|  |
| --- |
| 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：121班  学 生：袁国文  学 号：20123151  指导教师：杨书新 职称：副教授 |

摘 要

本文旨在为零售商选择促销商商品提供一些商品组合，这些商品组合要尽可能的带动其他商品的销售。基于这样的场景，商品的选择是一个组合最优化的问题。通过分析购物篮研究的国内外研究现状，发现购物篮常用方法是关联规则分析，而商品网络的研究仍比较少，并且社交网络中影响力最大化问题可以解决组合最优化问题，所以本文尝试探究社交网络影响力最大化算法在商品网络寻找影响力最大的商品组合。

本文的算法设计主要分为三个步骤，第一是先使用apriori算法，求取商品之间的关联性。然后以商品作为节点，在同一频繁项集的商品认为存在边的关系，关联性的方向决定了边的方向，关联性的强度最为边的权值，那么就可以得到一个有向图。第二是使用独立级联模型和热量传播模型对商品影响力传播的建模。第三是效仿2013年hui li等人18的求解影响力最大化问题的贪心算法，先进行社区划分，然后在从各个社区选出能使当前影响力增益最大的商品加入到商品集合中。

在真实的交易数据中对本文设计的算法进行了实验探究，发现基于独立级联模型的影响力最大化算法寻找的商品呈现出对邻居节点的影响较强且销量较高的特点，而基于热量传播模型的影响力最大化算法寻找的商品同样呈现出对邻居节点的影响较强，但是商品的销量较低的特点。

**关键字：**购物篮分析；关联规则；商品网络；社团划分；影响力最大化

**ABSTRACT**

This paper aims to provide some combination of goods for retailers to choose promotional commercial goods, these goods should be combined to greatly drive other merchandise sales. Based on this scenario, the choice of goods is a combinatorial optimization problem. Through the analysis of market basket research at home and abroad research status, find that the general method of market basket analysis is association rules analysis, and network product research is still relatively less, and the influence maximization algorithm in social networks can solve the combinatorial optimization problem, therefore, this paper attempts to explore social networks influence maximization algorithm in the network of commodity for selecting commodity combination which had the greatest influence.

The algorithm design in this paper has three steps, the first is to use the apriori algorithm to obtain the correlation between the goods. Then take the commodity as a node, two commodity has a edge when they are in the same frequent item set, the direction of association determines the direction of the edge and the strength of the association will be the edge weights, then you can create a directed graph. The second is to use the independent cascade model and the heat diffusion to modeling the influence spreading between goods. The third is follow 2013 Hui Li et al. 18 for using the greedy algorithm to resolve the influence maximization problem, the first is community division and then select commodity which will bring the maximization of gain from various communities and added to collection of items.

The algorithm designed in this paper was test in the real transaction data.The result show that influence maximization algorithm based on the independent cascade model select the commodity which effect his neighbor nodes strongly and better sales，while influence maximization algorithm based on the heat diffusion model select the commodity which effect his neighbor nodes strongly but poor sales.

**Keywords:** market basket analysis;commodities network;commodity division; influence maximization algorithm

**目录**

[第一章 绪论 1](#_Toc451546420)

[1.1研究背景与意义 1](#_Toc451546421)

[1.2国内外研究现状 2](#_Toc451546422)

[1.2.1基于商品网络的购物篮分析问题的研究现状 2](#_Toc451546423)

[1.2.2社交网络影响力最大化的研究现状 3](#_Toc451546424)

[1.3本文的主要工作 4](#_Toc451546425)

[1.4本文组织结构 5](#_Toc451546426)

[第二章 基于商品网络的购物篮分析概述 6](#_Toc451546427)

[2.1引言 6](#_Toc451546428)

[2.2购物篮的定义和应用研究 6](#_Toc451546429)

[2.3apriori算法 7](#_Toc451546430)

[2.3.1定义和性质 7](#_Toc451546431)

[2.3.2apriori的计算过程 8](#_Toc451546432)

[2.3.3apriori减枝思想 8](#_Toc451546433)

[2.4常见构建商品网络的方法 10](#_Toc451546434)

[2.5商品网络分析常见方法 11](#_Toc451546435)

[2.5.1社团发现 12](#_Toc451546436)

[2.4.2重要商品的发现 12](#_Toc451546437)

[2.6本章小结 13](#_Toc451546438)

[第三章 社交网络影响力最大化概述 13](#_Toc451546439)

[3.1引言 13](#_Toc451546440)

[3.2预备知识 13](#_Toc451546441)

[3.2.1社交网络概述 13](#_Toc451546442)

[3.2.2影响力最大化问题的定义 14](#_Toc451546443)

[3.3影响力传播建模 15](#_Toc451546444)

[3.3.1线性阈值模型 15](#_Toc451546445)

[3.3.2独立级联模型 15](#_Toc451546446)

[3.3.3热量传播模型 16](#_Toc451546447)

[3.3.4其他传播模型 17](#_Toc451546448)

[3.4影响力最大化关键技术 18](#_Toc451546449)

[3.4.1影响力最大化问题求解算法概述 18](#_Toc451546450)

[3.4.2有向图快速层次聚类算法 19](#_Toc451546451)

[3.5 本章小结 21](#_Toc451546452)

[第四章 影响力最大化算法在商品网络中的应用 22](#_Toc451546453)

[4.1引言 22](#_Toc451546454)

[4.2商品网络的构建 22](#_Toc451546455)

[4.2独立级联模型在商品网络的应用 25](#_Toc451546456)

[4.3基于热量传播模型的影响力最大化算法设计 26](#_Toc451546457)

[4.4基于贪心思想的影响力最大化算法 28](#_Toc451546458)

[4.5本章小结 30](#_Toc451546459)

[第五章 实验 31](#_Toc451546460)

[5.1引言 31](#_Toc451546461)

[5.2实验背景 31](#_Toc451546462)

[5.3.1实验环境 31](#_Toc451546463)

[5.3.2实验数据集 31](#_Toc451546464)

[5.3.2商品特征的选择 32](#_Toc451546465)

