|  |
| --- |
| 江 西 理 工 大 学  本 科 毕 业 设 计（论文）  题 目：基于商品影响力分析的购物篮方法研究  学 院：信息工程学院  专 业：计算机科学与技术(计算机软件)  班 级：091班  学 生：欧阳显斌  学 号：20092183  指导教师：杨书新 职称：副教授 |

摘要

随着经济的发展，越来越多的人会去中大型超市购物，超市的数量也在增加，也意味着超市竞争压力的增加。在超市的运营过程，会积累大量的购物篮数据。为了提高顾客的满意度和超市利润，零售商非常关注如何利用购物篮数据来刻画用户的购物行为。目前对购物篮分析的一种常用方法是关联规则分析，然而关联规则得到的是用户购买一些商品再购买另外一些商品的可能性。国外有很少的文献研究利用商品网络来分析购物篮数据，并且这些文献是直接对商品网络进行表面分析，度数和边的分布等。当前仍没找到利用社交网络的影响力分析的方法对商品网络进行分析的应用研究，所以本文会使用经典的影响力分析方法来进行这个探究性实验。

该文将利用两种经典的影响力最大化方法对商品网络进行分析，利用基于热量传播模型和快速层次聚类的影响力最大化算法和基于独立级联模型和层次快速层次聚类的影响力最大化算法对商品网络进行分析，并对找到的影响力最大化的商品的特征进行分析。具体所做的工作如下：

（1）关联规则分析的一种常用方法是apriori算法。本文利用apriori算法，求取商品之间的关联性。以商品作为节点，以关联性作为有向边，那么就可以得到一个有向图。

（2）这一步是为求取影响力最大的商品做铺垫，为了避免所有的影响力最大化商品都集中在一个小社团中，并且在这里我们具有较强关联性的商品会在同一个社团中，商品影响力只会在所属的社团中传播，在不同的社团中商品不会相互影响。本文将无向图的快速层次聚类扩展到有向图中，并设计算法的伪代码，并通过实验分析社团的一些特点。

（3）热量传播和影响力的传播具有一定的相似性，且热量传播模型中时间参数可以在一定程度上模仿促销的临时性。所以本文将设计基于热量传播模型的影响力最大化方法的算法，并从商品成团中选出影响力前K大的商品。

（4）独立级联模型是社交网络中常用的影响力传播模型之一，该模型简单直观，在影响力最大化中表现出良好的特性，如子模特性，方便扩展模型和进一步分析。所以本文将设计基于独立级联模型的影响力最大化方法的算法，并在商品网络的多个社团中寻找影响力前K大的商品。

（5）实验均使用一个大超市在1月份的交易记录作为实验数据，实验的算法实现用的是python语言，结果分析中图表绘制用python，网络绘制使用gephi软件。整个实验包括数据筛选，关联规则计算，社团划分和对上述两种影响力最大化算法的得到的结果进行对比和分析。

**关键字：**购物篮；关联规则；商品网络；社团划分；影响力最大化

ABSTRACT

最后在补充

目录

[基于商品影响力分析的购物篮方法研究 1](#_Toc449712888)

[摘要 1](#_Toc449712889)

[第一章 绪论 1](#_Toc449712890)

[1.1 课题来源及研究背景： 1](#_Toc449712891)

[1.2 国内外研究现状 2](#_Toc449712892)

[1.3 研究内容、意义、思路及架构 3](#_Toc449712893)

[1.4 本文主要创新点 4](#_Toc449712894)

[第二章 本文的相关理论 5](#_Toc449712895)

[2.1网络的定义 5](#_Toc449712896)

[2.2 网络的性质 5](#_Toc449712897)

[2.3 社交网络影响力传播模型 7](#_Toc449712898)

[2.4 影响力最大化问题定义 7](#_Toc449712899)

[2.5 关联规则算法 7](#_Toc449712900)

[第三章 基于购物篮的商品网络分析 8](#_Toc449712901)

[3.1 数据筛选 8](#_Toc449712902)

[3.2 构建商品网络 8](#_Toc449712903)

[3.3 关联规则算法设计： 9](#_Toc449712904)

[3.4 商品网络的特点 10](#_Toc449712905)

[第四章 基于有向图快速层次聚类的商品社团划分 10](#_Toc449712906)

[4.1 社区发现 11](#_Toc449712907)

[4.2 层次聚类算法 11](#_Toc449712908)

[4.3 商品网络社团划分 14](#_Toc449712909)

[第五章 基于社交网络影响力最大化的购物篮方法研究 14](#_Toc449712910)

[5.1 影响力传播模型 14](#_Toc449712911)

[5.2 影响力最大化问题的定义 16](#_Toc449712912)

[5.3 商品网络K个影响力最大化商品的发现 17](#_Toc449712913)

# 基于商品影响力分析的购物篮方法研究

## 绪论

随着经济水平的提高，人们在闲暇时间逛逛超市，不管是商品的购买量和光顾超市的频率都在快速增加，人们的需求带动了超市的发展。我们很容易想到这样的一个场景，超市进行着促销活动，在pos机前人们手中提着购物篮排着上上的队伍，收银员手忙脚乱的录入着顾客购物篮中商品的信息。在超市经营中积累的大量的这些数据叫做购物篮数据。

