SOM聚类技术在读者行为分析中的应用

马芳
（烟台工程职业技术学院图书馆，烟台 264006）

摘要：随着数字图书馆技术的发展，图书馆每天产生大量的数据，针对这些海量数据，采用数据挖掘技术中的自组织映射神经网络（SOM）算法，根据读者借阅行为特征对读者进行聚类，得到不同阅读兴趣和需求的读者群，并通过测试验证该算法是有效可行的。

关键词：数据挖掘；RFM模型；SOM聚类
分类号：G250
DOI：10.3772/j.issn.1673—2286.2014.06.010

1 引言

图书馆是图书、期刊、学位论文、音像等文献信息资源的收藏中心。随着数字图书馆技术的飞速发展，图书馆信息管理系统中每天积累大量的文献数据和读者数据，这些数据既包含读者的真实行为特征，也隐含着读者的个性化信息需求。面对如此海量数据，图书馆很难从中发现读者的信息需求，读者也很难从中找到自己所需的信息资源。因此，需要引入特定的信息处理技术，从大量数据中发现其中隐藏的、有价值的信息，数据挖掘技术正是实现该任务的关键技术。本文采用数据挖掘技术中的聚类算法对图书馆海量信息进行分析，从中找出隐藏在其中的内在规律，不仅可以预测未来可能出现的借阅行为，在很大程度上可以提高图书馆信息服务的质量以及加强图书馆资源的有效利用。

2 相关研究综述

数据挖掘是指从大量的、有噪声、不完全的、模糊的、随机的数据库中，提取隐含在其中的、人们预先不知道的、但又潜在有用的信息和知识的过程[1]。它起源于20世纪80年代后期，早期主要应用于商业领域。后来，随着数据挖掘技术的不断发展及广泛应用，也成为图书馆界的研究热点。探讨数据挖掘在图书馆中应用的论文最早出现于1997年[2]，国内外多位学者针对图书馆的特定应用提出了数据挖掘应用的架构和模型，比较有代表性的有：

加州大学Michael Cooper教授对加州大学数字图书馆的流通数据进行挖掘分析，并设计了相关模型，采用聚类、时间序列分析等方法，发现不同读者在查询时间、次数等方面具有不同的特点，从中分析出读者的阅读兴趣，预测读者的行为规律。

南京理工大学的孙健波，将数据挖掘技术与图书馆个性化服务结合在一起，采用聚类分析和关联规则分析，挖掘出读者对图书馆图书资源利用的关联内容，了解读者对图书馆服务的使用程度，为个性化服务提供决策依据。

数据挖掘是用于指定数据挖掘任务要求的模式类型，在实际应用中，根据不同的任务需求采用不同的功能。其功能主要有关联和相关分析、分类、预测、聚类、离群点分析和演变分析等。其中，聚类是人类一项最基本的研究事物分类的方法。具体地讲，就是根据事物本身的属性特征进行分类，划分的原则是在同一类中的对象相互之间具有较高的相似度，而不同类中的对象差别较大。类内的相似性越大，类间差别越大，聚类就越好[3]。聚类分析是一个活跃的研究领域，目前，已经存在许多经典的算法。常用的聚类方法主要有：基于划分的方法，如K-Means算法；基于层次的方法，如BIRCH算法；基于模型的方法，如人工神经网络（如SOM）方法。每种算法都有其自身的功能特点，因此，不同算法的选用，对聚类结果质量会产生一定的影响。与已有聚类方法相比，SOM神经网络的优点是能够实现自学习，

无须外界给出评价函数[4]，并且该网络采用各神经元（特征参数）之间的自动组织去寻找各类型间固有的、内在的特征，对于解决特征参数交错混杂、非线性分布的类型识别问题是非常有效的，而图书馆读者行为识别本身也是一种复杂多变的问题，所以其对读者聚类研究也是相当有效的。即本文采用SOM神经网络算法对读者进行聚类分析。