[5.3实验结果 32](#_Toc451546466)

[5.3.1社区划分结果 32](#_Toc451546467)

[5.3.2基于独立级联模型的影响力最大化算法实验结果 34](#_Toc451546468)

[5.3.3基于热量传播模型的影响力最大化算法实验结果 36](#_Toc451546469)

[5.3.4对比简析 37](#_Toc451546470)

[5.5本章小结 37](#_Toc451546471)

[第六章 总结与展望 38](#_Toc451546472)

[附录.. . 40](#_Toc451546473)

[独立级联模型的主要代码 40](#_Toc451546474)

[热量传播模型的主要代码 42](#_Toc451546475)

[社区划分的主要代码 46](#_Toc451546476)

[参考文献 50](#_Toc451546477)

1. **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。

啤酒和尿布的故事体现了MBA研究的商业价值，MBA的研究有很多，但是基于商品网络来进行MBA研究仍比较少，本文将利用apriori算法求取商品关联性构建商品网络，并新颖的用社交网络的影响力最大化算法从商品网络寻找影响力延展度最大的前K个商品。

## 1.1研究背景与意义

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[[[2]](#endnote-2)]提出了在购物篮发现关联规则的算法，这个算法利用条件概率的思想，从零售数据中发现潜在关联规则，这个技术后来被很多研究者所使用和优化。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）。或许最著名的关联规则是啤酒和尿布，因为家庭在周末到来时候经常一起购买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来提高购物篮分析效率。

因商品网络的直观特性，如商品网络的边能直观的反应商品之间的相关性，还能利用可视化的技术和图论的知识站在整个网络的角度进行购物篮分析。2012年Hyea Kyeong Kim等人[[[12]](#endnote-12)]构建了商品网络并分析商品网络的边权值分布，节点的中心度数与热销的关系等。通过了解近几年购物篮分析技术的发展，利用商品网络来进行购物篮分析的研究仍比较少，并且只是对商品网络的简单性质进行分析，如度数分布等。

利用商品网络进行购物篮分析因其研究仍比较少，所以存在很大的探究空间。在本文，为了更深入的对购物篮进行分析。在商品网络基础上，新颖的利用社交网络影响力最大化方法来寻找延展度最大的前K个商品。商品影响力的延展度在一定程度上反映了这些商品如果作为促销商品，它们可能在全局性的带动其它商品的销售。因此本文的研究可以辅助零售商进行商品促销的决策。

## 1.2国内外研究现状

因基于影响力最大化的购物篮分析，研究较少，目前未能找该方面的研究文献，所以国内外现状分成两部分阐述，一是基于商品网络的购物篮分析问题的研究现状。二是社交网络影响力最大化问题的研究现状。

### 1.2.1基于商品网络的购物篮分析问题的研究现状

研究者渐渐意识到商品网络所具有的直观性质，商品网络的节点可以用来反映商品，而边可以用来反映商品之间的关联性。复杂网络理论研究目前是较为成熟的理论，能够给商品网络的研究提供很多的理论支持。除此，复杂网络的可视化技术能够给商品网络分析带来很多的便利。

2012年Hyea Kyeong Kim等人[12]为了能从整体的视角上来分析网络的，构建了商品网络。商品网络的节点代表商品，节点之间的边代表商品一起被购买过。

利用了关联规则算法求得的商品之间的关联性作为商品之间的边的权重，边权重越大则代表商品之间关联性越强。在该商品网络上，Hyea Kyeong Kim等人[12]对商品网络的基本属性进行了分析，如边权分布，节点度数分布以及节点的degree centrality与商品热销程度的关系等。

2014年戚戚[[[13]](#endnote-13)]将一起购买的商品均认为存在边的关系，构建起了无权值无向图。戚戚在无向图上使用层次聚类进行了社团研究，并借助igraph挖掘社团中的核心商品。除此，为了在商品网络中给商品的重要性排序，戚戚将无向边转化为两条方向相反的有向边，边权值均取值为1，将无向图转化为有向带权图，并使用pagerank算法挖掘重要的商品。

2015年Zhu Z等人[[[14]](#endnote-14)]以周为时间窗口，寻找出用户每隔一段时间会购买的商品组合。根据这些商品组合，他们认为在同一组的商品存在边的关系，由此构建了一个无向的商品网络。为了刻画用户的消费行为，他们提出从无向的商品网络自动抽取商品小类的方法。

基于商品网络的购物篮分析的研究仍比较少，从2012年Hyea Kyeong Kim等人13对商品网络的进行简单的度数统计到2015年Zhu Z等人[15]的抽取商品小类别。基于商品网络的购物篮分析方法日趋丰富，但仍有很多的发展空间。

### 1.2.2社交网络影响力最大化的研究现状

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。

影响力传播研究有三大支柱，第一是影响力传播模型，主要是分析影响力在社交网络中传播的特点。第二是影响力传播学习，即根据机器学习等方法挖掘出影响力的传播模型和模型的参数。第三是影响力传播优化，即寻找方法来增大想要传播的影响，或者减小不想传播的影响。

影响力最大化（influence maximization）问题一个被广泛研究的影响力传播优化问题。影响力最大化是指在网络中寻找K个节点，这K个节点能使影响力的延展度最大，所以该问题是一个组合最优化的问题，在某些模型下（独立级联模型，线性阈值模型），是一个NP难的问题。

2003年，Kempe等人[[[15]](#endnote-15)]指出在独立级联模型和线性阈值模型下,影响力最大化节点集合的选取是一个NP难问题。因此Kempe等人使用贪心算法去求解该问题的近似解。2007年Leskovec J等人[[[16]](#endnote-16)]发现在独立级联模型和线性阈值模型下，影响力最大化问题呈现出很好的子模特性，在使用贪心算法每次选取影响力最大的种子节点的时候没必要更新所有节点的影响力，由此提出了给节点增加影响力更新标识的CELF算法。该算法减少了节点影响力更新的次数，在一定程度上提高了贪心算法的效率。2010年wei chen等人[30]发现独立级联模型可用随机级联模型来模拟，通过随机移除边，然后计算节点的可达集合大小来代表节点的影响力。2012年Borgs C等人[[[17]](#endnote-17)]提出了反向蒙特卡洛算法，该方法能在近似线性时间求解出影响力最大的种子集合，同时仍然有()的近似比保证。2013年，Hui Li等人[[[18]](#endnote-18)]提出了基于Conformity-Aware的贪心算法，该算法先使用社区划分的方法进行社区划分，然后对所有社区进行种子集合的选取。