在居民区总能方便的找到超市，这也意味着人们可选择的超市在增多，超市的竞争压力也在增加。因此零售商们想着各种方法来提高顾客的购物体验，其中一种常用的方法就是通过分析购物篮数据来刻画顾客的购物行为，然后零售商再根据顾客的特点来制定策略改善顾客体验。

购物篮分析的英文叫做Market Basket Analysis，简称为MBA。商品关联性分析是MBA的一个重要的组成部分，其中一个经典的故事是啤酒和尿布的故事。这个故事讲述的是美国年轻父亲在周五上下班期间会在沃尔玛超市中同时购买尿布和啤酒这两样很难联系到一起的商品。这个现像在20世纪90年代被美国沃尔玛超市所发现，超市通过调整尿布和啤酒的布局带来了尿布和啤酒销量的提升。

针对购物篮分析的研究中针对商品分析的

### 1.1 研究背景与意义：

超市在线上和线下的经营过程中，用户的购买信息在付款的同时会被记录下来。大量的购买数据能够反映用户的行为和喜好，零售商对顾客信息分析的越透彻，也就能制定更好的经营策略。

购物篮分析可以帮助零售商解决很多的问题，比如用户的喜欢购买哪些商品，在购买一些商品后想去购买哪些商品等，商品在不同时间的需求量等。

关联规则分析是解决用户购买一些商品后想去哪些商品的问题，一种经典的算法是Apriori算法，这个算法能够得购买商品组合的频率以及商品之间的直接关联关系，然而这个算法并不能从整个网络分析，挖掘更多有价值的信息。

本研究利用的apriori的条件概率思想，构建起一个商品网络，然后分析商品的社团结构，以及研究社交网络的传播模型在商品网络的适用性，并从中找出能够让商品购买量最大的前K个商品。

### 1.2 国内外研究现状

#### 1.2.1超市购物分析问题的研究现状

购物篮分析（MBA）是一种广泛使用的技术，可以用来分析顾客同时购买的某些商品或类别的可能性。1992年，Julander首次提出购物篮分析[[1]](#endnote-1)方法，该方法被瑞典的一个超市所使用。1993年Agrawal等人提出apriori算法[[2]](#endnote-2)从交易记录中发现关联规则的方法，这个技术被很多论文所引用。这个技术典型的应用是评估零售店的商品关联性。基于数据挖掘的思想，购物篮分析的目的是通过从pos机上的交易记录发现顾客购买商品的关联性（Chen等人 [[3]](#endnote-3)2005；Berry和Linoff等人[[4]](#endnote-4) 2004）。或许最著名的关联性规则是啤酒和尿布，因为家庭在周末到来时候经常在购买啤酒和婴儿尿布。关联规则分析还被应用到很多的场景中，如cross-selling，辅助决策，顾客行为分析和客户关系管理（tang等人[[5]](#endnote-5)2008，Haughton 等人[[6]](#endnote-6)2003等）。在2010年，Ting和Steve等人[[7]](#endnote-7)将购物篮的关联分析应用到了餐馆的菜单选择上。

MBA的目的是利用商品的关联性诱导客户购买非计划外的商品。2009年Bell等人[[8]](#endnote-8)发现非计划购买不是一种常见现象。顾客购买行为更多的以来于用户的计划和习惯，还有收集的商品信息。关联性本身可能与直觉相反，例如MBA方法有时错误的辨别互补商品（顾客喜欢一起购买的商品）或者是替代品（多购买一个可能是为了迎合别人的意愿）。2011年Bogdan Hoanca等人[[9]](#endnote-9)用MBA方法根据顾客的“价格敏感度和关联的饱和度”来估算一个大学小书店的潜在收入增长。同一年Chrndra等人[[10]](#endnote-10)研究如何设计一个好的样本子集来进行购物篮分析，因为在交易量巨大的企业中很难直接在整个数据集合采用购物篮分析。2013年Wiley Interdisc等人[[11]](#endnote-11)使用MapReduce来进行购物篮分析。

#### 1.2.2 社交网络影响力传播研究现状

在社会生活中，个体之间都存在相互影响的关系，一个人的偏好会受到其朋友，家人等的影响，同时也会影响到其他人。比如对某件商品的评价，可能会影响其他人对该商品的评价，同样对音乐的评价，餐馆的评价等也存在这种影响的传播，在社交网络中叫做口碑传播，也叫做影响力的传播。

深入理解人类的行为和影响力的传播，商家可以给顾客提供更加个性化和舒适的服务，及迅速提高品牌价值。政府可以根据影响力的传播特性，迅速找到部分高影响力的人群快速推广政策，也可以截断高影响传播路径迅速控制谣言的传播等。

社会影响力的研究在社会科学和市场已经有很长的研究历史了，Christakis和Fowler利用美国上万人的32年的医疗数据，验证了肥胖症和吸烟行为会在社交网络中相互影响和传播。随着在线社交媒体的发展和大数据技术的应用，让更大规模的社交行为数据的研究称为可能。21世纪基于互联网的大数据的影响力传播开始兴起，并成为数据挖掘和社交网分析的热点。影响力传播研究有三大支柱，第一是影响力传播模型，主要描述影响力在社交网络中如何传播、有何特点和性质。第二是影响力传播的学习，即如何利用网络大数据挖掘学习影响力传播模式和具体传播模型的参数。第三是影响力传播优化，着重考虑在不同的传播模型下，如何通过施加外部作用来扩大希望传播的影响力或者控制和减弱不希望传播的影响力，也包括有效地监控影响力的传播等。

