文章编号: 1001-0920(2018)06-1117-06 DOI: 10.13195/j.kzyjc.2017.0307

零售商品关联大数据稀疏网络的快速聚类算法

李桃迎†, 李 峰, 陈 燕, 吕晓宁

(大连海事大学 交通运输管理学院，辽宁 大连 116026)

摘 要: 关联规则方法被广泛应用于分析零售企业交易数据,以此指导品类管理、门店布局陈列和商品促销等运营决策,但面对电子商务网站非常巨大的数据量,仍存在效率低下的问题. 对此,提出商品关联大数据稀疏网络快速聚类算法. 首先,利用单步链表结构存储零售商品的共同购买关系矩阵;其次,对商品关联大数据稀疏网络的低度商品节点进行剪枝,降低搜索空间;再次,利用模糊 k 均值聚类对商品关联大数据稀疏网络进行快速聚类,并利用高连接度值商品节点被低连接度值商品节点分割的思想对剩余节点聚类;最后,将所提算法应用到亚马逊网站商品交易数据分析中,取得了良好的效果.

关键词: 聚类分析；大数据；关联规则；稀疏网络

中图分类号: C931.9 **文献标志码:** A

Fast clustering for sparse network of retail products associated big data

LI Tao-ying†, LI Feng, CHEN Yan, LYU Xiao-ning

(College of Transportation Management，Dalian Maritime University，Dalian 116026，China)

Abstract: The association rules method is widely used in the analysis of retail trading data so as to guide the operational decision-making for category management, store layout and the commodity sales promotion of products. However, the data of the electronic commerce website is very huge, which leads to the low inefficiency of the associate rule. The fast clustering for sparse network of retail products associated big data is proposed. Firstly, the structure of the one step linked list is used to store the co-purchasing matrix. Then, the nodes with the low degree in the sparse network of retail products associated big data are pruned, which reduces the search space. Furthermore, the nodes of the sparse network of retail products associated big data are grouped by fuzzy k clustering with the idea that the nodes with high connectivity value are partitioned by the nodes with low connectivity value. Finally, the proposed algorithm is applied to analyze the trading data of the Amazon website, and the results show the effectiveness of the proposed method.

Keywords: clustering analysis；big data；associate rule；sparse network

0 引 言

随着信息系统和信息化建设的全面应用,零售企业在运作管理中产生大量的数据记录,挖掘这些数据之间的经营视点和市场规律,具有相当大的应用价值. 通过研究零售企业的购物交易数据,挖掘商品之间存在的关联关系和频繁项目集,可以指导零售企业作出更加科学合理的品类管理、门店布局陈列和商品促销等运营决策,大幅度提高企业的经营业绩[1],啤酒与尿布[2] 的故事就是其中一个经典案例. 随着电子商务的快速发展,亚马逊、京东商城、一号店等电子商务零售企业基于商品关联性分析,在平台上提供高品质的商品推荐、关联促销和关联导购等营销服务,为顾客带来了便利的购物体验,同时极大地提高了客单价和商品销量[3].

人们对零售企业购物交易数据的分析多采用关联规则方法[4]. 张志宏等[1] 利用关联规则的特点,建立了一个多目标优化模型,将商品直接受益和由于交叉销售因素产生的直接利润作为两个独立的优化目标,并设计了多目标遗传算法进行求解;琚春华等[5] 应用关联规则挖掘商品之间存在的关联关系和频繁项目集;杨丰梅等[3] 应用邻接矩阵和截矩阵技术,提出了挖掘商品关联性的算法和求频繁项集的算法;

收稿日期: 2017-03-22；**修回日期:** 2017-06-22.

基金项目: 国家社会科学基金项目 (15CGL031)；国家自然科学基金项目 (71271034)；大连市高层次人才创新支持计划项目 (2015R063)；中央高校基础科研业务费专项基金项目 (3132017085, 3132016306).

责任编委: 唐万生.

作者简介: 李桃迎 (1983−), 女, 副教授, 博士, 从事方向数据挖掘、复杂网络的研究；陈燕 (1952−), 女, 教授, 博士生导师, 从事方向数据挖掘、多维信息组织与管理等研究.

