基于相似性的商品陈列研究

杨通辉,高 玲,臧 丽

(山东师范大学 信息科学与工程学院,山东 济南 250014)

摘 要:利用聚类的基本知识,根据不同顾客购买商品的相似性的大小,提出了运用 K-means 聚类算法。利用相似度代替欧氏距离,对该网络进行聚类分析,划分出相似性大的顾客群体,并根据每个群体中顾客购买每类商品占总商品数的比例进行排序,从而为商品陈列提供依据。

关键词:聚类;K-means 聚类算法;相似性;商品陈列

中图分类号:TP399 **文献标识码**:A **文章编号**:1674-7720(2012)05-0059-03

Commodity display research based on similarity

Yang Tonghui, Gao Ling, Zang Li

(College of Information Science and Engineering, Shandong Normal University, Ji'nan, 250014, China)

Abstract: This paper, by using the clustering of basic knowledge, according to different customer to purchase the commodity magnitude of similarity, using similarity instead of euclidean distance, the commodity network clustering analysis, divides the similarity in the types of goods, and according to the customer to purchase each group for each type of goods accounted for the proportion of the number of sorted goods, and thus provides the basis for commodity display.

Key words: clustering; K-means clustering algorithm; similarity; commodity display

随着经济的发展,商品的种类越来越多,作为顾客自由购物场所的商店,可利用有限的营业空间,在顾客浏览商品时,刺激顾客的购买欲望,达到扩大销售的目的。商品的陈列在销售过程中扮演着重要的角色,是商品沉默的推销员[1]。因此如何合理地对商品进行陈列[2],成为商店推销过程的一个必须要考虑的问题。由于不同顾客购买的商品之间具有一定的相似性,可以根据不同商品间的相似性,构造具有关联性的商品网络[3]形成聚类,并根据不同顾客购买商品的相似性的大小,运用 K-means 聚类算法,利用相似度代替欧氏距离,对该商品网络进行聚类分析[4],划分出相关性大的顾客群体,并根据每个群体中顾客购买每类商品的均值占总商品数得比例进行排序[5],从而得到商品陈列的依据,这样顾客在浏览商品时,便会刺激其购买欲望,进而达到扩大销售的目的。如图 1 所示。

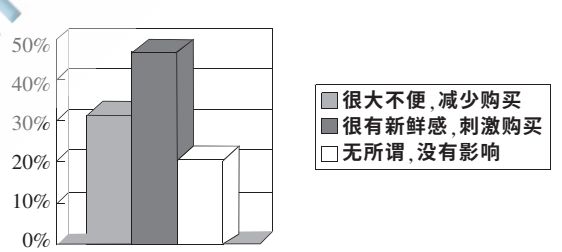


图 1 商品陈列变化对消费者的影响

1 聚类分析的理论基础

1.1 聚类简介

聚类[6](Clustering)是数据挖掘中一种重要的挖掘方法,它是将物理或抽象对象进行分组并将相似的对象归为一类的过程。聚类分析将物理或抽象对象分为几个群体,在每个群体内部,对象之间具有较高的相似性,而在群体之间相似性则比较低。聚类算法大体可以划分为:划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法[7]。

1.2 K-means 聚类算法简介

K-means 算法[8]属于聚类方法中的一种划分方法,该算法具有较好的可伸性和很高的效率,适合处理大文档集。K-means 算法将一组物理的或抽象的对象,根据它们之间的相似程度分为若干组,其中相似的对象构成一组。它采用欧式距离作为相似性的评价指标,即认为两个样本的距离越近,其相似度越大。其以最大欧式距