SOM（Self Organization Map）即自组织映射神经网络，它是由芬兰神经网络专家Kohonen提出的一种无监督学习的神经网络模型[5]。最初SOM神经网络主要用于机械工程中的仪器检测，后来该网络被作为一种数据分析的标准广泛应用于数据挖掘领域，经过不断的发展研究和实践证明，其在聚类分析及数据可视化方面有着广阔的应用前景。

基于以上研究现状及背景分析，在图书馆对读者进行个性化服务环境下，采用数据挖掘技术中的聚类算法，对图书馆中的流通数据进行聚类分析，将读者划分为不同阅读兴趣的群体，分析不同读者群的信息需求，对提高图书馆的服务质量是非常有益的。

3 数据来源及研究方法

3.1 数据来源

实验样本数据来自本校的图书馆自动化管理系统，从该系统的流通日志中可以导出读者的借阅历史记录，从中提取与聚类任务相关的字段，如读者条码、借书日期、还书日期，统计出借阅频率、第一次借阅日期、最近一次借阅日期，生成读者聚类信息库。本文选取2010级的读者借阅信息，时间范围在2010年9月15日至2012年12月31日，数据选择的时间跨度近三年，接近高职专科学生大学学习生活的完整时期（不包含实习期）。为了便于数据分析，删除零借阅读者，共得到可利用的读者记录680条，从中选取300条作为训练样本，其余380条作为测试样本。

3.2 研究方法

3.2.1 SOM神经网络训练算法

SOM神经网络将输入对象映射至最理想的输出层结点中，若输入对象具有较高的相似度，其对应的输出层结点应该相同或具有较小的欧氏距离。训练算法步骤如下：

步骤1：初始化。对输入神经元和输出神经元相连接的权值赋予较小的随机数。

步骤2：输入样本。从训练样本集中随机选取样本输入网络的输入层，对每一个样本执行下述步骤。

步骤3：寻找获胜神经元。计算欧氏距离，从中找出距离最小的获胜节点。

步骤4：定义优胜邻域。以获胜节点为中心确定t时刻的权值调整邻域，邻域大小随训练时间逐渐降低。

步骤5：调整权值。

步骤6：结束检查。所有样本输入并计算后完成一次迭代，用学习率是否减少到零或者某个预定的正小数为条件，当满足条件时结束并退出，否则回到步骤2进行下一次迭代。

3.2.2 基于SOM算法的读者聚类模型

SOM聚类是一个无监督的机器学习问题，其工作流程主要包含训练、测试和评估三个阶段，在此期间需要进行数据表示、聚类算法进行训练产生稳定的网络模型、用训练达到最优的网络模型进行测试、聚类性能评估等步骤。本文设计SOM读者聚类模型如图1所示。

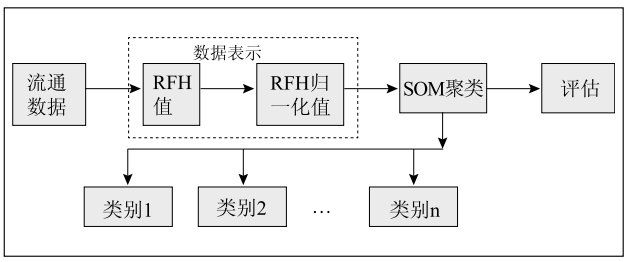

图1 SOM读者聚类模型

（1）数据表示

图书馆的服务方式以信息传递为主，其与市场营销模式有很大的相似性。本文引用客户关系管理领域中的RFM分析模型来进行数据描述，该模型分别通过近度R、频度F、值度M三个属性值来描述客户的重要程度和客户类型[6]。依据以上模式构建，针对读者的借阅特性和需求，可将传统 RFM 模型重新定义为：读者最近借阅日期R（Recency）、读者借阅频率F（Frequency）和读者借阅总时间T（Time）。由于图书馆经常会出现图书超期未还的现象，因此T值较大的读者不一定是活

跃读者。所以，根据实际情况，并通过征求专家意见，本文对指标进行适当调整，采用读者借阅保持日期H（Hold time）来代替读者借阅总时间T。

将读者借阅行为通过RFH模型表示成对应数值后，由于三个数值所代表的意义及范围不同，需要将它们进行标准化处理，本文采用归一化处理方法。如，对R值的归一化处理采用如下公式：

$$s\_r\_value = \frac{r\_value - r\_value_{min}}{r\_value_{max} - r\_value_{min}} \quad (1)$$