†通讯作者. E-mail: ytaoli@126.com

彭敦陆等[6]以分析用户行为为基础,通过分析电子商务平台上结构化的数据,计算实体之间的内在关联,并利用这一关联找到一些有着紧密相关、但共现次数不多的商品推荐给用户;Wong等[7-8]将交叉销售因素结合到商品选择中,建立了MIPS(Maximal-profit item selection)模型,提出了损失规则的概念,利用关联规则的置信度来刻画目标商品与非目标商品之间的关联关系,并给出了相应的优化求解算法. 国内外现有的文献中也有很多针对商品个性化推荐的研究,如唐晓波等[9]提出了3种客户聚类方法解决协同过滤推荐的“稀疏性”问题,并结合相关性产品和协同过滤进行推荐;Bach等[10]针对社交网络媒体中的评论数据进行个性化推荐,引入一种新的混合推荐方法,结合评论的合作功能和内容给出排序;朱国玮等[11]基于非线性逐步遗忘函数建立用户兴趣模型,预测用户未评价商品评分,引入“领域最近邻”处理方法查找目标用户的最近邻,预测未评价商品评分,以此为基础作出推荐. 李永立等[12]提出了基于图论的推荐方法,将人和物的相似性信息结合起来,构成综合的评估图模型,并转化为与之等价的评估矩阵,在最大化保留评估信息的优化目标下,以评估矩阵为基础建立推荐算法;胡润波等[13]结合现有移动商务感知信任研究,从移动技术、Wap网站、移动商家和商品、制度环境5个方面多角度地分析移动商务感知信任影响因素,并以此建立了移动商家信任度评价指标体系. 以上这些方法对零售商品交易大数据进行分析时的效果都不理想. 关联规则方法需要多次访问购物交易数据,且由于零售商品交易数据量特别巨大(可称为大数据),导致分析挖掘效率特别低下.

商品销售记录挖掘的主要目的是作商品推荐或商品绑定销售使用,近年也有人提出采用个性化推荐的常用挖掘方法——复杂网络方法来挖掘零售商品之间的关系,但采用复杂网络方法对购物交易数据进行挖掘时,需要在建立并存储所有商品的同时购买关联图或同时购买关联矩阵(Co-purchasing matrix). 因为零售商品种类特别多,相对所有商品而言,任何消费者同时购买商品种类的数量远远小于所有商品种类的数量,所以该矩阵为高稀疏矩阵. 目前广泛使用的邻接矩阵和邻接链表存储方式用于分析挖掘潜在规律时需要多次访问数据,所以不适用于零售商品的大数据关联分析. 面向零售商品关联大数据稀疏网络的聚类需要以提高效率为核心,常用的复杂网络聚类需要预先给出有效的目标函数和最优解搜索策略,搜索过程牺牲效率来提高聚类的精度,不适用于对零售商品关联大数据稀疏网络进行聚类.

虽然大型零售超市、电子商务网站的商品种类众多,但多数消费者一次购买不超过20种,所以商品之间的共同购买关系网络为稀疏网络——网络中节点的平均度不远大于节点总数的自然对数,用关联规则、复杂网络进行挖掘时,效率非常低下,因此急需一种能够处理高稀疏数据且快速挖掘商品销售记录中潜在规律的算法. 受文献[14]的启发,本文提出商品关联大数据稀疏网络快速聚类算法,同样采用基于密度的快速聚类算法,即基于这样的假设:类簇中心被具有较低局部密度的邻居点包围,且与具有更高密度的任何点有相对较大的距离. 本文算法首先利用单步链表结构存储零售商品的共同购买关系矩阵;然后对商品关联大数据稀疏网络的低度商品节点进行剪枝,降低搜索空间;最后利用模糊k均值聚类对商品关联大数据稀疏网络的快速聚类,并借用高连接度值商品节点被低连接度值商品节点分割的思想对剪枝节点聚类.

1 商品关联大数据稀疏网络快速聚类算法

由于仅购买1种商品的记录对研究没有意义,本文只考虑同时购买多种商品的情况.

1.1 构建商品关联大数据稀疏网络

构建商品关联大数据稀疏网络的过程包括如下3步.

1) 令V为商品节点集合,商品节点数目为N,则$V=\{v_1,v_2,\cdots,v_i,\cdots,v_N\}$,$v_i$表示第$i$种商品节点(商品节点为某一具体商品,如康师傅绿茶).

