值。

【关键词】移动学习；数据挖掘；用户聚类；协同过滤；混合推荐技术；时序预测算法

【中图分类号】G420 【文献标识码】A 【文章编号】1009—458x(2011)01—0031—05

在这个网络通讯技术高速发展、知识信息大爆炸的时代，移动学习呈现出了强大的发展势头。然而，海量数据信息中的“信息过载”问题也越来越尖锐地摆在教育学者和网站运营商面前：如何有效提高知识信息的利用率？如何准确洞悉用户情况和使用需求趋势？本文通过对一个实际运行的移动学习网站进行数据挖掘分析，分别对“移动学习使用者特征”、“移动知识资源的个性化推荐”、“移动课件资源需求趋势”这三个热门话题进行网站客观数据的实证研究，希望对相关同行有借鉴和参考作用。

实验背景

移动百科网站（www.hellobaike.com）是一个基于片段式学习特点而设计的M-learning知识资源库，提供适合手机学习的各类百科知识，该知识库资源WEB系统具备三个特点：它是一个提供视频课件资源的知识库系统，其中的信息资源很难进行基于内容的特征分析；信息资源的内容覆盖面很广，而且资源数量会不断增长，没有固定的特性分布；系统的使用人群不固定，不断有新用户通过注册加入进来。这也是当前很多WEB系统或网站所共有的特点，如Amazon，eBay，YouTube等等[1]，因此下面三个问题成了本次实验研究的关键所在：

第一，对新注册用户的快速聚类问题——系统新用户的历史行为信息（包括评价信息）是空白的，对于很多需要基于用户评价信息进行计算的具有推荐服务的系统来说，比如协同过滤系统和基于网路结构的推荐系统，就很难对此类用户进行产品推荐。

第二，将资源个性化地推荐给用户的问题——用户与所有产品可以建立一个评价向量：$E_i=(x_1,x_2,\cdots x_j)$，若用户i只对为数很少的产品作出了评价，就会导致评价向量E_i中有效评分值x_j很稀少，这同样会导致需要基于用户评价信息进行推荐的推荐系统无法得到较好的推荐效果，或是产生的推荐结果集很小。

第三，学习资源需求量预测的问题——要想得到用户的资源需求，我们通常的做法是对用户使用资源的满意度情况进行问卷调查，而这种方式成本较高，耗时较长。如何根据用户下载知识资源的网站客观数据，实时地预测未来一周移动学习资源需求量的趋势信息是一个值得研究的问题。

新注册用户的特征聚类

针对前文提到的新用户可用信息稀疏的问题，我们引入人口统计信息分析的技术，把所有用户按照个人特征信息进行聚类，通过用户聚类来找到目标用户的近邻用户，然后以其作为协同过滤的计算用户集。

本研究使用EM（期望最大化）算法来进行用户聚类，根据目标对象与目标族隶属关系的概率来分配对象，它的孤立数据不敏感性、算法收敛的稳定性和高效性对于依据用户人口统计信息的聚类有较好效果[2]。

1. 聚类特征维度的选取和数据预处理

有研究通过决策树技术发现用户的性别、年龄、职业、文化背景等特征信息对其兴趣偏好产生相对较大的影响因子[3]。本文在进行用户聚类时使用的特征维度是：性别、年龄、职业、文化程度和收入水平。

表 1 用户聚类维度的数据量化表

性别	年龄	职业	文化程度
1 男 2 女	1 14 岁以下 2 15~24 岁 3 25~34 岁 4 35 岁 ~44 岁 5 45 岁 ~60 岁 6 60 岁以上	1 中小学生 2 理工科大学生 3 文科大学生 4 事业单位职工 5 事业单位干部 ……	初中及以下 高中及中专 大专 本科 硕士 博士

各个用户特征维度上数据的预处理工作能有效提高用户属性的表示能力，并有助于提高算法的收敛速度。首先把相关的用户维度信息从数据库的不同位置抽取出来，对其进行数据清理和量化操作，最后再将规整化的数据装载至数据库的特定位置。