离原则选取新的聚类中心，以最小欧式距离原则进行模式归类。

欧式距离公式：$d(x,y)=\{\sum_i |x_i-y_i|^2\}^{\frac{1}{2}}$

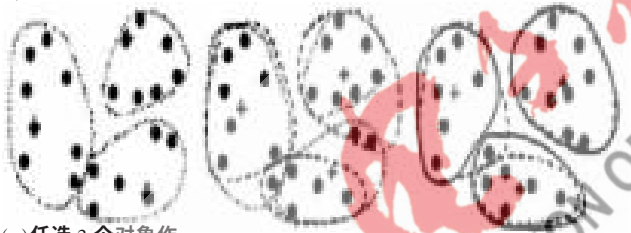
平方误差准则函数形式：$JC=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\|x_i-m_j\|^2$，$k$ 为要形成聚类的个数，n_j 是第 j 类中样本的个数，m_j 是第 j 类样本的均值，代表此类型数据集合的中心，即：

$$m_j=\frac{1}{n_j}\sum_{j=1}^{n_j}X_j,j=1,2,\cdots k$$

算法描述如下：

随机选取 k 个点作为初始聚类中心，然后根据各个样本到各聚类中心的距离把样本分到各类；重新计算每个类的中心（即类中所有点平均值，也就是几何中心），再次将各样本根据与聚类中心的距离归类，如此循环迭代，直到平方误差准则函数稳定在最小值。

如图 2 所示，当 $k=3$ 时，即需要将数据对象分为 3 个聚类，根据以上算法描述，任意选择 3 个对象作为 3 个初始聚类中心，聚类中心在图中用“+”来标注。根据与聚类中心的距离，每个对象被分配给最近的一个聚类，这样的分布形成了虚线所描绘的图形。



(a)任选 3 个对象作为初始聚类中心，得到的划分　(b) 第一次计算每类的中心，得到的划分　(c) 经过 n 次迭代，最终的划分

图 2　基于 K-means 算法聚类过程

2 顾客购物行为的向量化表示

由于一个顾客的购物行为可以用购买商品的种类来表示，为了便于进行聚类分析，为每个顾客建立一个 n 维向量[9]用来描述顾客的行为，把每个顾客的购买记录转变为向量，可以看做实现了从数据空间到向量空间的一种映射。比如：用 2 个向量 $X=(x_1,x_2,\cdots\cdots x_n)$、$Y=(y_1,y_2\cdots\cdots y_n)$ 代表顾客的购买行为，其中 X、Y 代表不同客户，x_n、y_n 代表每种商品的数量。若没有购买某种商品，便记其数量为 0。

3 顾客间的相似度

为了比较 2 个向量 $X=(x_1,x_2,\cdots\cdots x_n)$、$Y=(y_1,y_2\cdots\cdots y_n)$ 的相似度的大小[10]，定义了相似度函数 $sim(X,Y)$，用其来计算两个顾客购买商品的相似度，公式如下：

$$Sim(X,Y)=\frac{\sum_{j=1}^{n}(x_i,y_i)}{\sqrt{x_1^2+x_2^2+\cdots x_n^2}\times\sqrt{y_1^2+y_2^2+\cdots y_n^2}} \quad (1)$$

2 个向量的相似度表示了 2 个顾客的兴趣爱好的相似度，值越大，表示 2 个顾客兴趣度越相似，反之，表示 2 个顾客的兴趣度差别越大[11]。

4 算法分析

4.1 顾客群体的划分算法

算法的具体步骤如下：

输入：包含 n 个顾客行为的数据集。

输出：聚类数目 k 和 k 个聚类的集合。

(1)聚类数 k 的取值范围为[2，k_{max}]，步长可以变化、不固定，k_{max} 为聚类数目的最大限定[12]。

(2)从 n 个数据对象中任选 k 个对象作为初始的聚类中心。利用兴趣相似度式(1)计算出任意 2 个顾客之间的相似度。

(3)根据顾客之间的相似度，对数据集中的顾客进行分类，对于任意的顾客 $X\in n$，寻找与其相似度最大的类心 c_k，然后 X 属于第 k 类。

(4)当所有的数据集中的顾客都确定其聚类的归属后，计算每个聚类的新的类心（即类中所有点相似度的平均值，也就是几何中心）(式(2))，再次将各顾客依据相似度分类，直到误差准则函数(式(3))稳定在最小值。从而得到不同聚类。