计算机科学上首先是2001年Domingos和 Richardson在提出了基于马尔科夫随机场（Markov random field）的社交网络影响力模型[[12]](#endnote-12)。2003年，Kempe、Kleinberg等人[[13]](#endnote-13)提出独立级联和线性阈值模型等离散递进性传播模型和它们的扩展模型。这些模型简单，符合人们对影响力传播的直觉理解，现在成为了影响力传播的经典模型。为了满足影响力最大化的不同需求很多的影响力传播模型被提出，Chen等人提出IC-N模型[[14]](#endnote-14)，一种同时考虑积极影响和消极影响的扩展IC模型。后来Borodin等人提出一种LT模型的扩展模型，叫做competitive LT模型[[15]](#endnote-15)。2008年Hao Ma等人提出热力传播模型HDM[[16]](#endnote-16)，该模型引入传导率和传播时间等概念，尝试去模拟复杂的社交网络影响力传播。2014年，Kim等人[[17]](#endnote-17)提出了一种新奇的模型，被叫做连续性激活和时间限制的IC(CT-IC)模型，在这个模型里，当传播模型符合这几种模式[[18]](#endnote-18)[[19]](#endnote-19)[[20]](#endnote-20)的时候，允许节点被多次激活。

影响力传播建模的一个主要目的是控制和优化影响力的传播，这其中被广泛研究的核心问题是影响力最大化(influence maximization)问题。影响力最大化问题是一个组合优化问题，也可以被理解为图覆盖的问题的扩展。因而与图覆盖问题一样,在这些影响力传播模型下，影响力最大化问题也是一个NP难的问题。有很多的算法被提出尝试去求近似最优解，2003年，Kempe等人[[21]](#endnote-21)用贪心算法去求解影响力最大化的问题。2007年Leskovec J等人[[22]](#endnote-22)基于影响力最大化问题的子模特性提出了CELF算法，这个算法可以减少节点影响力的求取次数，原始的蒙特卡洛方法应用CELF方法就有上百倍的提高。2014年Borgs C等人[[23]](#endnote-23)提出了反向蒙特卡洛算法，这种算法随机的选取图上的节点，从该节点出发以所有边的相反方向进行蒙特卡洛模拟，达到世界上最可能影响该节点的集合。这个集合被称为反向可达集合。如果节点经常在RR集合中出现，那么该节点就是一个影响力最大的节点，这种算法接近线性时间的同时仍然有()的近似比保证。

### 1.3 研究内容、意义、思路及架构

#### 1.3.1 研究内容

（1）了解商品数据，并清洗数据。

（2）在阅读相关文献的基础上，寻找求取商品关联性的方法，并构建商品网络。

（3）学习复杂网络的一些理论知识，并借助复杂网络分析软件分析商品网络。

（4）学习社区分析的一些方法，对商品网络进行社区划分，并分析每个社区的特点。

（5）学习影响力分析的一些方法，探究影响力传播模型在商品网络中的表现。并利用影响力最化方法从中选取影响力最大的K个商品。

1.3.2 研究意义

随着经济的发展，超市的数量也在增加，一方面超市每天的交易量是很大的，尤其是节假日。另一方面超市超市数量的增加，说明超市的竞争也越大，能获得顾客喜爱的超市才能在激烈的竞争中生存下来。

在日常生活中，我们可以看到，超市通过各种促销活动，调整商品摆放等，这些都是精心为顾客带来更多的便利以及引导顾客消费。那么问题来了，促销的时候选择那些商品商品来进行促销呢？顾客的购买习惯是怎样的呢，如何进行商品的摆放才能给顾客带来更多的便利？

本文尝试根据顾客的购买记录构建一个商品网络，每件商品是网络的节点，边表示商品间存在关联，边权表示商品关联性强弱。利用社区划分，分析那些商品的联系更加紧密，可以帮助调整商品的摆放。利用影响力最大化分析，尝试找出影响力最大的商品，可以给商品促销提供更多的信息。

#### 1.3.3 技术路线

本文研究的主要目的是应用复杂网络分析的思想以及社交网络影响力分析的思想对商品网络进行建模分析。本文技术路线如下：

（1）数据清洗

理解原始数据各个字段的含义，抽取所需的字段并进行数据清洗。可视化数据的分布直观理解数据的特征。

（2）构建商品网络，并了解该商品网络的特征。

使用apriori的思想求取两两商品之间的关联性，然后构建起一个商品网络。使用复杂网络的思想，分析该网络的特征。

（3）对超市购物篮网络进行社团划分。

2008年Vincent等人提出了无向图的快速层次聚类。本研究因为使用的是有向图，将结合Leicht等人使用有向图聚类思想，推导出有向图的快速层次聚类。该步骤将研究商品社团的一些特点。