表 2 数据降维规则表

字段名称	规则描述	转换值	说明
F_AGE	F_AGE<=16 16<F_AGE<=40 40<F_AGE<=65 65<F_AGE	0 1 2 3	少年 青年 中年 老年
F_EduLevel	高中以下 (含) 大、专科 硕士以上 (含)	0 1 2	低学历 中等学历 高学历
F_JobType	…	0 1 2 3 …	务农 技工 文职 …
F_SpareTime	F_SpareTime<=1 1<F_SpareTime<=3 3<F_SpareTime<=6 6<F_SpareTime	0 1 2 3	忙 较忙 较闲 闲
F_Income	F_Icome<=1000 1000<F_Icome<=3000 3000<F_Icome<=6000 6000<F_Icome<=10000 10000<F_Icome	0 1 2 3 4	低 较低 中 较高 高
F_FavTime	24:00<F_FavTime<=5:00 5:00<F_FavTime<=8:00 8:00<F_FavTime<=11:00 11:00<F_FavTime<=13:00 13:00<F_FavTime<=17:00 17:00<F_FavTime<=20:00 20:00<F_FavTime<=24:00	0 1 2 3 4 5 6	深夜 清晨 上午 中午 下午 傍晚 晚上

2．EM 算法实现

对于所有的用户数据 x，并不知道各自来自哪个分支（聚类簇），如果把用户完整数据表达成 (x,y)，其中有 y 表示 x 所属分支的标签，取值为 y∈ (1，…，g)，那么整体数据的概率密码就可以表示为：

$$f(x,y;\theta)=\sum_{i}^{g} r_i f_i(x,y;\theta_i) \quad ①$$

公式①中的 g 为密度分支的个数，$r_1,r_2,\cdots,r_g$ 是各分支点总体分布的比例，f_i 是第 i 个分支的密度，θ_i 则是相应分支的未知参数。有了用户数据集 $\{x_1,x_2,\cdots,x_n\}$ 之后，就可使用极大似然估计的方法算得 $\hat{\theta}_{MLE}$：

$$\hat{\theta}_{MLE}=\arg\max\prod_{i=1}^{n} f(x_i,y_i;\theta) \quad ②$$

EM 算法本质上是迭代算法。它从初始解 θ_0 开始，迭代地得到 $\theta_1,\theta_2,\cdots,\theta_t$。在每步迭代中，似然函数单调增加。算法的执行步骤如下：

(1) 给定初始解 θ_0；

(2) 对 t=0,1,2,… 时，重复地进行下面两步操作：

E 步骤：在用户数据和当前解 θ_t 给定的情况下，计算完整数据的对数似然函数期望值。

$$Q(\theta|\theta_t)=\sum_{i=1}^{n} E_y[\log f(x,y;\theta)|x_i,\theta_t] \quad ③$$

公式③中的 E_y 是关于随机变量 y 求期望。

M 步骤：选取一个新的参数 θ_{t+1}，使对数似然函数期望值最大化。

$$\theta_{t+1}=\arg\max_{\theta} Q(\theta|\theta_t) \quad ④$$

(3) 如此往复迭代，计算得到 θ_t，直到算法达到收敛。

3．用户聚类结果

通过 EM 迭代自适应的计算，可得到各类别用户簇，各个类别的分布特征如图 1 所示。

图 1 移动学习数据聚类特征分布图

知识资源的个性化推荐

个性化资源推荐模型主要由 3 个功能模块组成：行为记录模块、模型分析模块和推荐算法模块。

行为记录模块是推荐模型的输入部分，主要负责用户注册信息和系统使用信息的记录，并将它们存储到数据库中的特定位置；模型分析模块负责用户信息的抽取、转换和重载操作，并完成对用户人口统计信

息的分析工作，以实现对新用户和产品评价信息稀疏用户的资源推荐；推荐算法模块是整个推荐模型的核心功能模块，它负责模型的大部门计算工作，通过产品资源的协同过滤来实现对目标用户的推荐服务。