除了基于独立级联模型和线性阈值模型的影响力最大化算法， 2008年Hao Ma等人[[[19]](#endnote-19)]认为热量在传播媒介上随时间的扩散与影响力的传播相似，因此创新性的提出使用热量传播模型（Heat Diffuse Model， HDM）来模拟复杂的影响力传播。该方法求节点的延展度时候，不需要进行多次的影响力传播模拟并取平均值来近似。

2014年Chen等人[[[20]](#endnote-20)]为了缩小影响力种子候选集合的大小以及避免影响力种子集中于一个社区，提出了适用于无向图的基于层次聚类的影响力最大化方法，该方法分为3步骤，第一社区发现，第二选取候选集合，第三种子集合的选取。

## 1.3本文的主要工作

基于商品网络的购物篮分析是近几年来出现的购物篮分析方法，国内外对该方法的研究仍比较少。虽然图论的研究已经很成熟了，近年来也涌现大量的复杂网络数据挖掘的方法。在商品网络中的应用，缺乏实际情况的检验，所以很多的研究仍处于探究阶段。使用社交网络影响力最大化方法在商品网络中寻找影响力最大的种子集合的研究仍比较少，本文从探究的角度进行了如下的工作：

1. 搜集基于基于商品网络的购物篮分析的方法和社交网络的影响力最大化的方法，厘清这两种研究的发展趋势和现状，了解主流方法的求解思路和特点，并选取合适的方法来进行本文的探究性实验。
2. 在选取商品影响力最大的种子集合时候，为了解决商品集中在一个社区的问题，本文采用了社团划分的方法，进行种子选取的时候可根据社区种子的选取情况，对社区的种子选取进行控制。并且可以根据社区特点帮助分析存在较强的相互影响关系的商品类别。
3. 本文选取独立级联模型和热量传播模型作为商品影响力的传播模型，并利用贪心算求解影响力最大化的问题。这两种影响力最大化方法有其各自的特点，本文将探究这两种方法在实际应用的表现。
4. 使用真实的购物篮数据分别进行基于独立级联模型和热量传播模型的影响力最大化实验，对这两种算法实验结果进行评估。

## 1.4本文组织结构

本文的正文本分全文分为六章，具体组织结构如下：

第一章：绪论。该章先用一个例子点出了购物篮分析的商业价值，然后介绍购物篮分析的背景以及本文的研究意义，接着总结本文所用理论的国内外研究现状，最后阐述本文的主要研究内容和组织结构

第二章基于商品网络的购物篮分析概述。该章介绍购物篮的定义和应用研究，然后介绍了购物篮分析中常用与分析商品关联性的apriori算法，为构建商品网络做一个知识铺垫。在此基础上介绍了商品网络常见的构建方法和商品网络的研究方法。

第三章社交网络影响力最大化概述。该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，影响力最大化问题的定义，影响力传播模型，以及求解影响力最大化问题的关键技术。

第四章影响力最大化算法在商品网络中的应用。该章介绍在本文的研究中构建商品网络的方法和伪代码设计，独立级联模型和热量传播模型在商品网络中的应用和伪代码设计，最后介适合上述两种传播模型求解影响力最大化问题的贪心算法伪代码设计。

第五章实验。该章主要介绍实验的背景和分析实验结果。

第六章总结与展望。总结本文的内容，包括研究问题的来源，研究过程，收获以及未来的计划。

**第二章 基于商品网络的购物篮分析概述**

## 2.1引言

购物篮分析是零售行业的一种重要的数据挖掘工具。这种方法可以帮助零售商挖掘隐藏在购物篮数据下的有价值信息，辅助商品分类，摆放，促销等经营决策。这一章主要介绍基于购物篮分析的定义和应用，以及商品网络的定义，常见的商品网络的构建方法，以及商品网络目前的研究方法。

## 2.2购物篮的定义和应用研究

随着经济水平的提高，因为人们的需求，超市的规模和数量都在增加。随着顾客可选的超市数量增加，超市的激烈越发激烈。超市对顾客消费行为的了解程度是超市在激烈竞争中胜出的关键因素。自从1992年Julander等人[1]提出购物篮分析方法，购物篮分析的研究便开始兴起。

购物篮是超市中供顾客装商品的一种篮子。在顾客付款的时候，收银员会将篮子中的商品信息会录入到POS机器中，一个客户一次购买的所有商品的流水号是相同的。购物篮分析是通过分析这些交易记录来发现顾客的消费行为。购物篮分析在零售行业的应用[[[21]](#endnote-21)]主要有：

1. 商品的布局设计。通过对购物篮的数据进行商品的关联性分析可以知道顾客在购买一些商品之后喜欢购买哪些商品。那么零售商可根据的顾客喜好进行商品的布局的调整。
2. 促销活动的有效性分析。从购物篮数据中分析以何种策咯进行商品的促销能带来更多的回报，包括促销商品的选择，促销价格的设置，促销时间长度设置等。
3. 顾客分类。根据顾客的消费特征对顾客进行分类管理，可以更有针对性的为顾客的推荐服务。
4. 顾客的忠诚度分析。顾客的会员卡记录者使用者的购买行为，通过时间序列上的购买分析可以了解顾客的消费喜好的变化和忠诚度的变化，那么零售商可以以此为依据制定策略留住老顾客。
5. 交叉销售。通过分析交易记录了解顾客的多种需求，在满足顾客的某一个需求的时候同时尝试去推荐相关的商品满足相关的需求。比如在满足顾客购买手机的需求下，尝试去满足顾客对屏幕保护的需求。
6. 降低库存成本。通过对超市的交易记录可以了解顾客在某些时间（季度等）对商品的销售量，那么商家可以针对性的调节进货量等。
7. 市场和趋势分析。主要是根据交易信息预测商品的需求量等，可以有针对性的调节商品的价格和库存等。