2) 依据商品购买记录,建立商品之间共同购买关系的商品关联大数据稀疏网络$G=\langle V,E\rangle$,所有商品间共同购买关系作为边,如商品节点v_i与v_j间的共同购买关系作为边e_{ij},边的集合为$E=\{e_{ij}\}$,同时所有商品共同购买关系的矩阵表达形式为$(e_{ij})_{N*N}$,v_j表示第j种商品节点,$j=1,2,\cdots,N$.

3) 存储商品节点v_i为一个单步链表,单步链表包含3部分:商品节点v_i的序号、商品节点v_i的连接度d_i、指向所有与该商品节点v_i相邻的所有商品节点序号的集合Ω_i. 商品节点v_i的连接度d_i表示与商品节点v_i存在共同购买关系的商品数目,即商品购买关系网络中与商品节点v_i相邻的商品节点数. 下面以具体例子来说明算法的实施过程.

例1 假设商品有8种,销售记录如表1所示. 其中:商品序号1~8分别代表康师傅麻辣牛肉面、散士力架、农夫山泉饮用天然水、康师傅绿茶、洽洽香瓜子、环保购物袋、蒙牛红枣酸牛奶和散东北珍珠米;

表格中的“√”表示购买了该商品.

表 1　8种商品的销售记录

	商品序号							
	1	2	3	4	5	6	7	8
小票1	√		√					
小票2		√	√	√				
小票3		√		√	√			
小票4	√					√		
小票5						√	√	
小票6						√		√

图1为依据销售记录构建的8种商品共同购买关系网络图,该网络为稀疏网络(节点的平均度2.25接近节点总数的自然对数2.08).

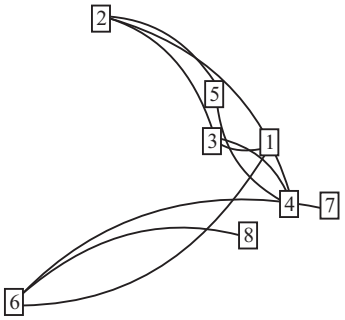
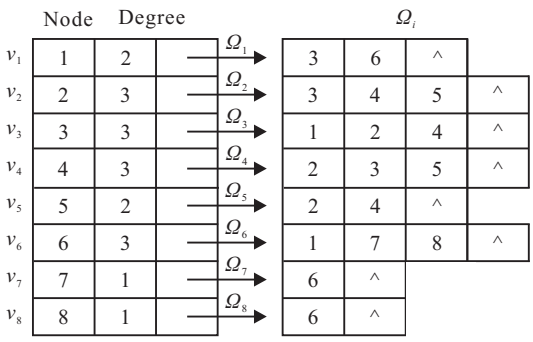
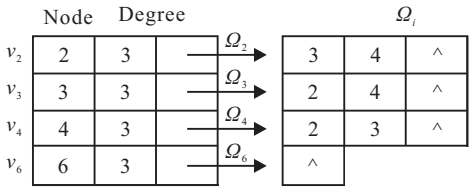


图 1　商品共同购买关系网络图

建立网络的链表存储结构,如图2(a)所示.



(a) 初始链表存储结构



(b) 剪枝后的链表存储结构

图 2　网络的链表存储结构

图2(a)中,每个商品节点作为一个链表,先存储v_i的序号,再存储v_i的度d_i (与v_i存在同时购买关系的商品数目,即商品购买关系网络中与节点v_i相邻的节点数),之后指向所有与该节点相邻的节点.

获取Ω_i的步骤如下:

Step 1: 设定当前商品节点序号$j=1,\Omega_i=\varnothing$,$\varnothing$为空集.

Step 2: 若$j\leqslant N$,则转到Step 3,否则转到Step 5.

Step 3: 若商品节点v_i与v_j相邻,即存在共同购买关系,则将商品节点v_j存入Ω_i,否则直接转入Step 4.

Step 4: 令当前商品节点序号$j=j+1$,返回Step 2.

Step 5: 停止,得到Ω_i.

1.2　剪枝操作

对商品关联大数据稀疏网络进行剪枝的具体步骤如下.

Step 1: 令连接度的阈值为d_c.

Step 2: 设定当前商品节点序号$i=1$.

Step 3: 如果商品节点v_i的连接度$d_i\geqslant d_c$,则转到Step 4,否则转到Step 5.

Step 4: 将商品节点v_i放入集合Ψ,Ψ为连接度大于等于d_c的商品节点集合.