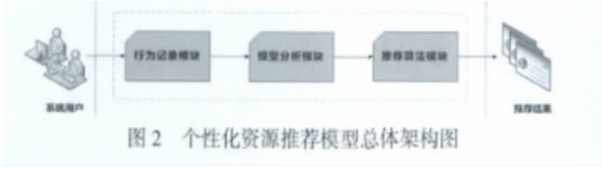

图 2　个性化资源推荐模型总体架构图

1. 协同过滤

合理利用用户之间的兴趣相似性，可以有效提高推荐的精确度，所以本研究中提出的混合推荐技术是以协同过滤技术为基础的。协同过滤分析用户兴趣，在用户群中找到与目标用户兴趣相似的用户，依据这些近邻用户的产品评价信息来计算出目标用户对其未接触过的产品的预测评分，将此作为衡量目标用户对产品的喜好程度，进而对其进行推荐。

⑴　选取近邻用户

通过计算目标用户与其他用户的评价向量 $E_i=(x_1,x_2,\cdots,x_j)$间的相似性，满足一定阈值 δ 要求的用户将被选取为目标用户的近邻用户，作为评分预测的计算用户集。本研究中使用 Pearson 系统来计算用户相似性，用它来表征用户之间的相似性有较好效果[4]。

$$Sim(x,y)=\frac{\sum_{j\in I_{xy}}(r_{xj}-\bar{r}_x)(r_{yj}-\bar{r}_y)}{\sqrt{\sum_{j\in I_{xy}}(r_{xj}-\bar{r}_x)^2}\sqrt{\sum_{j\in I_{xy}}(r_{yj}-\bar{r}_y)^2}} \quad ⑤$$

公式⑤中的$\bar{r}_x$是用户 x 对产品的评分均值，r_{xj}是用户对产品 j 的评分，I_{xy}代表用户 x 和用户 y 都评价过的产品集，Sim(x,y)则是用户 x 和用户 y 的相似度系数。对于产品评价信息稀疏的用户或新用户，上文中引入人口统计信息的分析技术，对用户进行聚类，为这类用户选取出了近邻用户。

⑵　预测评分

本研究中使用全局数值算法，将计算用户集中用户对指定产品的评价以相似度作为权值，组合产生预测值[5]。

$$P_{xj}=\bar{r}_x+\left(\sum_{y=1}^{n}Sim(x,y)\times(r_{yj}-\bar{r}_y)\right)^{-1} \quad ⑥$$

公式⑥中的 n 是计算用户集中用户数量，P_{xj}为用户 x 对产品 j 的预测评分。

2. 模型工作流程

用户登录知识库资源系统后，会根据其产品评价数据的粘稠度，判断是否需要通过用户聚类操作来为其选取近邻用户，进而以此作为协同过滤算法的计算用户集，为目标用户进行资源推荐。

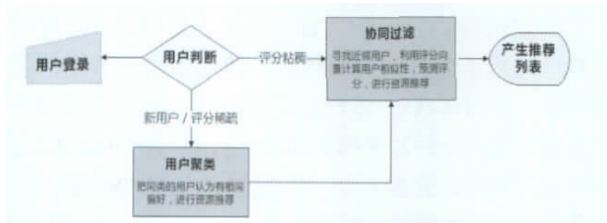

图 3　推荐模型工作流程图

3. 推荐结果的呈现

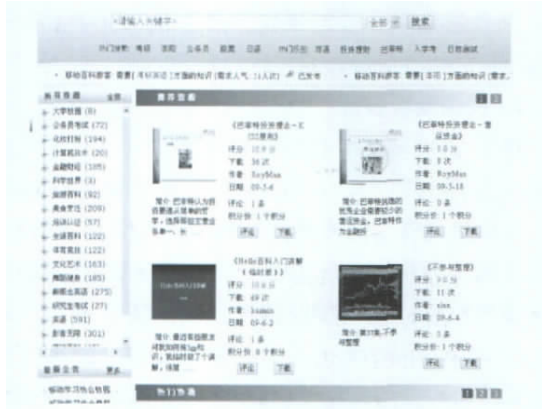
图 4　个性化资源推荐的结果呈现截图

学习资源需求量预测

移动百科网站为移动学习爱好者提供课件资源下载和上传服务，并把所有资源分为 20 大类 70 余小类。对于资源上传者来说，知道哪类资源受欢迎，特别是未来哪类资源受欢迎是十分重要的，很有可能据此提高自己上传资源的下载量，同时也满足了下载者的未来需求。基于课件资源的历史下载记录对该类课件的未来下载量进行预测，实际上就是典型的基于时间序列数据的数据挖掘问题。