其中，r_value_{min}是最小日期差，r_value_{max}是最大日期差。F、H值的归一化与R值相同。

（2）SOM聚类

将归一化处理后的RFH数值表中的各项作为数据源，进行SOM聚类。本模型分为训练和测试两个模块。训练模块是将训练样本集输入到创建好的网络模型中，对SOM网络进行训练，通过不断调整网络节点权值直至输出最优解。聚类模型建立后，需要用测试样本集对模型性能进行检测，即将测试样本集输入到训练模块训练好的网络模型中，对输出的聚类结果和实际类别进行比较，验证其聚类精度和实用性。

（3）评估

对聚类网络进行性能评估的方法比较多，本文采用准确率（Precision）、召回率（Recall）和F_1三个指标量化值测试聚类结果，由于准确率和召回率有时会出现矛盾的情况，所以需要对它们进行综合考虑，即采用综合评价指标F_1作为准确率和召回率的调和平均值。

$$F_1 = \frac{Precision \times Recall \times 2}{Precision + Recall} \quad (2)$$

4 读者聚类实验及结果分析

对以上提出的聚类方法，采用MATLAB软件的SOM神经网络工具箱，利用其提供的初始化、训练、激活等函数构建网络模型，通过具体实验数据进行网络训练与测试和评估。

4.1 实验数据

本文选取2012年12月31日为分析点，统计RFH模型中的各指标值。将读者信息通过RFH模型表示，然后将RFH数值进行归一化处理。经查询，r_value_{min}为2，r_value_{max}为795；f_value_{min}为1，f_value_{max}为86；h_value_{min}为0，h_value_{max}为792。部分结果如表1所示。

表1 RFH模型指标值及其标准化

reader_id	r_value	s_r_value	f_value	s_f_value	h_value	s_h_value
201010002	376	0.4716	21	0.2353	434	0.5480
201010005	49	0.0693	38	0.4353	783	0.9886
201010006	33	0.0391	10	0.1059	631	0.7967
201010009	194	0.2421	1	0	0	0
201010010	62	0.0757	18	0.200	427	0.5391

4.2 SOM训练与测试

（1）SOM网络结构的确定

由于采用RFH值作为网络输入变量，所以输入层的神经元个数为3个。

在SOM网络中，竞争层也是输出层，竞争层的神经元个数与训练样本类别的多少有关，神经元过多或过少会出现聚类过细或过粗的情况。而训练步数的大小也影响着网络的聚类性能，可以分别设置不同的训练步数，观察其聚类性能，然后根据具体需要逐步确定竞争层神经元的个数。本文根据图书馆读者行为特性，将读者分为活跃读者、一般读者、懒惰读者3种类型，由此可确定输出层为2×2的二维结构，即竞争层神经元的个数为4个。

（2）SOM网络的训练

网络结构确定后，根据SOM神经网络训练算法进行网络的训练。首先，对网络的初始权向量进行设定，通常的做法是采用对输出层各权向量赋一个较小的随机数，但是由于初始权向量与输入向量越接近，会使网络权值的调整范围越小。所以，本模型中，先计算出全体样本的中心向量（平均值），然后在该中心向量基础上加一个小随机数作为权向量的初始值。各权向量初始化后，将归一化处理后的300份训练样本逐一输入网络的输入层进行训练，初始训练次数设为500。对输入的每条读者的RFH记录，SOM网络自动为其寻找最相近的神经元，找到获胜神经元后，将该读者条码记录在该神经元所代表的读者簇类号下。在经过270次训练后，网络权值不再发生变化，网络将所有的训练样本完