Step 5: 当前商品节点序号$i=i+1$.

Step 6: 若$i\leqslant N$,则转到Step 3,否则转到Step 7.

Step 7: 停止剪枝,得到剪枝后的商品节点集合Ψ.

Step 8: 设定当前商品节点序号$i=1$.

Step 9: 如果$v_i\in\Psi,\exists v_j\notin\Psi$且$v_j\in\Omega_i$,则从$\Omega_i$中删除$v_j,j=1,2,\cdots,N$;否则转到Step 10.

Step 10: 求新的商品节点序号$i=i+1$.

Step 11: 若$i\leqslant N$,则转到Step 9,否则转到Step 12.

Step 12: 得到剪枝后的商品关联大数据稀疏网络$G'=\langle V',E'\rangle$,V'为剪枝后的商品节点集合,E'为剪枝后边的集合;设置连接度的阈值$d_c=3$,例1中剪枝后G'的存储结构如图2(b)所示.

1.3　基于模糊k均值的快速聚类

本文对所有剪枝后节点的Ω_i集合,运用模糊k均值算法进行快速聚类,采用的成本函数如下所示:

$$F(T,W,C)=\sum_{l=1}^{k}\frac{\sum_{j=1}^{n}\sum_{i=1}^{m}\tau_{lj}\omega_i(c_{li}-x_{ji})^2}{\sum_{i=1}^{m}(c_{li}-\bar{x}_i)^2}+\gamma\sum_{i=1}^{m}\omega_i\log\omega_i. \quad (1)$$

其中

$$\sum_{l=1}^{k}\tau_{lj}=1,1\leqslant j\leqslant n,\tau_{lj}\in\{0,1\};$$

$$\sum_{i=1}^{m}\omega_i=1,0\leqslant\omega_i\leqslant 1;$$

k、n和m分别是簇、对象和属性的数目;x_{ji}是第j个对象第i个属性的值;$C=[c_{li}]$是$k\times m$矩阵,c_{li}是第l个簇中心的第i个属性的值;$T=[\tau_{lj}]$是$k\times n$的矩

阵,τ_{lj}是第j个对象属于第l个簇的隶属度;$W=[\omega_i]$是m维向量,ω_i是第i个属性的权值;γ是大于1的参数;$\bar{x}$是所有对象的均值,$\bar{x}_i$是$\bar{x}$的第i个属性的值,即$\bar{x}_i=\dfrac{1}{n}\sum\limits_{j=1}^{n}x_{ji}$,如果$n>1$,则成本函数有效,否则$\sum\limits_{i=1}^{m}(c_{li}-\bar{x}_i)^2=0$,导致$F(T,W,C)$无法获得.其中式(1)中第一项的分母为各类的均值与所有类均值之差的平方和.

F的最小化包含一系列未知解决方法的条件非线性优化问题.一般而言,为了求取T、W和C的值,需要将优化问题转换成部分优化问题,文献[15-17]中的方法可以用作参考.首先固定T和C,寻找使$F(T,W,C)$最小的W;然后固定T和W,寻找最适当的C;最后固定C和W,寻找合适的T.重复上述步骤直到目标函数不再减少.

定理1 在T和C的值保持不变时,当且仅当下式成立时,F是最小值:

$$\omega_t=\frac{\exp\Big(\dfrac{-\psi_t}{\gamma}-1\Big)}{\sum\limits_{i=1}^{m}\exp\Big(\dfrac{-\psi_i}{\gamma}\Big)},\tag{2}$$

其中

$$\psi_t=\sum_{l=1}^{k}\frac{\sum\limits_{j=1}^{n}\tau_{lj}(c_{lt}-x_{jt})^2}{\sum\limits_{i=1}^{m}(c_{li}-\bar{x}_i)^2}.\tag{3}$$

证明 采用文献[15-17]中的方法可以获得无条件的最小值优化问题,如下所示:

$$\begin{aligned}\min\ \ & F(\{\omega_i\},\xi)=\\ &\sum_{l=1}^{k}\frac{\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{m}\tau_{lj}\omega_{li}(c_{li}-x_{ji})^2}{\sum\limits_{i=1}^{m}(c_{li}-\bar{x}_i)^2}+\\ &\gamma\sum_{i=1}^{m}\omega_i\log\omega_i-\xi\Big(\sum_{i=1}^{m}\omega_{li}-1\Big).\end{aligned}\tag{4}$$