## 2.3apriori算法

apriori算法是关联规则的一个经典算法，该方法常用于购物篮分析求取商品间的相关性质。在本文中将会用到apriori 算法来构建商品网络，所以在这里提前介绍基础的知识。

### 2.3.1定义和性质

**1、定义**

apriori有两个阶段，第一阶段是求满足最小支持度的频繁项集，第二阶段是从频繁项集构造满足最小置信度的规则。该算法有四个的关键的信息，支持度（support degree）、置信度（confident degree）、最小支持度（minimize support degree）和最小置信度（minimize confident degree）。为了方便描述算法先引入，文中通用符号对应含义如表2-1所示。

表2-1 airiori算法的符号

|  |  |
| --- | --- |
| 符号 | 含义 |
| num(A) | 商品A在商品A中出现的次数 |
| T | 交易记录的数量 |
| support(A) | 商品A的支持度 |
| confident(B|A) | 购买A商品后购买B商品的可能性 |
| minSupport | 最小支持度 |
| minConfident | 最小置信度 |
|  | 表示集合大小为k的商品集合，称为k项集 |

1. 支持度（support degree）表示商品或商品组合在交易记录中出现的频率。可用公式（2.1）计算商品A的支持度：

（2.1）

1. 置信度（confident degree）表示购买一些商品后购买另外一些商品的可能性。

已经知道商品A，B，{A,B}的支持度，可用公式（2.2）求商品A后购买商品的B的置信度：

(2.2)

1. 最小支持度（minimize support degree）和最小置信度（minimize confident degree）是为了提高算法效率和约束求解结果设置的。大于等于最小支持度的项集被称为频繁项集合。

**2、重要性质**

（1）如果S是频繁项集，那么S的任何一个子集也是频繁项集。

（2）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。

### 2.3.2apriori的计算过程

**1、支持度计算**

计算项集的支持度的过程如表：

表2-2 计算支持度的过程

|  |
| --- |
| （1）计算含有一个元素的项集出现的频率，然后筛选出support()>= minSupport的项集记为’。  （2）将’两两组合得到集合大小为k的项集合，并求support()，然后筛选support()>= minSupport的项集记为。循环处理直到没有更大的频繁项集合产生。 |

**2、关联规则计算**

对于每一个频繁项集，将的集合大小为m的子集记，记。

计算项集间的关联性的过程如表：

表2-3 计算项集间的关联性的过程

|  |
| --- |
| 对每一个  （1）根据公式（xxx）计算，然后筛选出的项集记为。  （2）将两两组合构成集合大小为2的，计算 ，然后筛选出的项集记为。重复这个步骤直到n = k。 |

### 2.3.3apriori减枝思想

**1、计算项集的支持度剪枝思想**

求项集利用minSupport进行剪枝的过程如图2-1所示：如图假设现在交易记录中出现过的商品有商品A，B，C。当求项集的支持度的时候，如果support(B) < minSupport，由式2.1可知含有B商品的项集支持度都会小于最小支持度，那么这些项集的支持度都没必要去算了，即如图2-1所示红色的节点{B,AB,ABC}剪枝过程会被去掉。

图 2-1 支持度计算的剪枝



**2、计算关联规则剪枝思想**

求关联规则的剪枝过程如图2-2所示，A，B，C表示交易记录中出现的商品，在这个例子中假设{A, B, C}是频繁项集，={A,B,C}，当。当求BC->A的置信度的时候，如果，由公式（xx）可知如果， 那么，的置信度就没必要去算了，即如图2-2所示红色的节点会被剪枝去掉。

图2-2 关联规则计算时的剪枝过程



## 2.4常见构建商品网络的方法

商品网络分析是购物篮分析的一种新的方法，网络的节点表示商品，网络的边表示商品间存在关系，网络的边权表示关系的强度。商品网络可以可以直观的表示商品间的关系，并且能将购物篮分析扩展到网络的分析上。

商品网络常见的构建方法有5种：

1. 根据购物篮的数据抽取出的商品信息构成表2-4，A,B,C代表商品，如果交易号中出现了该商品则在相应位置填充1，否则填充0。

表2-4 商品网络构建的例子

|  |  |  |  |
| --- | --- | --- | --- |
| 交易号 | A | B | C |
| 1 | 1 | 1 | 1 |
| 2 | 0 | 1 | 1 |

网络1。2014年戚戚[14]等人把商品 作为网络中的节点，如果两个商品在同一条交易记录中出现，那么就认为这两个商品节点之间存在一条边。根据表2-4的信息可构建的无向商品网络如图2-3所示：

图2-3 商品网络1



1. 网络2。为了使用pagerank算法对商品的重要性进行打分，2014年戚戚[14]等人将商品网络的无向边转化方向相反的有向边，边的权值均置为1。以图2-3为例可转化成图2-4：

图2-4 商品网络2



（3）网络3。2012年Hyea Kyeong Kim等人[13]以商品作为节点，如果两个商品在同一天被购买，那么就认为这两个商品节点之间存在一条边。然后将每天一起购买的商品当作一个购物篮，利用apriori算法求取商品间的关联性作为商品间边的权值。

（4）网络4。2012年Hyea Kyeong Kim等人[13]使用的另外一种构建购物篮的方法是，将一个用户在一个时间段购买的商品当作是一起购买，然后以购买的商品作为为节点，然后一个购买则认为存在一条无向边，构建起了无向商品网络。

（5）网络5。2015年Zhu Z等人[15]对喜欢一起购买购买的商品定义为顾客每个时间周期（如每周等）会一起购买的商品。按这个要求从原始交易记录进行筛选出交易记录的子集。在这个新数据集上，把商品看作节点，把一起购买的商品建立边，构建起了一个无向商品网络。这个方法与（4）的区别是（4）是以（客户，时间段）作为分组依据，而这里不仅仅是分组还有周期购买的约束。

