文章编号：1008-8717（2007）09-0079-02

如何在 Web 挖掘中使用聚类算法

张 丽 霞
（菏泽学院计算机与信息工程系，山东 菏泽 274000）

摘 要：首先介绍了聚类的概念，然后提出了用模糊聚类算法对 Web 事务进行聚类。在聚类的相似性度量上，不再单纯地以访问次数或浏览时间来度量，而是采用用户浏览离散化时间为度量。该算法比传统算法准确性高，运行时间少，扩展性好。

关键词：聚类；模糊聚类；日志挖掘
中图分类号：TP3 **文献标识码**：A

1. 聚类的概念

聚类是把一组个体按照相似性归成若干类别，即“物以类聚”,它的目的是使得属于同一类别的个体之间的距离尽可能的小，而不同类别上的个体间的距离尽可能的大。在数据挖掘领域，聚类方法包括统计方法、机器学习方法、面向数据库的方法。

1.1 聚类分析的分类

聚类分析按照某个相似测试将未标记的样本集分成若干个类，使同一类中的样本尽可能地相似，不同类中的样本尽可能地不相似。按照聚类结果表现方式的不同，现有的聚类分析算法可以分为：硬聚类算法、模糊聚类算法和可能性聚类算法。在硬聚类算法中，分类结果用样本对各类的隶属度表示。样本对某个类别的隶属度只能是 0 或 1。样本对某个类别的隶属度为 1，表示样本属于该类;样本对某个类别的隶属度为 0，则表示样本不属于该类。早期的聚类算法都是硬聚类算法，硬聚类算法容易陷入局部极值。

模糊聚类产生于 60 年代末，是聚类分析与模糊集理论相结合的产物。模糊聚类算法与硬聚类算法相比，提高了算法的寻优概率，但模糊聚类的速度要比硬聚类慢。

可能性聚类算法是聚类分析与可能性理论的结晶。可能性聚类算法也容易陷入局部极值，但可能性聚类算法抑制噪声能力很强。

1.2 Web 日志挖掘对聚类算法的特别要求

聚类是一个富有挑战性的研究领域，它的潜在应用提出了各自特殊的要求。数据挖掘对聚类的典型要求如下[1]：

（1）可伸缩性

许多聚类算法在小的数据对象集合上工作得很好；但是一个大规模服务器日志库可能包括几百万条记录，在这样的大数据集合样本上进行聚类可能会导致有偏差的结果。我们需要具有高度可伸缩性的聚类算法。

（2）处理不同类型属性的能力

许多算法被设计用来聚类数值类型的数据。但是，应用可能要求聚类其他类型的数据，如二元类型、分类/标称类型、序数型数据，或者这些数据类型的混合。

（3）发现任意形状的聚类

许多聚类算法基于欧几里德距离或者曼哈坦距离度量来决定聚类。基于这样的距离度量的算法趋向于发现具有相近尺度密度的球状簇。但是，一个簇可能是任意形状的。

（4）处理噪声数据的能力

绝大多数现实世界中的服务器日志库中都包含了孤立点、空缺、未知数据或者错误的数据。一些聚类算法对于这样的数据敏感，可能导致低质量的聚类结果。

（5）对于输入记录的顺序不敏感

同一个数据集合，当以不同的顺序提交给同一个算法时，可能生成差别很大的聚类结果。开发对数据输入顺序不敏感的算法具有重要的意义。

（6）高维性

一个数据库或者数据仓库可能包含若干维属性。在高维空间中聚类数据对象是非常有挑战的，特别是考虑到这样的数据可能非常稀疏，而且高度偏斜。

（7）可解释性和可用性

用户希望聚类结果是可解释的，可理解的，和可用的。

2. Web 日志挖掘的算法描述

2.1 从 Web 日志生成 Web 事务集合

Web 日志记录了用户访问站点的信息，通过对日志的处理，可以得到一个用户在一段连续时间范围内的访问页面序列——Web 事务。我们利用最大时间间隔法来获取 Web 事务。

2.2 用户浏览时间的离散化表示方法

收稿日期：2007-03-23

在聚类的相似性度量上，不仅要考虑在 Web 事务中对某页面的访问次数，而且要考虑在该页面上的浏览时间，故本文提出将离散化技术应用到用户浏览时间的表示上，将时间属性域划分为区间，用区间的标号来代替实际的时间值。可按照用户在网页上的浏览时间，将 Web 访问分成经过、普通阅读和详细阅读三种，浏览时间离散化见表 1。