因为各簇之间是相互独立的,所以

$$\begin{aligned}F(\omega_i,\xi)=&\sum_{l=1}^{k}\frac{\sum\limits_{j=1}^{n}\sum\limits_{i=1}^{m}\tau_{lj}\omega_i(c_{li}-x_{ji})^2}{\sum\limits_{i=1}^{m}(c_{li}-\bar{x}_i)^2}+\\ &\gamma\sum_{i=1}^{m}\omega_i\log\omega_i-\xi\Big(\sum_{i=1}^{m}\omega_i-1\Big).\end{aligned}\tag{5}$$

$F(\omega_i,\xi)$是可导的,设置导数为零,即

$$\frac{\partial F(\omega_i,\xi)}{\partial\xi}=\sum_{i=1}^{m}\omega_i-1=0,\tag{6}$$

且

$$\begin{aligned}\frac{\partial F(\omega_t,\xi)}{\partial\omega_t}=&\sum_{l=1}^{k}\frac{\sum\limits_{j=1}^{n}\tau_{lj}(c_{lt}-x_{jt})^2}{\sum\limits_{i=1}^{m}(c_{li}-\bar{x}_i)^2}+\\ &\gamma(1+\log\omega_{lt})-\xi=0.\end{aligned}\tag{7}$$

由式(7)可以得到

$$\omega_t=\exp\Big(\frac{-\psi_t+\xi-\gamma}{\gamma}\Big),\tag{8}$$

其中ψ_t如式(5).

将式(8)代入(6)可得

$$\exp\Big(\frac{\xi-\gamma}{\gamma}\Big)=\frac{1}{\sum\limits_{i=1}^{m}\exp\Big(\dfrac{-\psi_i}{\gamma}\Big)}.\tag{9}$$

将式(9)代入(8)可得(2). □

同理,固定W和C可以求取T的值.众所周知,如果第j个对象到第l个簇的距离最小,则它将属于第l个簇,即

$$\tau_{lj}=\begin{cases}1,\ \sum\limits_{i=1}^{m}\omega_{li}(c_{li}-x_{ji})^2\leqslant\sum\limits_{i=1}^{m}\omega_{zi}(c_{zi}-x_{ji})^2;\\ 0,\ \text{Otherwise}.\end{cases}\tag{10}$$

$\tau_{lj}=1$表示第j个对象完全属于第l个簇,否则表示不属于第l个簇.

固定T和W,采用数学平均值的方法获取C的值,有

$$c_{li}=\sum_{j=1}^{n}\tau_{lj}x_{ji}\Big/\sum_{j=1}^{n}\tau_{lj}.\tag{11}$$

其中:$1\leqslant l\leqslant k,1\leqslant i\leqslant m$.

聚类过程结束,输出各类中的商品节点,例1中最终得到2个类,分别为$C_1=\{v_2,v_3,v_4\},C_2=\{v_6\}$.针对所有被剪枝的节点,即未分类的节点,将其划分到与各类中节点连接边的数目最多的类,当存在与两个或多个类的连接边数相同时,制定统一的策略,如属于节点数较少或下标较小的类等.针对例1中的8种商品,最终得到的聚类结果为$C_1=\{v_2,v_3,v_4,v_5\},C_2=\{v_1,v_6,v_7,v_8\}$,具体如图3所示.图3中,商品2(散士力架)、商品3(农夫山泉饮用天然水)、商品4(康师傅绿茶)和商品5(洽洽香瓜子)归为一类,所以当零售商在进行货品摆放时,应当考虑到消费者可能会同时购买这些商品,可以将这些商品摆放在零售商店(或电子商务零售网站)相近位置,

方便顾客选购,以此来增加商品销量. C_2 中的商品同样可以摆放在相近的位置.