假设原始交易记录如表2-5，A,B,C代表商品，如果交易号中出现了该商品则在相应位置填充1，否则填充0。

表2-5 交易记录例子

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 交易号 | 顾客 | 时间 | A | B | C |
| 1 | 1 | 20150103 | 1 | 1 | 1 |
| 2 | 2 | 20150103 | 0 | 1 | 1 |
| 3 | 1 | 20150108 | 1 | 0 | 1 |
| 4 | 2 | 20150108 | 1 | 1 | 1 |

以及7天为时间周期，如表顾客1在2015年01月03日购买了商品A,B,C后，7天内又购买了A,C，重复购买的商品是A,C。同理顾客2重复购买的是B,C。筛选出新的交易信息如表2-6所示，根据表2-6构建的网络如图2-5：

表2-6 购买记录

|  |  |  |
| --- | --- | --- |
| A | B | C |
| 1 | 1 | 0 |
| 0 | 1 | 1 |

图2-5 商品网络3



## 2.5商品网络分析常见方法

基于商品网络的购物篮分析，常见的分析有两类，一是社团结构发现，二是重要的商品的发现。这里基于2.4节所述的商品网络展开介绍方法，因为不是本文的研究重点，所以只介绍基本的思想。

### 2.5.1社团发现

1、层次聚类[14]。

一种非重叠社区发现算法，层次聚类的过程是先将每个节点看作是一个社区，然后选取距离最近（或相似度最大）的两个社区进行合并，当整个网络的模块值最大停止合并。戚戚[14]使用该方法对2.4节的网络1进行社区划分,这些社区可以给商品分组布局提供依据

2、K-团发现[15]。

这是一个重叠社区发现的方法，可以寻找出网络中的所有K个节点的完全图。Zhu Z等人[15]使用该方法对2.4节的网络5进行了K-团发现。在对这些K-团分析的是否发现，K-团中的商品存在互补的关系。如报警器和烟雾探测器是家庭安全设备，它们在同一团中。

### 2.4.2重要商品的发现

1、pagerank算法[14]。

PageRank算法的作用是对按重要性对网页排序。pagerank算法的核心思想是一个网页的重要性由两点来决定，一是越多网页的链接指向该网页，那么该网页越重要，即pagerank值越大。二是该页面越重要，那么的链接指向的页面也就越重要。

戚戚[14]在网络2中使用pagerank算法按重要性对商品进行排序，发现排在首位的是时尚杂志。结果具有一定的合理性，顾客休息或等待同伴时候喜欢观看杂志，并且时尚杂志比起其他专业书籍在更多年龄段的人群受欢迎。

2、度中心性（Degree Centrality）[13]。

度中心性是节点中心性最直接的度量，一个节点的度数越大，度中心性也会越大，反映了该商品节点越重要。

2012年Hyea Kyeong Kim等人[13]发现网络3度中心性大的商品是一些购买较频繁的日常必需品，而网络4中度中心性大的商品是普通家庭的非经常否买的生活必须品。

## 2.6本章小结

本章介绍了购物篮的定义和购物篮分析的常见应用，然后介绍了关联规则分析算法中的apriori，包括该算法的定义，核心思想以及执行过程，为下文的商品网络的构建做了知识铺垫。接着以戚戚[14]、Hyea Kyeong Kim等人[13]和Zhu Z等人[15]的商品网络研究为例子的介绍常见商品网络构建方法以及介绍常见的商品网络分析方法，如社区发现和重要商品发现等。

# 第三章 社交网络影响力最大化概述

## 3.1引言

影响力最大化问题是社交网络影响力传播研究的一个热点问题。影响力最大化是一个组合最优化问题，求解有3个步骤，1、节点的间相互影响强度的度量，2、选取合适的影响力传播模型，3、选择影响力最大化求解算法。

因为本文探究社交网络影响力最大化算法在商品网络中的应用，所以在本章以求解影响力最大化问题的步骤为顺序，预先介绍基础的理论知识，包括社交网络的概述，影响力最大化问题的定义，影响力传播建模，影响力最大化问题研究的关键技术（包括影响力最大化问题求解算法概述和本文求解影响力最大化问题用到的有向图层次聚类）。

## 3.2预备知识

### 3.2.1社交网络概述

社会生活中，个体之间会在某种社会关系，信息会通过这些关系进行传播。为了方便的对社交关系进行研究，常用的方法是对社会关系进行网络建模。网络的节点集}表示社会中的个体的集合，网络的边的集合}表示节点间的关系，表示节点i对节点j存在某种关系。

在社交网络中，关系存在很多种，相应的网络也有很多种。在本文中将讨论个体间的影响关系，因为个体间的相互影响关系是不同，所以使用一个有向带权的网络来表示。网络可简单的定义为G(V,E,W)，其中V是节点集合，E是边的集合，而W是边上的权值。表示社会的一个个体，表示个体可影响个体，影响大小用W来表示。与存在影响关系的个体我们称为邻居，用N()来表示。影响节点的节点称为节点入邻居，用表示，入邻居对影响力总和为。相反，被vi节点影响的节点称为节点用，节点节点邻居节点总的影响为。

为了方便进行相关理论的描述，这里先定义通用的符号如表3-1所示：

表3-1 社交网络通用符号的含义

|  |  |
| --- | --- |
| 符号 | 含义 |
| G(V,E,W) | 社交网络图 |
| V | G(V,E,W)中的个体集合 |
| n | G(V,E,W)的个体数量 |
|  | G(V,E,W)中第i个节点 |
|  | 对出邻居的影响的影响总和（带权出度） |
|  | 入的邻居对的影响总和（带权入度） |
|  | 节点对出邻居的影响 |
| E | G(V,E,W)中的个体间关系集合 |
| M | G(V,E,W)中W中边权的和 |
|  | 编号为i的边 |
|  | 以为起点，为终点的边 |
| S | 初始用户集合，或称：种子节点集 |
| ­seed\_num | S中所能包含的最大的节点数量 |
|  | 的邻居集合 |
|  | 节点出邻居集合 |
|  | 节点的入邻居集合 |