成聚类训练。

（3）SOM网络的测试

把经过网络训练后稳定下来的权值作为聚类测试知识，将选出的380条测试样本输入到已经训练好的网络模型中，进行聚类。

4.3 实验结果评估

分别对每个类群的三个聚类效果评估指标——召回率、准确率、F_1值进行实验结果评估，评估结果如表2所示。

表2 SOM网络聚类测试结果

类别	样本数	召进	错误	召回率	准确率	F_1值
1	146	158	28	94.5%	87.3%	90.7%
2	198	190	32	89.8%	93.7%	91.7%
3	36	39	11	88.8%	82.1%	85.3%
合计	380	387	71	91.6%	89.9%	90.7%

从表2可以看出，本模型根据读者行为进行聚类的召回率、准确率以及F_1值都在80%以上，其中有部分值达到90%以上，目前大多数性能较好的聚类模型的召回率、准确率以及F_1值都在80%以上，从而说明，本文提出的模型对图书馆读者行为聚类是可行、有效的。

5 聚类结果分析

通过以上实验，SOM网络将参与聚类的读者样本分为3个聚类，每类读者所占比例如图2所示，部分读者所在类群如表3所示，对各聚类的读者群进行分析如下：

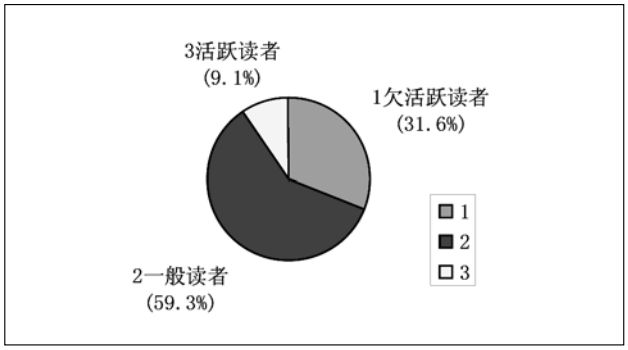

图2 每类读者所占比例

表3 部分读者类表

Reader_id	Cluster
201010002	2
201010005	3
201010006	2
201010009	1
201010010	2

由图2可知，聚类为3的读者只占参与分析的总读者数量的9.1%，但这类读者的借阅量高，是图书馆的重要支持者，即该类为活跃读者。对这类读者，图书馆可以适当延长借阅期限、增大借阅册数，进一步激发读者的读书热情。聚类2的读者所占比例为59.3%，虽然这部分读者个人借阅量较低，但包含人数较多，总借阅量较高，是图书馆的重要使用者，即该类为一般读者。对于这部分读者，图书馆需要进一步挖掘他们的读书潜力，比如进行主动推荐服务，让读者及时了解图书馆各项服务动态，鼓励读者参与图书荐购活动，增大读者感兴趣的图书藏量，提高读者借阅图书的积极性。聚类1的读者所占比例为31.6%，通常这类读者的学习成绩普遍较差，缺乏学习和读书的兴趣，即该类为懒惰读者。对于这类读者，图书馆需要通过开展多种讲座、座谈等活动吸引他们进入图书馆，使读者感受到图书馆特有的文化氛围，懂得学习知识的重要性，从而，唤醒他们的读书兴趣。

从表3可看出，条码为201010002的读者，被划分到一般读者中，虽然该读者的借阅频率（21次）较高，高于借阅频率平均值（经统计，借阅频率平均值为18次），但其未借阅图书的时间长达一年。针对这种情况，需要根据实际情况找出具体原因。通过调查发现，随着校园网络覆盖率的提高，该类读者的课余时间大多用于上网。对于这部分读者，图书馆需要积极采取措施，加强引导，从他们感兴趣的方向入手，比如加强图书馆网站建设，通过网站加大图书馆各项服务的宣传，做好新书推荐、参考咨询、在线问答等栏目，加强与读者的沟通与互动，有效地调动读者读书的积极性，让他们回归到活跃读者群中。