1. 时序预测算法

一般数据以时间区分，可分为截面数据和时间序列数据两种。时间序列数据指的是同一变量在不同时间点的观测值，如日数据、周数据、月数据等。CRISP-DM（cross-industry standard process for data mining，跨行业数据挖掘标准过程模型），是用于描述、定义、开发和实现数据挖掘项目的步骤的方法，可使数据挖掘项目的设计开发部署更加快速，成本更加低廉，系统更加可靠，项目更易于管理，因而在各种知识发现过程模型中占据领先位置[6]。

ARTXP 算法[7]，是基于自回归决策树模型的时

序预测算法。ARTXP 算法将数目可变的过去项与要预测的每个当前项相关。该算法是在 SQL Server 2005 中引入到 SSAS，针对预测序列中的下一个可能值进行了优化，特别适合于短期预测。ARTXP 算法首先使用标准的“开窗术”把时间序列数据集转化为便于回归分析的事例集。然后，使用转换好的数据集学习对目标变量产生一个决策树。这个决策树在叶节点产生线性回归，因此，产生了分段的线性自回归模型，然后使用贝叶斯技术学习决策树的结构和参数。ARTXP 算法有以下优势：从数据中学习 ART 模型计算方法效率高；结果模型产生准确预测；ART 模型简单的线性分段预测，比较容易解释。

2. 预测模块设计

基于 CRISP- DM 模型，根据学习者下载学习资源的时间序列数据，以及 ARTXP 算法建立挖掘模型，在后台建立“需求预测”模块，按照资源类别，指出未来一周用户对各类学习资源需求量,用以指导安排课件制作。预测模块结构框架如图 5 所示。其中，Analysis Server 是核心，里面封装了数据挖掘的各类接口和我们所需使用的时序分析算法；数据库用来存放网站数据和预测表；预测辅助程序用于定时更新预测表和重定型挖掘模型；时序预测 Web 服务主要封装与预测有关的数据挖掘结果及数据库中信息的查询，并通过 SOAP 消息为客户端输出预测结果。预测功能模块所使用到的开发工具为 Microsoft Visual Studio .NET 2005，要求安装 Microsoft SQL Server 2005，.NET Framework 2.0 和数据分析工具 SSAS。

⑴ 建立辅助程序

该程序为 Windows 计划任务，每周执行一次，功能是预测表的首次填充和数据更新，从相关表中提取数据归纳汇总填充入预测表，并借助 AMO 完成时序模型的重定型，以保证预测结果的即时性。

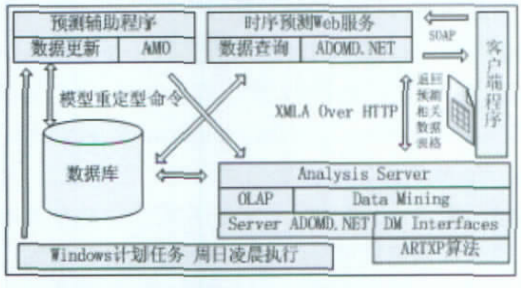

图 5 预测模块解决方案结构框架

首先建立如图 6 所示的预测表，它的数据来源为移动百科网站后台管理系统中的资源下载记录表和课件表。由于下载记录表中的数据是按照时间顺序记录，而预测表按照周记录，并且要对各类型课件进行下载计数汇总，所以不能直接联合查询填充，需要先在内存中建立一个中间表再使用辅助程序的 FillTableAll()或 UpdateTableAll()方法完成周汇总，然后将结果写入预测表。这两个方法的区别在于：前者是第一次运行时填充所有汇总记录，后者只每周执行的时候填充预测表中需要更新的记录。之后调用 AMO 的 Process 方法处理挖掘模型，以达到重定型的目的。最后在 Windows 计划任务中设置辅助程序 .exe 定时执行。