### 3.2.2影响力最大化问题的定义

在给定的网络G=(V,E,W)，影响力传播模型和所需的参数的情况下，每个节点的影响力，即能够成功影响多少的节点是可以通过计算得到的。我们先假设()是一个计算影响力的函数，那么集合S的影响力可用(S)来表示。

影响力最大化问题是从网络中选择seed\_num个种子节点的集合，这个种子集合的比其他seed\_num个节点所组合的集合影响力大。我们可用公式xxx来表示：

。 (3.1)

## 3.3影响力传播建模

在影响力传播模型中最常用的传播模型是线性阈值模型和独立级联模型，热量传播模型则跳出了这个两个模型的思维。在该节中将讲述这三个模型的思想，而伪代码只列出与本文研究有关的独立级联模型和热量传播模型伪代码。

### 3.3.1线性阈值模型

2003年，Kempe、Kleinberg等人[16]将线性阈值模型引入到计算机科学中来对影响力传播进行建模。线性阈值模型中用户有两种状态，活跃状态和非活跃状态。活跃状态可以去影响非活跃状态的邻居，并且用户从非活跃变到活跃状态，就会一直保持活跃状态。线性阈值模型的特点是该模型认为影响是可以积累的，且站在被影响者的角度的来考虑被影响程度。线性阈值模型有一个约束，对于，满足不等式(3.2)。

(3.2)

线性阈值模型中每一个节点会分配一直激活阈值，越大说明节点越难被激活。让从最初种子集合到第t轮已经被激活的节点，初始的节点集合，。在第t+1轮不活跃节点的活跃入邻居是的，如果能在该t+1轮被激活需满足条件(3.3)：

(3.3)

线性阈值模型的过程是每次用活跃节点去激活新节点得到活跃节点，直到不再有新的节点被激活为止。

### 3.3.2独立级联模型

2003年，Kempe、Kleinberg等人[16]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[[[22]](#endnote-22)]改进了独立级联模型，通过概率1-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,W)，V是节点集，E是边集合，W是边的权值集合，表示从节点到节点的边的权重。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V，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(3.4)来表示：

(3.4)

按照上面的迭代步骤，直到为空。

节点只会尝试一次通过激活节点，因此我们可以换个角度，根据2009年W. Chen等人[23]的思路，可以预先通过一个随机函数判断节点能否通过边成功影响。如果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。求取强连通组件C,其中。对于一个节点,在的可达节点均在，此时。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，就将的影响力记作0，否则为。独立级联模型求取节点影响力的时候往往需要进行多次模拟取节点的平均影响力。

### 3.3.3热量传播模型

为了捕捉市场临时的信息传播，2008年Hao Ma等人[19]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,W),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[19]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,经过时间i节点的温度如(3.5)所示，该公式可化简为(3.8)。

(3.5)

公式(3.5)可化简为(3.6)

(3.6)

其中 (3.7)

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(3.8)：

(3.8)

求积分后可得(3.9)：

(3.9)

### 3.3.4其他传播模型

除了上述的三种模型，还有很多的影响力模型，这里只简单介绍两个其他的影响力模型。

（1）传染病模型(epidemic model)

传染病模型最初用作对病毒在人群中的传播进行建模，现在已经扩展到对信息传播研究上了。在该模型节点可能存在三种状态，易感S(susceptible)，感染I(infected)，治愈R(susceptible)。该模型根据状态根据节点状态的变化可区分为三种模型，SI模型，SIS模型，SIR模型。SI模型中节点可由易感变成感染状态并一直保持感染状态。SIS模型中允许节点在易感状态和感染状态来回变化。SIR模型模拟了节点的免疫过程，节点由易感状态可切换到感染状态，当节点从感染切换到治愈状态后便不会再被感染。

（2）选举模型(voter model)

选举模型的原本用在统计物理中，现在被扩展到社交网络影响力的传播上了。该模型影响过程是，一个节点的状态由的状态来决定，通过随机从中选取一个状态作为的当前状态。这种变化模拟了在选举过程中意见在个体间的传播，一个节点通过与其他节点的交流并接受了其中一个节点的观点。这个模型常用于了解在某一个时刻网络中节点的状态分布和相关性质。

## 3.4影响力最大化关键技术

这一节主要讲述影响力最大化求解算法概述和在选取影响力最大的商品集合中会用到的有向图快速层次聚类算法。

### 3.4.1影响力最大化问题求解算法概述

影响力最大化问题是一个组合最优化的问题，很难找到全局最优解，所以很多研究者基于启发式算法和贪心算法来求解影响力最大化问题。

1. **基于贪心算法**

贪心算法不从整体考虑问题的最优解，它的特点是通过不断求取当前的最优解来取得局部最优解。

2003年，Kempe等人[16]指出在独立级联模型和线性阈值模型下,影响力最大化节点集合的选取是一个NP难问题，因此Kempe等人使用贪心算法BasicGreedy每次选择一个使影响增益最大的节点加入到种子集合S中，最终求得含K个节点的影响力最大化种子集。在求解过程中，种子集合S的影响力增益是非负且单调递减的，即满足子模特性，又因，所以该方法所求的近似解至少是最优解的(1-1/),其中e是常数。

2008年Hao Ma等人[[[23]](#endnote-23)]认为热量在传播媒介上随时间的扩散与影响力的传播相似，因此创新性的使用热量传播模型（Heat Diffuse Model， HDM）来模拟复杂的影响力传播。在该模型中，Hao Ma等人求第I轮种子集合时候，使用第I-1次求取的种子集S与一起求取影响力，然后能与S一起影响力最大的加入到种子集S中。