表 1　浏览时间与离散值对照表

离散值	访问情况
0	没有访问
1	0min<浏览时间≤X_1min
2	X_1min<浏览时间≤X_2min
3	X_2min<浏览时间

其中，X1 和 X2 作为输入参数，表示浏览时间边界值，可以在初始化的时候对输入参数进行赋值，这样可以针对用户浏览习惯灵活设置。采用了这种浏览时间离散化的表示方法，用户只要访问了页面，即使时间再短也有离散化时间（离散值为 1）；用户在页面上的浏览时间即使很长，也有离散化时间（离散值为 3）。这样就有效地避免了在进行 Web 事务相似性度量时，在采用连续时间情况下忽略用户浏览次数的情况，也避免了单纯地考虑访问次数而不考虑访问时间的问题。

2.3 Web 站点用户访问矩阵表示

根据 Web 事务建立 Web 站点的用户访问矩阵[2]，即

$$M_{URL-Session}=\underbrace{\left.\begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,l} & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,l} & s_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{l,1} & s_{l,2} & \cdots & s_{l,l} & s_{l,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{m,1} & s_{m,2} & \cdots & s_{m,l} & s_{m,n} \end{bmatrix}\right\}URL}_{Session}$$

式中：m 为网站 URL 的个数；n 为 Web 的事务数；S_{ij} 是在第 j 个 Web 事务上对第 i 个 URL 的离散化浏览时间。

2.4 原始数据标准化

要构造模糊相似矩阵，必须对数据进行标准化处理，使数据压缩到[0,1]闭区间内。设有 n 个对象 E1，E2，…，En，每个对象具有 m 个对象指标 y1，y2，…，ym。xij 表示第 i 个对象的第 j 个指标。

n 个对象第 j 个指标的平均值和标准差分别为：

$$\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}X_{ij}$$

$$s_j=\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2\right]^{\frac{1}{2}}$$

原始数据标准化为

$$x'_{ij}=(x_{ij}-\overline{x_j})\big/S_j$$

运用极值标准化公式，将标准化数据压缩到[0，1]内，即

$$\overline{x_{ij}}=\frac{x'_{ij}-x'_{\min j}}{x'_{\max j}-x'_{\min j}}$$

式中：$x'_{\min j}$ 和 $x'_{\max j}$ 分别是 $x'_{1j},x'_{2j},\cdots,x'_{nj}$ 中的最小值和最大值。

2.5 构建模糊相似矩阵

构建模糊相似矩阵 R^F：

$$R^F=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}$$

式中：r_{ij} 为两个对象 E1 与 E2 之间相似程度的变量，r_{ij} 越接近于 1，表明这两个对象越相似。r_{ij} 可以用距离法、夹角余弦法、相关系数法、主观评分法等来确定。常见的距离函数有 Hamming 函数、Minkowski 函数、Euclidean 函数和 Maximum 函数等，其中最常使用的是 Hamming 函数：

$$r_{ij}=\begin{cases}1-c\sum_{i=1}^{m}\left|x_{ik}-x_{jk}\right| & i\neq j\\ 1\end{cases}$$

常数 c∈[0，1]，可以选取适当的值即可。

2.6 模糊相似矩阵的传递闭包构造算法

根据上面步骤得到的模糊相似矩阵 RF 满足自反性、对称性，但它不满足传递性，因此必须求其传递闭包。求模糊矩阵的传递闭包是一项复杂的工作，所需的时间和空间巨大。为了解决这个问题，国内外学者研究了一些方法，如直接聚类法、编网法、最大树法[3]和图论法[4]。本文采用图论法，将模糊相似度与图论相结合，来得到一种求模糊相似矩阵传递闭包的简单有效的新方法。它主要有以下步骤组成：

步骤 1：对模糊相似矩阵 R^F 中对称相似度的一半依次从大到小排列，并记下行数和列数。

步骤 2：生成只有结点没有边的无向图 G，每一个结点对应论域中的每一个个体对象。

步骤 3：按生成的相似序列依次在图 G 上连接相应的结点，并标上权重，即两个个体对象之间的相似度。如果一条边加到图 G 后，图 G 出现回路，就不加入该边。整个加入过程到图 G 连通为止，这样得到图 G 是一颗树。

（下转 83 页）

没有任务时自动到聊天室与其他机器人甚至人类聊天。比如棋类聊天室机器人甚至可以和深蓝这样的国际象棋高手过招或者交流怎样下好国际象棋。

4．2．9 智能搜索引擎

智能机器人可以利用该搜索引擎在整个互联网上搜索他所需要的相关信息。

5．实验

湖北民族学院陶勇老师和笔者在华南理工大学计算机系进修学习，华南理工大学计算机系的肖南峰博士在日本留学和澳大利亚做访问学者期间完成有关智能机器人方面的研究，取得了丰硕的成果。现在华南理工大学领导着一个关于智能机器人方面的应用。目前关于网格在智能机器人方面的研究已经取得了很大的进展。