（4）尝试寻找影响力最大的K件商品

探究与传播时间有关的热量传播模型和经典的独立级联模型在商品网络中的表现，并利用贪心算法求取影响力最大的K件商品。

1.3.4 论文架构

第一章，绪论

本章介绍本研究的课题及选题背景，国内外研究综述，讨论研究内容、意义、思路及架构，并描述本文的创新点。

第二章， 社交网络的相关理论

本章介绍本研究中会用到的社交网络的基础知识，包括社交网络的定义，相关统计特性，影响力传播模型，影响力最大化问题定义。本章作为知识铺垫。

第三章，基于购物篮的商品网络分析

本章将会介绍数据的筛选和清洗，购物篮数据特点，如何构建商品网络，商品网络的特点。

第四章，基于有向图快速层次聚类的商品社团划分

本章首先介绍社区划分的理论知识，包括传统的层次聚类思想，Vincent等人提出的无向图的快速层次聚类，以及如何应用到有向图中。在理论知识支撑下，对商品网络进行社团划分和分析社团的特点。

第五章，基于社交网络影响力最大化的购物篮方法研究

本章将会介绍社交网络的两个影响力传播模型，与时间有关的热量传播模型，经典的独立级联模型。还会介绍影响力最大化问题的定义，并且探究影响力最大化方法在求取K个影响力最大化商品的适用性。

### 1.4 本文主要创新点

（1）利用商品网络和社团划分从整体的角度研究购物篮特点

传统的购物篮方法大多基于关联规则的，该方法只能得到商品之间的直接关系，不适用于从整体上分析购物篮的特点。国外虽然有使用商品网络来分析购物篮特点的文献，如2012年Hyea Kyeong Kim等人（A product network analysis for extending the market basket analysis），但是基于社团划分的购物篮分析方法较少。

本研究将基于关联规则构建商品网络，进而使用快速层次聚类进行划分，从整体上去了解商品的特点。

（2）利用社交网络影响力最大化的思想挑选影响力最大的K件商品

社交网络影响力最大化的研究现在是社交网络分析的一个热点，理论研究也比较成熟，但是在商品网络中仍没找到相应的应用研究文献。在本文中将探究社交网络影响力最大化方法在商品网络中的适用性。

## 第二章 本文的相关理论

### 2.1网络的定义

网络（图）的简单的形式可用G = (V, E)表示, 其中是节点集合， E是边的集合，所谓的边就是连接两个节点的线。如果边上有权值的话表示带权图，如果边有方向则成为有向图。

当给节点和边赋予不同的含义的时候，图可以表示不同的网络。如果节点是神经元，边是神经连接，则可以表示神经网络。如果节点是物种，边是捕食关系，则可以表示食物链网络。如果节点是人，边表示人与人之间的联系，则可以表示社交网络。

### 2.2 网络的性质

#### 2.2.1 邻接矩阵

网络的表示方法有很多种，这里使用邻接矩阵来表示。对于一个图G = (V, E),可用邻接矩阵A来表示，其中表示存在一条从节点i和节点j的关联情况。

对于边上没有权值的无网络，

对于边上没有权值的有向网络

#### 2.2.2 权值矩阵

我们将图的边权矩阵记作W，类似于邻接矩阵表示，不同的是节点i与节点j边上的权值，而不是0和1。

#### 2.2.3 度数和带权度数

对于图G = (V, E)，记节点vV，

无向图节点v的度数

deg(j) =

有向图节点v的出度数，入度数，度数分别为：

deg\_out(i) = ，

deg\_in(i) =

deg(i) = deg\_out(i) + deg\_out\_in(i)

如果是一个有向带权图的话，

deg \_with\_weight\_out (i) = ，

deg \_with\_weight\_in (i) =

deg\_with\_weight (i) = deg \_with\_weight\_out (i) + deg\_out \_with\_weight\_in (i)

#### 2.2.4 网络直径

网络中两个节点的距离是两个节点的最短路径长度，记作。而图的直径是网络中任意两个节点最短路径长度的最大值，

#### 2.2.5 有向图的模块值

模块值是网络结构中的一种度量。可以用来度量社团划分的结果，图的模块值越大说明社团内部的连接越紧密，而社团之间的连接越稀疏。

模块值的定义是，给定一些社区，社区内的边减去边随机分布时候边的期望。有向图的模块值[[24]](#endnote-24)如下：

### 2.3 社交网络影响力传播模型

为了方便阐述，简单的将社交网络定义为有向图G = (V，E，W)，其中V是节点的集合，E是边的集合，W是表的权值。在社交网络中，一个人可以用一个节点来表示，u节点对v节点的影响关系可以用一条边来表示，影响强度用表示。

通常会给每个节点两个可选的状态，活跃状态和非活跃状态。网络中传播的信息，想法或者产品统称实体。那么一个节点接受了相应的实体，我们称该节点出于活跃状态，否则该节点处于非活跃状态。一个节点由非活跃状态变成活跃状态，我们称该节点被激活，他就成为了实体的新传播源。

影响力模型刻画了影响力在网络中的传播模式，包括一个节点的状态如何影响到其他节点的状态以及影响力在网络中如何扩散的。

影响力的传播模型分为很多种，随机模型，博弈论模型，传热病模型和热量传播模型等。随机模型按传播特点有可分为离散时间/连续时间，递进/非递进。传播模型在商品网络中的研究较少，所以本文以探究实验为主，选择了两种经典的传播模型，2010wei chen等人的基于随机级联的独立级联模型[[25]](#endnote-25)，热量传播模型5。