2007年Leskovec J等人[17]发现BasicGreedy贪心在独立级联模型和线性阈值模型求解影响力最大化问题过程中，影响力增益呈现出很好的子模特性，该性质下选取影响增益最大的种子节点时候，可以不更新所有节点的影响力增益，因此在BasicGreedy基础上提出了给节点增加影响力更新标识的CELF算法。该算法的求解精度与BasicGreedy相同，而求解速度正常情况比BasicGreedy算法快700倍。

2013年，Hui Li等人[[[24]](#endnote-24)]认为在社会活动中人们会表现出从众意识（Conformity Aware）。在对从众行为进行影响力传播建模后，Hui Li等人证明社区划分后，原始的贪心算法在从众意识影响力传播模型上同样满足子模特性，且贪心解依然是最优解的(1-1/)。所以Hui Li等人提出一种基于社区划分的贪心算法，该算法分为两个阶段第一阶段使用类似于CELF算法对每个社区进行求解影响力增益最大的节点，然后不断的从所有社区的影响力增益最大的节点作为种子节点，并只在被选取种子节点的社区中，更新节点影响力增益，重复步骤直到选取K节点的影响力种子集合。

1. **基于启发式算法**

启发式算法利通过分析节点的特性，依据经验选取更符合种子节点特性的节点来构造种子集合。

2011年jiang等人[[[25]](#endnote-25)]使用模拟退火算法来求解影响力最大化问题，首先设置一个初始温度T,临界温度并随机选择一个初始的K节点种子集合S，然后随机的从邻居节点中选择一个节点来替换S中的一个节点得到新的种子集合，如果影响增益，则让S=，否则再通过一个收敛的随机函数（与T有关）来决定是否仍用替换S作为当前的种子集合。重复该步骤并让T=T-，直到T<结束程序。该算法的时间复杂度是O(IRM)，其中I是迭代的次数，M是边数，R是计算影响力增益的时间复杂度。

2012年Wang C等人[[[26]](#endnote-26)]使用局部的影响力来近似作为节点在图中的影响力的PIMA算法。该算法首先通过dijkstra算法构建节点间最大影响力路径(maximization influence path, mip)，然后忽略概率小于阈值的mip,这样可以高效的得到节点的局部影响力。这种方法大大减小的求解影响力的时间花费，PIMA算法比当时经过优化的蒙特卡洛算法快1000倍。

2014年Chen等人[21]为了缩小影响力种子候选集合的大小以及避免影响力种子集中于一个社区，提出了基于热量传播模型的影响力最大化方法。该方法根据社区的种子选取情况，节点的度数和在网络中的位置等特征来选取种子节点集合，方法分为3步骤，第一社区发现，第二选取候选集合，第三通过不断试探性的替换种子集合的节点优化种子集合。

### 3.4.2有向图快速层次聚类算法

本文求影响力最大化商品集合过程中会用到有向图的快速层次聚类，原因有两个，第一社区划分可以防止选取的种子集合中的商品集中于一个社区，第二可社区的划分能够反映哪些商品的类别存在较强的影响关系。

在本文使用的聚类算法是有向图的快速层次聚类，因为聚类不是本文的重点，所以只介绍在算法实现过程中用到的公式(3.18)和(3.20)，以及介绍算法的流程。

给定一个有向图G(V,E,W)，节点的带权出度用表示,带权入度则用来表示，根据2008年E. A. Leicht1等人[[[27]](#endnote-27)]的文献可知有向图的模块值计算公式为(3.10)。

(3.10)

其中

(3.11)

(3.12)

(3.13)

当节点v和节点w在同一个社区的时候，为1否则为0。

引入(3.14)，其中x和y是社区的编号。当x=y的时候。

(3.14)

社区x和社区y连接的所有边的边权占图总的边权和比例用表示，则有：

(3.15)

社区x中节点入度和占图总的边权和比例用为表示，则有：

(3.16)

社区x中节点的出度和占图总的边权和比例用表示，则有

(3.17)

有向图的模块值计算(3.10)可根据(3.14)~(3.17)简化为：

(3.18)

由有向图公式模块值计算公式(3.18)可得一个孤立点加入社区后的模块值的增益的计算公式(3.19)：

(3.19)

(3.19)可简化为(3.20)

(3.20)

表示当前社区C内的边权和，

是节点x连接C内部节点的边的权值和，

是终点是社区C的节点的边的权值和，

是边的起点是社区C的节点的边的权值和。

仿照2003年Blondel等人[[[28]](#endnote-28)]的无向图快速层次聚类算法，设计有向图的快速层次聚类算法fhc的伪代码如表3-2所示。步骤1~2把每个节点当作一个社区，步骤3计算当前社区划分的模块值。步骤5~16将每个节点加入其中一个邻居社区，加入后带来模块值的增益最大。步骤4~19 不断调整节点到不同社区，直到模块值的增益变化小于值。步骤20~24根据每个社区的节点集合，从原图抽取出相应的子图。步骤25返回子图列表。

表3-2 fhc算法伪代码

|  |
| --- |
| Algorithm fhc  input：G,=0.001  output：subg  1.for each len(G.V)  2. cluster[i] #将每个节点看作是一个社区  3.Q2cluster划分状态下，用公式xxx计算模块Q值  4.do  5. for each in G  6. max\_irc -9999  7. best\_cid = 0 |

续表3-2

|  |
| --- |
| 8. i\_id所属社区的编号  9. remove from cluster[i\_id]  10. nei\_cluster的邻居社区编号  11. for each cid in nei\_cluster  12. irc用公式xxx计算节点加入cid社区后模块值变化值  13. if ircmax\_irc  14. max\_ircirc  15. best\_cidcid  16. cluster[best\_cid].append()  17. QQ2  18. Q2  19.while Q2 - Q >  20.subg empty list  21.for each v\_set in cluster  22. if len(v\_set)>0  23. new\_g根据v\_set节点集合从G中抽取子图  24. subg.append(new\_g)  25.return subg |