(5)对聚类数目为 k 时的有效指数 $Validity(k)$(式(4))进行计算，选择 $Validity$ 值最大的 k 只保留下来。

(6)输出聚类数目 k 和 k 个聚类的集合。

平均相似度公式：

$$m_i=\frac{1}{n_i}\sum_{i=1}^{n_j}Sim(X,Y),i=1,2,\cdots k \quad (2)$$

误差准则函数形式：

$$J_c=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\|Sim(X,Y)-m_i\| \quad (3)$$

式中，k 为要形成聚类的个数，n_i 是第 i 类中样本的个数，m_i 是第 i 类样本的均值。

有效指数定义[13]：

$$Validity(k)=\frac{\sum_{i=1}^{k}Sim(x,c_i)}{\sum_{i\neq j}^{k}Sim(c_i,c_j)} \quad i,j=1,2\cdots k \quad (4)$$

式中，c_i 表示第 i 个聚类的中心。

4.2 商品的陈列算法

依据上面算法分成的 k 个顾客群体，在每类群体中，计算每种商品占商品总数的比例，依据比例的大小，由近到远对商品进行排列，从而得到商品的排列次序。

本文根据顾客的购买记录，根据其购买的商品间的相似性，划分出相似性大的顾客群体，再根据每个群体中的每种商品占商品总数的比例大小进行排序，从而得到商品排序的理论依据，进而使商品得到合理排序，这样顾客在浏览商品时，便会刺激其购买欲望，达到扩大销售的目的。但是每种商品，由于其品牌不同，知名度、

信誉度等也不同，并且商品陈列时还要考虑场地位置，颜色搭配等，从而为商品陈列带来新的问题，因此在为其提供基础的同时为下一步工作指明了方向。

参考文献

[1] 傅强.超市商品陈列对消费心理的影响[J].中国商贸，2010(3).

[2] 朱海红，江庭友，司丹丹，基于数据挖掘技术的商品陈列研究[J].商场现代化，2010(12).

[3] 王金龙，徐从富，徐娇芬，等.利用销售数据的商品影响关系挖掘研究[J].电子科技大学学报，2007(2).

[4] 崔春生，吴祈宗，王莹，用于推荐系统聚类分析的用户兴趣度研究[J].计算机工程与应用，2011(7).

[5] 刘金岭. 数据挖掘技术在商品销售预测方面的应用[J].商场现代化，2008(2).

[6] BERRY M, LINOFF G. Data mining techniques for marketing, sales, and customer relationship management [M]. 2nd ed. [S.l.]: John Wiley & Sons, Inc, 2004.

[7] 黄韬，刘胜辉，谭艳娜.基于k-means聚类算法的研究[J].计算机技术与发展，2011(7).

[8] 安建成，德增.一种改进的K-means算法[J].电脑开发与应用，2011(4).

[9] 韩瑞凯，孟嗣仪，刘云，等.基于兴趣相似度的社区结构发现算法研究[J].计算机应用，2010(10).

[10] Han Jiawei, KAMBER M.数据挖掘概念与技术[M].北京：机械工业出版社，2001.

[11] 王德荣，李卫华.网络号百用户兴趣模型挖掘算法[J].现代计算机，2010(4).

[12] 赵凤霞、福鼎，基于K-means聚类算法的复杂网络社团发现新算法[J].计算机应用研究，2009(6).

[13] 樊宁.K均值聚类算法在银行客户细分中的研究[J],.计算机仿真，2011(3).

（收稿日期：2011-11-07）

作者简介：

杨通辉，男，1986年生，硕士，主要研究方向：数据挖掘，复杂网络。

高玲，女，1965年生，副教授，主要研究方向：移动计算，智能控制，复杂网络。

臧丽，女，1986年生，硕士，主要研究方向：数据挖掘，复杂网络。