### 2.4 影响力最大化问题定义

对于一个给定网络G = (V，E)，影响力传播模型和所需的参数（如激活阈值，边的权值等），影响力最大化就是寻找由个节点组成的种子集合，使得的延展度最大，可表示为。影响力最大化问题属于组合最优化问题，在不同的影响力传播模型下该问题解决难度不同，如果在独立级联模型下，该问题属于的NP难问题，常见的采用贪心算法来逼近最优解。

### 2.5 关联规则算法

关联规则算法是购物篮分析的一种常用算法，它可以用来分析商品被一起购买的可能性，也可以用来分析购买一些商品后购买另一些商品的可能性。关联规则的一种经典算法是apriori算法，这个算法的剪枝思想是如果S是频繁项集，那么S的任何一个子集也是频繁项集，如果S是非频繁项集合，那么S的任何超集也是非频繁项集。

支持度：项集在数据库中出现的频率。

置信度：在一个项集出现的情况下，另一个项集出现的概率

Apriori计算关联规则主要是利用了条件概率：

P(AB)是AB项集出现的频率，P(B)是B项集出现的频率，A项集B项集的置信度是。

## 第三章 基于购物篮的商品网络分析

本章将会介绍数据的筛选和清洗，购物篮数据特点，如何构建商品网络，商品网络的特点。

### 3.1 数据筛选

原始数据中含有的字段有日期，销售流水号，商品ID，商品名称等字段，在这个实验中，并不需要所有的数据，根据时间字段筛选其中的一个月数据，再从中选择所需要的销售流水号，商品小类。以销售流水号作为连接的键，就可以得到每个交易的商品组合了。

这里可以贴上统计的商品小类在交易中出现的次数

### 3.2 构建商品网络

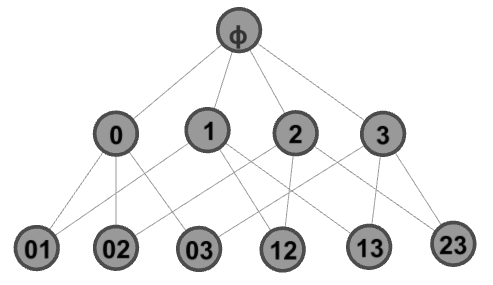
构建商品网络我们将商品小类作为网络的节点，只要商品小类在一次交易中同时出现就认为其存在关系，然后使用apriori算法的条件概率的思想来构建边。

我们可以使用apriori算法来求取项集频率和商品小类的关联规则。使用一个简单的例来理解这个过程，假设有4个商品小类{0, 1, 2, 3}，交易记录如下：

|  |  |
| --- | --- |
| Transaction | Items |
| T1 | 0,1,3 |
| T2 | 1,2 |
| T3 | 1,2,3 |

假设项大小为1和2的项集分别记作L1和L2。

项集出现频率的求取过程是，先求L1出现的频率，然后再将L1组合成L2，求L2的出现频率。



两两的关联规则好求写，知道了L2项集，那么可以将L2项集的每个商品小类集合拆分开，根据条件概率求彼此的关联规则，如{0,1}商品小类集合：

### 3.3 关联规则算法设计：

求项大小为1的项集伪代码如下：

|  |
| --- |
| Algorithm：createL1(dataSet)  输入：dataSet  输出：项大小为1的项集L1  L1←empty list  for each transaction in dataset  for each item in transaction  if [item] not in C1  insert [item] to C1  输出C1 |

求项大小为K的项集伪代码如下：

|  |
| --- |
| Algorithm：aprioriLn(Lk)  输入: 项大小为K的项集Lk  输出：项大小为K+1的项集Ln  Ln←empty list  lenLk←the length of Lk  n←Lk中项的大小 + 1  for each i in [0,1...lenLk-1]  for each j in [i+1,i+2...lenLk-1]  C1 = list(Lk[i])[:n-2]  C2 = list(Lk[j])[:n-2]  if C1 equal to C2  insert Lk[i]∪Lk[j] to Ln  输出Ln |

计算项集的出现频率伪代码如下：

|  |
| --- |
| Algorithm：scanD(dataset, Lk, minSupport)  输入：交易记录集dataSet，项集Lk，最小支持度minSupport  输出：满足约束条件的项集ret\_Lk，项集的中项出现频率supportData  ItemsetCnt←empty dictionary  for each tid in dataset  for each itemset in Lk  if itemset is not a subset of tid  if itemset is not one of key in ItemsetCnt  ItemsetCnt [ itemset ]←1  else  ItemsetCnt [ itemset ]←ItemsetCnt [ itemset ] + 1  numItems←float(length of dataSet)  ret\_Lk←empty list  supportData←empty dictionary  for each key in ItemsetCnt  support←ItemsetCnt[key] /numItems  if support >= minSupport  insert key to ret\_Lk  supportData[key]←support  输出：ret\_Lk，supportData |

计算任意两个商品小类间的关联规则伪代码如下：

|  |
| --- |
| generateRules(Lk, supportData, minConf)  输入：项集列表L，输入每一项的支持度supportData，最小置信度minConf  输出：关联规则ruleList  ruleList←empty list  len\_Lk←length of Lk  for i in [1...len\_Lk-1]  for each itemset in L[i]  for item in itemset  confidence←supportData[item]/supportData[itemset-item]  if confidence >= minConf  insert (itemset-item, item, confidence) in ruleList  输出：ruleList |