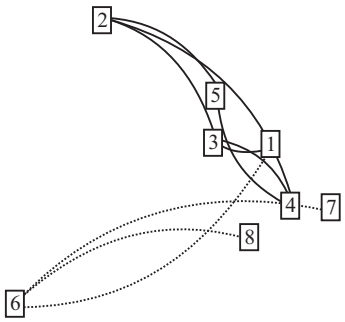


图 3　8个商品的聚类结果

2　实验分析

2.1　亚马逊200个商品的关联网络快速聚类

首先,依据亚马逊200个商品的销售记录数据,构建共同购买关系网络,并建立该网络的链表存储结构;其次,对商品关联大数据稀疏网络进行剪枝,剪枝过程中选择连接度的阈值为 $d_c = 5$,剪枝后网络的节点有56个,剪枝后的节点集合

$$V' = \{v_0, v_1, v_2, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12}, v_{13}, v_{14}, v_{19}, v_{22}, v_{23}, v_{24}, v_{26}, v_{27}, v_{32}, v_{34}, v_{35}, v_{36}, v_{38}, v_{40}, v_{41}, v_{43}, v_{44}, v_{46}, v_{47}, v_{48}, v_{49}, v_{53}, v_{56}, v_{65}, v_{68}, v_{71}, v_{74}, v_{86}, v_{87}, v_{88}, v_{92}, v_{96}, v_{97}, v_{105}, v_{112}, v_{121}, v_{122}, v_{123}, v_{126}, v_{127}, v_{128}, v_{129}, v_{177}, v_{190}\};$$

再次,对商品关联大数据稀疏网络进行快速聚类,得到各类的高连接度节点集合,集合的数目即为类的数目,针对 V' 最终得到3类商品节点集合,即

$$C_1 = \{v_0, v_1, v_2, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}, v_{32}, v_{34}, v_{35}, v_{36}, v_{38}, v_{40}, v_{41}, v_{43}, v_{44}, v_{46}, v_{47}, v_{48}, v_{49}, v_{53}, v_{56}, v_{86}, v_{87}, v_{88}, v_{92}, v_{96}, v_{97}, v_{105}, v_{112}, v_{177}\},$$

$$C_2 = \{v_{11}, v_{12}, v_{13}, v_{14}, v_{19}, v_{22}, v_{23}, v_{24}, v_{26}, v_{27}, v_{65}, v_{68}, v_{71}, v_{74}, v_{126}, v_{127}, v_{128}, v_{129}\},$$

$$C_3 = \{v_{121}, v_{122}, v_{123}, v_{190}\};$$

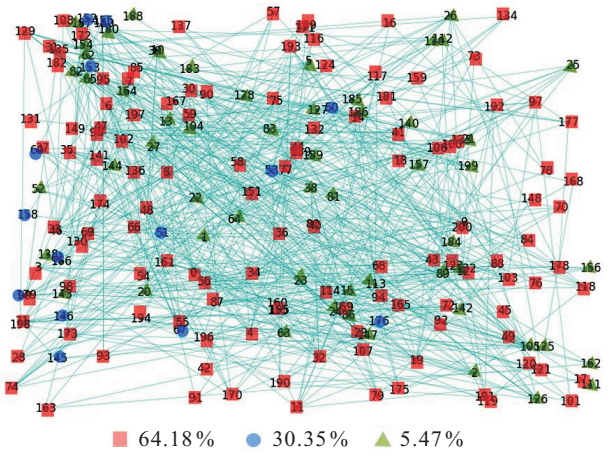


图 4　亚马逊200个商品的聚类结果

最后,对低连接度节点进行划分,得到最终的聚类结果,如图4所示,其含义与图3相似.

2.2　亚马逊500个商品的关联网络快速聚类

首先,依据亚马逊500个商品的销售记录数据,构建共同购买关系网络,并建立该网络的链表存储结构;其次,对商品关联大数据稀疏网络进行剪枝,剪枝过程中选择连接度的阈值为 $d_c = 10$,剪枝后网络的节点有71个,剪枝后的节点集合为

$$V' = \{v_0, v_2, v_4, v_5, v_6, v_{12}, v_{13}, v_{14}, v_{19}, v_{20}, v_{22}, v_{24}, v_{26}, v_{27}, v_{29}, v_{30}, v_{31}, v_{34}, v_{35}, v_{38}, v_{40}, v_{41}, v_{43}, v_{44}, v_{46}, v_{47}, v_{71}, v_{82}, v_{86}, v_{89}, v_{96}, v_{103}, v_{105}, v_{106}, v_{121}, v_{122}, v_{123}, v_{126}, v_{130}, v_{151}, v_{155}, v_{193}, v_{198}, v_{209}, v_{211}, v_{212}, v_{214}, v_{241}, v_{250}, v_{257}, v_{258}, v_{264}, v_{281}, v_{324}, v_{326}, v_{329}, v_{336}, v_{338}, v_{347}, v_{353}, v_{374}, v_{375}, v_{377}, v_{378}, v_{395}, v_{406}, v_{409}, v_{414}, v_{432}, v_{436}, v_{443}\};$$