列名	数据类型	允许空
id	bigint	☐
Weekdate	datetime	☑
Category	int	☑
downs	int	☑

图 6 预测表设计

⑵ 建立分析服务项目

在这个项目中根据预测表建立数据源、数据视图、创建和处理预测挖掘模型，并设置安全级别，允许.net 程序有权访问模型。所有操作可在 vs 中或 sql 企业管理器中执行。

使用 DMX 语言来描述建立该挖掘模型：

```
Create Mining Model ForcastCategory(
Weekdate key time,
Category long key,
Downs long continuous Predict
)Using Microsoft_Time_Series
```

然后使用预测表数据训练模型：

```
Insert into ForcastCategory(Weekdate, Category, Downs)
OPENQUERY([ForcastCategory],
    'SELECT [Weekdate]、[Category]、[Downs]
    FROM [dbo]. [ForcastCategory]')
```

部署模型后即可查看到模型的预测结果如图 7 所示。为使客户端页面顺利访问，需增加 IIS user 角色，默认为 IUSR_ 计算机名，并给予数据库及模型读权限。

⑶ 建立后台预测结果浏览页面

主要使用 AdomdClient 类库进行模型预测查询，供管理者浏览。设计一个公共类，用于封装与分析服务器处理事务 (连接、获取 DMX 查询数据等) 的各类操作，便于代码复用。页面后台代码使用 DMX 查

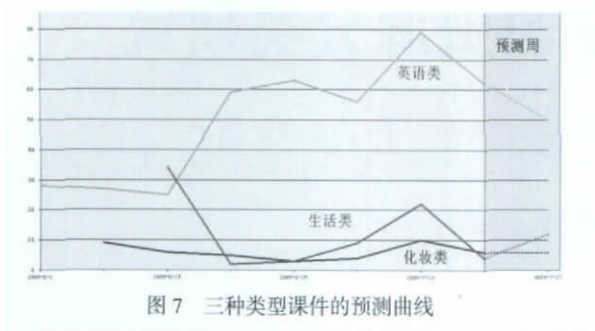

图 7　三种类型课件的预测曲线

询语言来检索预测结果。如查询英语类（d=）课件下周预测下载量，客户端得到的返回表结构的数据，包含预测的日期（'9/20/2010'）和下载量（'50'）：

```
SELECT Category ,PredictTimeSeries(Downs, 1)AS Downs
FROM ForcastCategory
Where Product='1'
```

表 3　客户端预测呈现（只选取三类）

课件类型名	预测下周下载量	所占百分比	变化情况
英语类	50	20%	↓
生活类	12	4. 8%	↑
化妆类	7	2.8%	↑

上表中只列出了三类移动知识资源的需求预测情况，同理，我们可以直观了解移动百科网站上的 17 大类学习资源的需求预测情况，并根据需求趋势，组织安排相应的移动知识资源的制作与上传。预测结果的呈现见图 8。

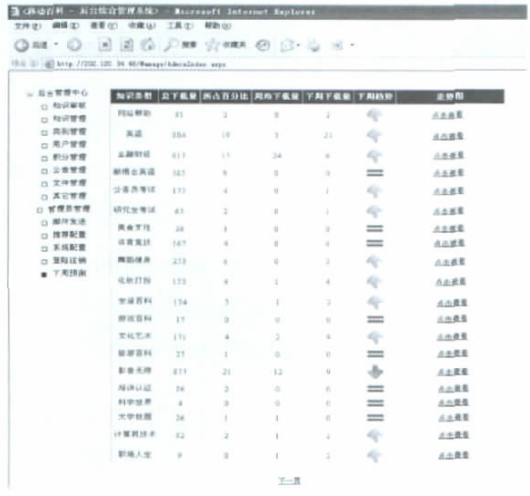
图 8　预测结果

结束语

数据挖掘技术虽然在教育领域应用已取得一定成绩，但往往针对某一局部问题进行孤立的分析解决。本文从一个实际运行的学习网站角度出发，利用"移动百科网站"的客观数据做了一系列分析研究，从人口统计学到移动学习者特征聚类，再由用户聚类推算用户偏好，从而对移动学习资源进行个性化推荐服务，并据此进一步进行资源需求量趋势预测。本文整合并利用了多种数据挖掘技术手段，对移动学习网站运行过程中实际产生的问题进行实验分析，这些问题互相联系，从而使得本次研究具有较强整体性。

[参考文献]

[1] Linden G, Smith B, York J., "Amazon.com recommendations: Item-to-item collaborative filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, 2003.