### 3.4 商品网络的特点

#### 3.4.1 商品网络的直径

#### 3.4.2 商品网络边的特点

#### 3.4.3 商品网络节点度数的分布

#### 3.4.4 商品网络带权度数的分布

## 第四章 基于有向图快速层次聚类的商品社团划分

### 4.1 社区发现

社区发现是复杂网络分析的一个经典问题。社区划分是将一个网络划分成多个社区，使社区内的节点联系紧密，而社区间的连接稀疏。将网络的一部分节点划分到一个社区划的质量好坏，可以用模块度量。模块值越大，说明这些节点更可能在一个社区。模块值最大约束下，将社区划分成多个社区的问题是一个NP难问题。社区发现的方式有很多种，比如1970年Kernighan等人[[26]](#endnote-26)基于贪心算法的二分法，1973年Fiedler等人[[27]](#endnote-27)基于谱二分法，2002年Girvan等人[[28]](#endnote-28)通过移除边聚集系数最大的边来划分社区，2004年Newman等人[[29]](#endnote-29)基于模块值的聚类方法。在2008年Blondel等人[[30]](#endnote-30)提出来高效的层次聚类。社区发现在近两年也有很多的研究，2014年，Zhou等人[[31]](#endnote-31)提出了一种结合博弈论和Shapley Value值去评估个体紧密连接的聚类方法。在2016年feiping nie等人[[32]](#endnote-32)提出了一种高效的聚类方式，这种聚类方法能通过不断学习相似度矩阵获得K个簇，而不需用K均值来获得聚类。

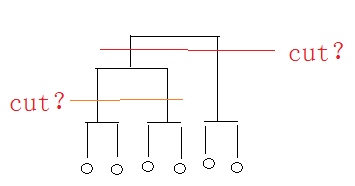
在本文中，依然采用经典的算法，使用2007年Newman等人[[33]](#endnote-33)提出的有向图模块值计算公式，将2008年Vincent D Blondel等人30的无向图快速层次聚类调整成适用于有有向图快速聚类方法。

### 4.2 层次聚类算法

#### 4.2.1 层次聚类思想

|  |
| --- |
| 1、将每个节点看作是一个社区  2、合并两个距离最近（或相似度最大）的社区为一个新的社区  3、直到最终成为一个社区 |

如下图，层次聚类需要知道在哪里停止聚类，即一个评估聚类好坏的值，有的算法使用模块值来衡量聚类结果，在4.2.3节会讲到该模块值。



#### 4.2.1 无向图快速层次聚类

给定一个无向图，是节点v的带权度，是节点w的带权度，是从v节点和w节点的边的权值。

2004年M. E. J. Newman等人提出的无向图模块值，计算公式如下：

当节点v和节点w在同一个社区的时候为1，否则为0。

根据2004年Aaron Clauset等人的文献对模块值进行简化：

引入：

=，

社区i和社区j连接的边权和与图总的边权和比值为，

社区i中节点的边权和与图总的边权和的比值为，

Q值计算公式可简化为：

一个孤立点i加入社区C的模块值增益：

其中表示当前社区C内的边权和，是节点i连接C内部边权和，是连接到社区C的节点的边权和，m是图的边权值和，无向图 。

在上面的公式基础上，快速层次聚类伪代码如下：

|  |
| --- |
| 1、将每个节点看作是一个社区  2、计算Q值  3、for vertex in Graph:  计算vertex加入与其关联的社区带来的值  选取 最大值，如果则，将节点  加入相应的社区，否则该节点保持原状。  4、计算Q值，比上次状态的Q值变化大于则跳至3  5、否则输出聚类结果 |

#### 4.2.2 有向图快速层次聚类

因为研究的商品网络是有向的，所以需要对算法进行调整，在2007年E. A. Leicht1等人提出了有向图的模块值计算：

给定一个有向图G，表示节点v的带权出度，表示节点w的带权出度，当节点

其中

当节点v和节点w在同一个社区的时候，为1否则为0。

类似于无向图模块值的简化，引入：

社区i和社区j连接的边权和与图总的边权和比值为，

社区i中节点入度和与图总的边权和比值为，

社区i中节点的出度和与图总的边权和比值为，

有向图的模块值可以简化为：

有向图中，一个孤立点加入社区后的模块值的增益：

其中：

表示当前社区C内的边权和，

是节点i连接C内部边权和，

是边的终点是社区C的节点的边权和，

是边的起点是社区C的节点的边权和，

有向图边权和 。

### 4.3 商品网络社团划分

## 第五章 基于社交网络影响力最大化的购物篮方法研究

第五章，基于社交网络影响力最大化的购物篮方法研究

本章将会介绍社交网络的两个影响力传播模型，与时间有关的热量传播模型，经典的独立级联模型。还会介绍影响力最大化问题的定义，并且探究影响力最大化方法在求取K个影响力最大化商品的适用性。

### 5.1 影响力传播模型

#### 5.1.1 热量传播模型

为了捕捉市场暂时的信息传播，2008年Hao Ma等人提出了使用热量传播模型来对复杂的社交网络的影响力传播模（HDM）型进行建模。HDM模型提供了很多可以模拟真实条件的参数，比如时间和传导率。在选择商品促销的时候，存在一个时间的概念，促销多久会带来比较好的结果，所以该实验使用HDM模型来对商品网络影响力传播建模。