将SOM神经网络输出结果与BP神经网络以及kNN算法输出结果进行对比，以检验网络性能，对比结果如图3所示。

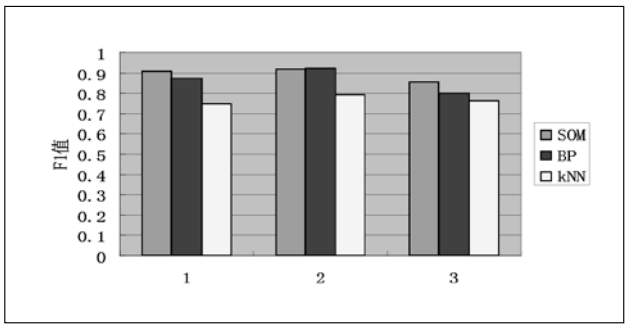

图3 不同聚类算法的F_1值比较

每种聚类都产生3种不同类别的读者群，由图3可以看出，对于1、2、3每种类别的F_1值，SOM网络相对于BP神经网络、kNN算法有明显提高，即SOM神经网络聚类效果更好。这是由于：BP算法本身具有可能会产生局部最小值的缺陷，使得BP神经网络在进行训练时可能收敛到局部极小点，但不能保证其为误差平面的全局最小值，也就是全局搜索能力较弱。kNN算法要事先给出期望生成类群的数量，而且该算法不适合发现大小差异较大的类群，并对“噪声”和孤立点数据非常敏感。因此，该算法通常只有局部最优解而非全局最优[7]。而SOM 神经网络的训练方式无需事先确定类群数目，可避免主观因素对聚类效果的影响，并且该网络具有一定的“鲁棒性”，少数单元的信息不会影响网络对整体信息的判断能力，从而为整体数据的合理分析提供保障。当然，SOM 神经网络也有它的不足之处，比如：需要不断地平衡网络的学习速度与权值向量之间的关系；神经网络中一个神经元的初始权值向量离输入向量太远，就会导致不能在竞争中获胜，因而得不到学习，形成“死”神经元等。

6 结语

本文将数据挖掘技术和图书馆自动化管理系统相结合，根据读者的借阅行为对流通日志中的大量数据进行聚类，并通过分析聚类结果掌握不同读者群的特性与需求，为服务读者和科学管理提供可靠的依据。基于图书馆数据的丰富性和多样性，有必要进一步加大对这些数据的挖掘工作，本文只对读者聚类进行研究，还有其他更多的内容需要开展。比如，图书MARC信息及图书流通信息的挖掘，其对馆藏建设、信息咨询等业务具有科学的指导意义。总之，随着图书馆数字化技术的飞速发展，加强图书馆的信息处理水平，促进读者的信息获取能力，是图书馆在信息时代发展的必由之路。

参考文献

[1] HAN Jiawei, KAMBER M.数据挖掘概念与技术[M].北京:机械工业出版社,2011:3-6.

[2] 曹美琴.数据挖掘在图书馆个性化服务中的应用[D].陕西:西北大学,2008:11-12.

[3] TAN Pang-Ning, STEINBACH M, KUMAR V.数据挖掘导论[M].北京:人民邮电出版社,2011:306-307.

[4] TSAO E C, BEZDEK J C, PAL N R. Fuzzy kohonen clustering networks [J].Pattern Recognition, 1994, 27(5): 757-764.

[5] KIELY M. Word-of-mouth marketing [J]. Marketing, 1993(9): 6-11.

[6] 林盛,肖旭.基于RFM的电信客户市场细分方法[J].哈尔滨工业大学学报,2006(5):759-760.

[7] 谢雄程,刘之家.基于聚类与分类混合算法的应用研究[J].广西师范学院学报:自然科学版,2011,28(3):82-87.

作者简介

马芳，女，1975年生，管理学硕士，烟台工程职业技术学院图书馆馆员，研究方向：数字图书馆技术、用户研究。E-mail：Lgmafang@163.com。

User Behavior Analysis in University Library by SOM

MA Fang
（Library, Yantai Engineering and Technology College, Yantai 264006, China）

Abstract: With the development of digital library technology, library produces a large number of data every day. Based on these data, using the algorithm of Self Organization Map (SOM) in data mining technology, clustering readers according to their borrowing behavior, we classify different reading interest and demand of readers. Through the validation, the algorithm proves feasible and effective.

Keywords: Data mining; RFM model; SOM clustering

（收稿日期：2014-02-18）