再次,对商品关联大数据稀疏网络进行快速聚类,得到各类的高连接度节点集合,集合的数目即为类的数目,针对 V' 最终得到13类商品节点集合,即

$$C_1 = \{v_0, v_2, v_4, v_5, v_6, v_{34}, v_{35}, v_{38}, v_{40}, v_{41}, v_{43}, v_{44}, v_{46}, v_{47}, v_{82}, v_{86}, v_{89}, v_{96}, v_{103}, v_{105}, v_{106}, v_{151}, v_{155}\},$$

$$C_2 = \{v_{12}, v_{13}, v_{14}, v_{19}, v_{20}, v_{22}, v_{24}, v_{26}, v_{27}, v_{29}, v_{30}, v_{31}, v_{71}, v_{209}, v_{250}, v_{281}, v_{324}, v_{326}, v_{329}, v_{347}, v_{353}, v_{374}, v_{375}, v_{377}, v_{378}\},$$

$$C_3 = \{v_{121}, v_{123}, v_{193}\}, C_4 = \{v_{126}, v_{130}, v_{198}\},$$

$$C_5 = \{v_{211}, v_{212}, v_{257}, v_{258}, v_{264}\}, C_6 = \{v_{214}\},$$

$$C_7 = \{v_{241}\}, C_8 = \{v_{336}\}, C_9 = \{v_{338}, v_{395}, v_{436}\},$$

$$C_{10} = \{v_{406}, v_{443}\}, C_{11} = \{v_{409}\},$$

$$C_{12} = \{v_{414}\}, C_{13} = \{v_{432}\};$$

最后,对低连接度节点进行划分,得到最终的聚类结果,如图5所示,其含义与图3相似.

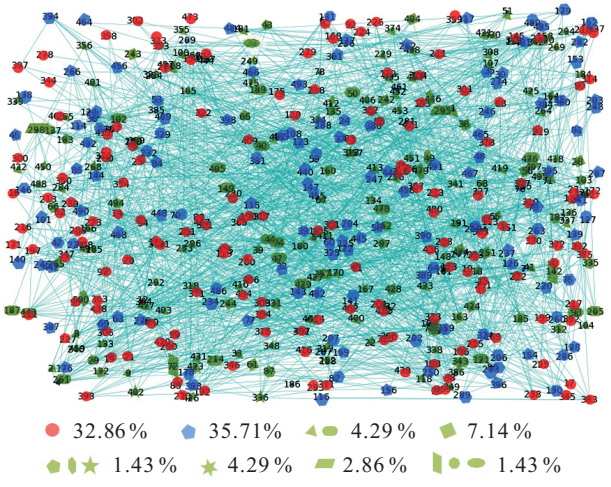


图 5　亚马逊500个商品的聚类结果

3 结 论

针对关联规则算法处理商品交易大数据时存在效率低下的问题,提出商品关联大数据稀疏网络快速聚类算法,有利于帮助零售企业或电子商务网站在海量交易数据中快速获得商品的关联. 首先,利用单步链表结构存储零售商品的共同购买关系矩阵;其次,对商品关联大数据稀疏网络的低度商品节点进行剪枝,降低搜索空间;再次,利用高连接度值商品节点被低连接度值商品节点分割的思想对商品关联大数据稀疏网络快速聚类;最后,利用该算法对亚马逊的销售记录进行分析挖掘,结果对于指导零售企业作出更加科学合理的品类管理、门店布局陈列、商品促销和大幅度提高企业的经营业绩具有重要的现实意义. 在电子商务迅速发展的今天,本文方法对于线上电子商务零售企业如亚马逊、京东商城等,在提供高品质的商品推荐、关联促销和关联导购等方面也具有很高的应用价值,有助于提高客单价和商品销量.

参考文献(References)

[1] 张志宏, 寇纪淞, 陈富赞, 等. 基于关联分析的多目标商品组合选择方法[J]. 系统工程学报, 2011, 26(1): 132-138.
(Zhang Z H, Kou J S, Chen F Z, et al. Multi-objective products selection method based on association analysis[J]. J of Systems Engineering, 2011, 26(1): 132-138.)