在给定的一个有向图G = {V,E,P},其中V是节点的集合，E是边的集合，P是边上的权值的集合。为了简化模型，需要先做出如下的假定：

（1）在时间，节点vi传播出去的温度记作DH(i, t, )。

（2）DH(i, t, )是均匀传播出去的，传播的温度按度数进行均分

（3）所有的节点有相同的热量传播能力

（4）热量的传播方向跟边的方向有关，且不受温度差影响。

根据Hao Ma等人的文章，节点i在t+时间内的变化等于它接受邻居的温度减去节点i传播给邻居的温度：

其中

代表节点j存在的概率，实验中会取1。用于标识节点i的出度，i的出度大于0时候，则为1，否则为0。

当时候有：

求积分后可得：

使用热量传播模型求节点影响力的伪代码如下：

|  |
| --- |
| Algorithm HeatDiffusion(g, S, h0, a, t, ):  输入：图g，种子集合S，初始温度h0，传导率a，时间t, 激活阈值  输出：影响力num  num←0  f←zeros array  for each v in S  ←h0  计算f(t) =  for each v in g  if  num += 1  输出：num |

#### 5.1.2独立级联模型

2003年，Kempe、Kleinberg等人提出独立级联和线性阈值模型等离散递进性传播模型和它们的扩展模型。这些模型简单，符合人们对影响力传播的直觉理解，现在成为了影响力传播的经典模型。2009年W. Chen等人改进了独立级联模型，通过概率1-P移除边，然后计算从每个节点出发的可达节点数作为相应节点的影响力，为了使结果更准确进行多次实现取平均可达数作为节点的影响力。本次实验将使用这个影响力传播模型。

对于一个给定的有向图G = (V,E,P)，V是节点集，E是边集合，P是边的权值集合，表示从节点u到节点v的边的权重。让在第t轮被激活的种子集合。 = S表示初始的种子集合。在t>0轮，一个给定的节点uV，被它在t-1步骤激活的邻居节点v影响的概率是独立的，其中。在这样的模型下节点u被激活的概率可以用下面的公式来表示：

按照上面的迭代步骤，直到为空。

u节点只会尝试一次通过激活v节点，因此我们可以换个角度，可以预先通过一个随机函数，边可以选择作为影响力传播的边，否则认为边不被选择作为影响力传播的边并移除该边，这样就可以得到一个新的图G’。在图G’中，种子集合S可达的节点集合记作。这样子做的好处是，我们可以通过求取强联通组件，得到相互可达的多个节点集合记作RS,表示第i个强联通组件。对于节点v，有。对于节点V\S，如果节点，那么在选择影响力最大节点时候为了避免选择影响重叠节点，就将v的影响力记作0，否则为。

### 5.2 影响力最大化问题的定义

影响力最大化问题的定义是，在给定的网络G=(V, E)，影响力传播模型和所需的参数的情况下，从网络中选择K个种子节点作为种子集合，种子集合的影响扩散后使被激活节点数最大，即。

在热量传播模型下，使用贪心算法求影响力最大种子集合伪代码如下：

|  |
| --- |
| Algorithm Enhanced k-Step Greedy Algorithm  输入：图g，初始温度h0，传导率a，时间t, 激活阈值  输出：影响力最大的K节点种子集和影响力influence  for each l in [1…k]  max\_influ -1  vertex = 0  for each v in V\S  influ HeatDiffusion(g, S{v}, h0, a, t, )  if influ max\_influ  max\_influinflu  vertexv  SS{vertex}  Influencemax\_influ  输出：S，influence |

独立级联模型下，使用贪心算法求影响力最大种子集合伪代码如下：

|  |
| --- |
| Algorithm NewGreedIC(G, k)  输入：有向图G，种子数k  输出：种子集合S  S，N  for each I in [1,2…k]  for each v in V\S  for i in [1,2..N]  从图G中移除边当  计算  计算  For each v in V\S  If v    end if  end for  for each v in V\S  S = S{}  end for  end for |