[2] Zhang Chi, Chen Gang, Wang Minjuan, Wang Huimin, "Cluster Analysis on Students in Mobile-Learning using EM Algorithm," in Distance Education in China, vol.9, no.5, pp. 68-71, 2009.

[3] Wang Huimin, Chen Zeyu, Wang Minjuan, Zhang Chi, "The application of Decision Tree Technology on Gender Diversity Research in Mobile Learning," in Modern Educational Technology, vol. 7, no.4, pp. 30-33, 2009.

[4] Breese J, Hecherman D, Kadie C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," In Proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence, May, 1998.

[5] Gabriela Polcicovaa, Peter Tino, "Making Sense of Sparse Rating Data in Collaborative Filtering via Topographic Organization of User Preference Patterns," in Neural Networks, vol.17, pp. 1183-1199, 2004.

[6] CRISP-DM consortium .CRISP-DM 1.0 .(2000-8-1)[2009-8-25] from http://www.crisp-dm.org/CRISPWP-0800.pdf.

[7] C.Meek, D.M.Chickering, D.Heckerman, "Autoregressive tree models for time-series analysis," in Proceedings of the Second International SIAM Conference on Data Mining, pp. 229-244, April 2002.

收稿日期：2010-11-15

作者简介 刘钢 张驰 王慧敏 陈笑怡。上海交通大学现代远程教育研究中心（200240）。

王敏娟，圣地亚哥州立大学教育技术系。

责任编辑　池　塘

Data Mining in Mobile Learning

Liu Gang, Wang Minjuan, Zhang Chi, Wang Huimin and Chen Xiaoyi

With the rapid development of network technology and communication technology, mobile learning is showing a strong momentum of development. Using a mobile learning website as a research platform, the study intends to investigate issues such as mobile learners' features, personalized recommendation technology for mobile learning materials and learning demand forecasting by analyzing objective data obtained with data mining technology. It is expected that findings from the study will be valuable to mobile learning researchers, website operators, and policy-makers with knowledge-based services.

Keywords: mobile learning; data mining; customer clustering; cooperating filter; blending recommendation technology; time-series analysis

Technology-supported Collaborative Deep Learning: From the Perspective of Higher-order Thinking

Duan Jinju

Higher-order thinking is a core feature of constructivism. How to facilitate learners' higher-order thinking is a priority in instructional design. Using the instruments of in-depth interview, online observation and content analysis, the study investigated learners' interaction in a network-based course. Findings indicate that the majority of their talk was higher order and that procedural, social and lower-order talk was less evident but present in their talk in reduced proportions. Factors which affected their higher-order thinking included the topic, teacher's monitoring, learning motivation, use of tools, and flexibility in learning time in reduced proportions. In terms of communication tool, BBS was their main means of interaction, which means low-tech tools were most popular among learners.

Keywords: technology; higher-order talks; higher-order thinking; instructional design

On the Impact of Animation in CSCL on Collaborative Learning

Yang Wenyang and Zhao Wenxia

Interaction is a basic activity which can stimulate and generate common knowledge in collaborative learning while conversation is the basis of collaboration. This study reports on an experimental research on the impact of animation on Computer Supported Collaborative Learning (CSCL) with 60 third-year undergraduate students of Educational Technology as subjects. Experiment findings show that the use of animation in CSCL proved to be more conducive to learning in certain circumstances in comparison with static course materials. Nevertheless, too much animation and repeated use of animation to introduce key messages turned out to impede collaborative learning. Research findings are relevant to the design of interaction as well as development of resources and tools in CSCL.

Keywords: animation; collaboration; Computer Supported Collaborative Learning; multimedia; interaction