[2] Tsur D, Ullman J D, Abiteboul S, et al. Query flocks: A generalization of association-rule mining[C]. Proc of ACM SIGMOD Int Conf on Management of Data. Seattle: ACM Press, 1998: 1-12.

[3] 杨丰梅, 李梦, 田歆, 等. 一种带记忆性的零售商品关联度分析方法[J]. 系统工程理论与实践, 2014, 34(11): 2872-2880.
(Yang F M, Li M, Tian X, et al. An approach for retail goods association rules analysis with memory property[J]. Systems Engineering — Theory & Practice, 2014, 34(11): 2872-2880.)

[4] Brijis T. Building an association rules framework to improve product assortment decisions[J]. Data Mining and Knowledge Discovery, 2004, 8(1): 7-23.

[5] 琚春华, 殷贤君.基于兴趣度的数据流频繁模式散列挖掘算法[J].系统工程理论与实践, 2012, 32(12): 2764-2773.
(Ju C H, Yin X J. Mining approximate frequency itemsets over data streams based on hash and interesting degree[J]. Systems Engineering — Theory & Practice, 2012, 32(12): 2764-2773.)

[6] 彭敦陆, 张书录. 考虑实体关联的商品推荐技术[J]. 小型微型计算机系统, 2016, 37(3): 464-468.
(Peng D L, Zhang S L. Entity-relationship aware approach for recommending relevant commodities[J]. J of Chinese Computer Systems, 2016, 37(3): 464-468.

[7] Wong R C, Fu A W, Wang K. MPIS: Maximal-profit item selection with cross-selling considerations[C]. Proc of the 3rd IEEE Int Conf on Data Mining. Melbourne: IEEE Computer Society, 2003: 371-378.

[8] Wong R C, Fu A W, Wang K. Data mining for inventory item selection with cross-selling considerations[J]. Data Mining and Knowledge Discovery, 2005, 11(1): 81-112.

[9] 唐晓波, 樊静. 基于客户聚类的商品推荐[J]. 情报杂志, 2009, 28(6): 143-146.
(Tang X B, Fan J. Recommendation based on customer clustering[J]. J of Intelligence, 2009, 28(6): 143-146.)

[10] Bach N X, Hai N D, Phuong T M. Personalized recommendation of stories for commenting in forum-based social media[J]. Information Sciences, 2016, 352/353: 48-60.

[11] 朱国玮, 周利. 基于遗忘函数和领域最近邻的混合推荐研究[J]. 管理科学学报, 2012, 15(5): 55-64.
(Zhu G W, Zhou L. Hybrid recommendation based on forgetting curve and domain nearest neighbor[J]. J of Management Sciences in China, 2012, 15(5): 55-64.)

[12] 李永立, 吴冲, 王崖声. 基于图论和信息最大化保留的在线推荐方法[J]. 系统工程理论与实践, 2011, 31(9): 1718-1725.
(Li Y L, Wu C, Wang Y S. On-line recommendation method based on graph model and maximizing information retention[J]. Systems Engineering — Theory & Practice, 2011, 31(9): 1718-1725.)

[13] 胡润波, 杨德礼, 祁瑞华. 移动商务中基于综合评价的推荐信任评估模型[J]. 运筹与管理, 2010, 19(3): 85-93.
(Hu R B, Yang D L, Qi R H. Recommended trust evaluation model in mobile commerce based on combination evaluation model[J]. Operations Research and Management Science, 2010, 19(3): 85-93.)

[14] Alex Rodriguez, Alessandro Laio. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[15] 李桃迎, 陈燕, 张金松, 等. 基于聚类融合的混合属性数据增量聚类算法[J]. 控制与决策, 2012, 27(4): 603-608.
(Li T Y, Chen Y, Zhang J S, et al. Incremental clustering algorithm of mixed numerical and categorical data based on clustering ensemble[J]. Control and Decision, 2012, 27(4): 603-608.)

[16] Jing L, Ng M K, Huang J Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data[J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

[17] Chan Y, Ching W, Ng M K, et al. An optimization algorithm for clustering using weighted dissimilarity measures[J]. Pattern Recognition, 2004, 37(5): 943-952.

(责任编辑：齐 霁)