### 5.3 商品网络K个影响力最大化商品的发现

1. Julander, C.-R. (1992). Basket analysis: A new way of analysing scanner data. *International Journal of Retail & Distribution Management , 20* (7), 10-18. [↑](#endnote-ref-1)
2. Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD International Conference on Management of Data (SIGNOD ’93)* (pp. 207-216). Washington DC: ACM. [↑](#endnote-ref-2)
3. Chen, Y. L., K. Tang, R. J. Shen, and Y. H. Hu. 2005. Market basket analysis in a multiple store environment. *Decision Support* *Systems* 40 (2): 339-54. [↑](#endnote-ref-3)
4. Berry, M. J. A., and G. S. Linoff. 2004. *Data mining techniques—For marketing, sales, and customer support*. 2nd ed. New York:Wiley. [↑](#endnote-ref-4)
5. Tang, K., Y. L. Chen, and H. W. Hu. 2008. Context-based market basket analysis in a multiple-store environment. *Decision Support* *System* 45:150-63. [↑](#endnote-ref-5)
6. Haughton, D., J. Deichmann, A. Eshghi, and S. Sayek. 2003. A review of software packages for data mining. *The American* *Statistician* 57 (4): 290-309. [↑](#endnote-ref-6)
7. Ting, P.-H., Pan, S., & Chou, S.-S. (2010). Finding Ideal Menu Items Assortments: An Empirical Application of Market Basket Analysis. *Cornell Hospitality Quarterly , 51* (4), 492-501. [↑](#endnote-ref-7)
8. Bell, D. R., Corsten, D., & Knox, G. (2009, January). Unplanned Category Purchase Incidence: Who Does It, How Often, and Why. *Knowledge@Wharton* . [↑](#endnote-ref-8)
9. Hoanca, B. & Mock, K. “Using Market Basket Analysis to Estimate Potential Revenue Increases for a Small University Bookstore.” Conference for Information Systems Applied Research, Vol.4, No.1822, 2011. [↑](#endnote-ref-9)
10. Chandra, B., & Bhaskar, S. (2011). A new approach for generating efficient sample from market basket data. *Expert Systems with Applications , 38*, 1321–1325. [↑](#endnote-ref-10)
11. Market Basket Analysis algorithms with MapReduce. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery 3(6): 445-452 (2013) [↑](#endnote-ref-11)
12. Domingos P, Richardson M. Mining the network value of customers. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, USA, 2001: 57~66 [↑](#endnote-ref-12)
13. Kempe D, Kleinberg J M, Tardos É. Maximizing the spread of influence through a social network. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ( K DD ), Washington D C, USA, 2003: 137~146 [↑](#endnote-ref-13)
14. W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu,D. Rincon, X. Sun, Y. Wang, W. Wei, and Y. Yuan.

    Influence maximization in social networks when negative opinions may emerge and propagate. In

    SDM, volume 11, pages 379{390. SIAM, 2011 [↑](#endnote-ref-14)
15. A. Borodin, Y. Filmus, and J. Oren. Threshold models for competitive influence in social networks. In Internet and network economics, pages 539-550. Springer, 2010 [↑](#endnote-ref-15)
16. H. Ma, H. Yang, M. Lyu, and I. King. 2008. Mining social networks using heat diffusion processes for marketing candidates selection. In *Proceedings of the 17th ACM Conference on Information and Knowledge* *Management (CIKM’08)*. 233–242. [↑](#endnote-ref-16)
17. J. Kim, W. Lee, and H. Yu. CT-IC: Continuously activated and time-restricted independent cascade model for viral marketing. Knowledge-Based Systems, 62:57-68, 2014. [↑](#endnote-ref-17)
18. W. Chen, W. Lu, and N. Zhang. Time-critical

    influence maximization in social networks with time-delayed diffusion process. arXiv preprint arXiv:1204.3074, 2012. [↑](#endnote-ref-18)
19. Y. Li, W. Chen, Y. Wang, and Z.-L. Zhang. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In WSDM, pages 657{666. ACM, 2013. [↑](#endnote-ref-19)
20. A. Goyal, F. Bonchi, and L. V. Lakshmanan. A data-based approach to social influence maximization. VLDB, 5(1):73{84, 2011. [↑](#endnote-ref-20)
21. Kempe D, Kleinberg J M, Tardos É. Maximizing the spread of influence through a social network. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ( K DD ), Washington D C, USA, 2003: 137~146 [↑](#endnote-ref-21)
22. Leskovec J, Krause A, Guestin C, et al. Cost- effective outbreak detection in networks. Proceedings of the 13rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Jose, USA, 2007: 420 ~429 [↑](#endnote-ref-22)
23. Borgs C, Brautbar M, Chayes J, et al. Maximizing social influence in nearly

    Optimal time. Proceedings of ACMSIAM Symposium on Discrete Algorithms

    (SODA), Portland, USA, 2014: 946~957 [↑](#endnote-ref-23)
24. Community structure in directed networks E. A. Leicht1 and M. E. J. Newman1, 2 （得改） [↑](#endnote-ref-24)
25. **Efficient Influence Maximization in Social Networks** [↑](#endnote-ref-25)
26. KERNIGHAN, B.W. and S. LIN (1970) An efficient heuristic procedure for structure graphs, Bell System Technical Journal, 49, 291–307. [↑](#endnote-ref-26)
27. FIEDLER, M. (1973) Algebraic connectivity of graphs, Czechoslovak Mathematical Journal, 23, 298–305. [↑](#endnote-ref-27)
28. GIRVAN, M. and M.E.J. NEWMAN (2002) Community structure in social and biological networks, Proceedings of the National Academy of Science, 99, 7821–7826. [↑](#endnote-ref-28)
29. NEWMAN, M.E.J. and M. GIRVAN (2004) Finding and evaluating community structure in networks, Physical Review E, 69, 026113, 1–16. [↑](#endnote-ref-29)
30. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast Unfolding of Communities in Large Networks. J. Stat. Mech. Theory Exp. 2008, P10008, 1–11. [↑](#endnote-ref-30)
31. ZHOU, L. and K. LÜ (2014) Detecting communities with different sizes for social network analysis, The Computer Journal. DOI:10.1093/comjnl/bxu087. [↑](#endnote-ref-31)
32. Feiping Nie, Xiaoqian Wang, Michael I. Jordan, Heng Huang. The Constrained Laplacian Rank Algorithm for Graph-Based Clustering. The 30th AAAI Conference on Artificial Intelligence (AAAI), Phoenix, USA, 2016. [↑](#endnote-ref-32)
33. Community structure in directed networks [↑](#endnote-ref-33)