6．结束语

将网格作为智能机器人的超级大脑，会使智能机器人获得超乎想象的计算能力和存储能力，同时互联网上丰富的信息资源使得智能机器人信息知识能力强大。可以想象拥有网格这样的超级大脑的智能机器人也许会像人类一样的聪明。

参考文献：

[1]周其方，陈万米，费敏锐.基于 Delaunay 三角形网格的 RobCup 路径规划算法研究[J].系统仿真学报，2004，第 16 卷（6）:1158~1161.

[2]徐德等.基于网络的机器人跨平台远程时控[J].武汉大学学报(信息科学版)，2003，第 28 卷(2):34~37.

[3]洪伟.智能机器人系统中局部环境特征的提取[J].机器人，2005，第 25 卷（3）: 45～50.

[4]王宇龙.智能机器人的研究[J].微纳电子技术，2003，7/8 期.

作者简介：胡坤华（1963-），男，湖北建始人，湖北民族学院信息工程学院讲师，硕士，主要从事计算机应用研究。

（上接 80 页）

步骤 4：建立模糊等价矩阵 R^*，其中 $r^*_{ij}(i,j=1,2,3,\cdots,n)$ 是矩阵 R^*的元素。

（1）在 r^*_{ij} (i=1，2，3，…， n)处记下 1。

（2）令 r^*_{ij} (i≠j)为树 G 上结点 i、j 通路上的最小权值。

2.7 采用 λ 截矩阵法进行模糊聚类分析

λ 截矩阵是联系模糊关系与普通关系的桥梁。模糊相似矩阵 R^F 经过改造得到模糊等价矩阵 R^*后，对于任意 λ ∈[0,1]截取的 λ 截关系所对应的 λ 截矩阵 R^*_λ，每一个 λ 截矩阵 R^*_λ可以决定一个 λ 水平的分类。随着 λ 从大到小取值，R^*_λ所对应的 X 的分类不断发生变化，使分类形成一个动态的聚类图，这种聚类分析方法称之 λ 截矩阵法。λ 介于 0 和 1 之间，λ 取得越大，分类精度越高，一个元素属于多个子类的可能性就越小，有可能使域内元素各成一类；反之，λ 取得越小，则分类越粗糙，一个元素同时局于多个子类的可能性就越大，有可能使域内所有元素都聚成一类。这样都失去了聚类的意义，所以 λ 的设定要适中[6]。

2.8 Web 事务聚类

如前所述，矩阵的列矢量反映了在一个 Web 事务中对站点 URL 的访问情况。聚类时可对矩阵列向量进行标准化处理后，再进行模糊相似矩阵的构筑，在此基础上用图论法得到模糊等价矩阵，最后设置置信水平进行聚类。在个性化设计中，常需要对在线的 Web 事务进行归类。先将 Web 访问矩阵按 Web 事务所属的类别组织成事务矩阵，然后将活动 Web 事务加到该矩阵中进行归类操作。将要归类的活动 Web 事务的浏览行为作为一列矢量加到该矩阵中，以形成新的矩阵，并对该矩阵的行矢量进行处理，形成相似矩阵，再采用 λ 截矩阵法进行聚类分析。Web 事务相似程度最高的类别矢量就是该活动 Web 事务的类别归属。

3. 结论

Web 日志挖掘是 Web 挖掘领域中一个重要的研究方向。它对于发现用户浏览网站的行为规律，改善页面之间的超链接结构，提高整个 Web 系统的性能等方面都具有十分重要的意义 。本文介绍的算法在数据量大时，聚类速度较慢，还有待于进一步提高。

参考文献：

[1]邢东山，宋擒豹，沈钧毅.一种新的从 Web 事务模糊聚类算法的研究.西安交通大学学报，2002，R44-47.

[2]张文升.基于 Web 日志的数据挖掘的研究.辽宁工程技术大学硕士学位论文，2005.

[3]彭祖赠等.模糊数学及其应用[M].武汉:武汉大学出版社，2002.

[4]骆洪青，吴小俊，曹齐英.模糊聚类分析的一种新方法研究[J].华东船舶工业学院学报，2000，14(3): 24-27.

作者简介：张丽霞（1979－），女，山东省菏泽市牡丹区人，菏泽学院助教，在读硕士，研究方向：计算机技术。