## 3.5 本章小结

本章先介绍了社交网络的定义，以及给出阐述社交网络理论会用的一些符号及其含义。在这个基础介绍社交网络影响力最大化化问题的定义，然后介绍线性阈值模型，独立级联模型，热量传播模型等传播模型。最后介绍求解影响力最大化问题的关键技术，包括求解影响力最大化问题的算法概述和文中会用到的快速层次聚类算法及其实现。

**第四章 影响力最大化算法在商品网络中的应用**

## 4.1引言

本章介绍如何将社交网络影响力最大化算法应用到商品网络中，涉及的内容有商品影响力是如何表现的，独立级联模型和热量传播模型如何对商品影响力进行描述的，构建商品网络的伪代码实现，在独立及级联模型和热量传播模型中设计影响力最大化算法求影响力最大的商品集合。

## 4.2商品网络的构建

在超市的中，顾客往往会把想买的东西一起放在购物篮中，收银员通过扫描条形码将购物篮中的商品录入到电脑中，一次购买的商品有一个共同的交易号。因此同一交易号的商品组合在一定程度上反映了顾客喜欢同时购买这些商品，使用apriori关联规则算法我们可以得到顾客购买了一种商品购买另外一种商品的可能性，即顾客购买一个商品后，该商品会激起该购买另一件商品的欲望。我们将apriori算法求得的这种商品关联性当作是商品之间的影响关系，关联性越强，商品间相互影响的关系也越强。

实验中我们将使用一个有向带权网络G(V,E,W)来表示商品的影响关系网络，其中表示商品网络中的一件商品，表示商品节点会影响到商品节点，商品的影响关系强度是。

根据2.3节apriori算法，求两两商品间的影响强度的由5部分组成，一生成1项集，二根据k项集生成k+1项集，三求项集的支持度，四根据项集的支持度求两两商品间的关联性。为方便理解伪代码，先给重要变量的含义如表(4-1)。

表4-1 独立级联模型算法伪代码中重要的变量的意义

|  |  |
| --- | --- |
| dataset | dataSet是一个二维列表，每一行代表一次交易的商品组合 |
| Lk | Lk是项集的列表，里面的每个项集的元素个数是相同的 |
| L2 | 2项集列表 |
| minConfident | 最小置信度 |
| minSupport | 最小支持度 |
| supportData | 每个项集的支持度 |

1. 构造1项集的伪代码如表4-2。

dataSet是一个二维列表，每一行代表一次交易的商品组合，步骤2~5是遍历所有商品，并且构建1项集的列表C1，列表中的项集是不重复的。

表4-1 createL1算法伪代码

|  |
| --- |
| Algorithm: createL1  Input: dataSet  output: L1  1.L1←empty list  2.for each transaction in dataset  3. for each item in transaction  4. if [item] not in C1  5. insert [item] to C1  6.return C1 |

1. 通过k项集构造k+1项集的伪代码如表(4-3)。

程序输入Lk项集列表，第1~2步骤是初始化标量，Ln是存储k+1项集的列表，lenLk是k项集的个数。第4~11步骤的作用是如果k项集的前k-1个元素相同，那么这两个项集就可以构建出k+1项集。如果构建的k+1项集不在k+1项集列表Ln中，则添加k+1项集到Ln中。

图4-3 aprioriLn算法伪代码

|  |
| --- |
| Algorithm: aprioriLn  input: 项大小为K的项集Lk  output：项大小为K+1的项集Ln  1.Ln←empty list  2.lenLk←len(Lk) # Lk项集的k项集的个数  3.n←len(Lk[0]) + 1 # Lk是k项集列表，项的元素个数加1赋值给n  4.for each i in [0,1...lenLk-1]  5. for each j in [i+1,i+2...lenLk-1]  6. Li ← sorted( list(Lk[i]) ) # 第i个项集的元素升序排列  7. Lj ← sorted( list(Lk[j]) ) # 第j个项集的元素升序排列  8. C1 ← Li[:n-2] # 将Li的前n-2个元素赋值给C1  9. C2 ← Lj[:n-2] # 将Li的前n-2个元素赋值给C1  10. if C1 equal to C2  11. if Lk[i]∪Lk[j] not in Ln  12. add Lk[i]∪Lk[j] to Ln  13.return Ln |

1. 计算项集的出现频率伪代码如表4-4。

步骤2~8步是统计所有1项集的出现次数，第11~12步是计算1项集的支持度，并筛选出大于最小支持度的项集。因为是计算两两商品之间的关联性，所以只需要求项集1的支持度。

表4-4 scanD算法伪代码

|  |
| --- |
| Algorithm: scanD(dataset, Lk, minSupport)  input：交易记录集dataSet，项集Lk，最小支持度minSupport  output：满足约束条件的项集ret\_Lk，项集的中项出现频率supportData   1. ItemsetCnt←empty dictionary |

续表4-4

|  |
| --- |
| 2.for each tid in dataset  3. for each itemset in Lk  4. if itemset is not a subset of tid  5. if itemset is not one of key in ItemsetCnt  6. ItemsetCnt[ itemset ]←1  7. else  8. ItemsetCnt[ itemset ]←ItemsetCnt[ itemset ]+ 1  9.numItems←float(length of dataSet)  10.ret\_Lk←empty list  11.supportData←empty dictionary  12.for each key in ItemsetCnt  13. support←ItemsetCnt[key] /numItems  14. if support >= minSupport  15. insert key to ret\_Lk  16.supportData[key]←support  17.return ret\_Lk，supportData |

1. 计算任意两个商品小类间的关联规则伪代码如表4-5所示。

算法generateRules的步骤1~2是初始化变量。步骤8中的itemset-item是集合itemset与集合item的差集。第3~7步骤是求频繁项集的子集与差集的关联性，因为是L2中的每个元素是2项集，所以只需求得的是两两商品间的支持度。

表4-5 generateRules算法伪代码

|  |
| --- |
| Algorithm: generateRules(L2, supportData, minConfident)  input：项集列表L，输入每一项的支持度supportData，最小置信度minConf  output：关联规则ruleList  1.ruleList←empty list  2.len\_Lk←length of L2  3.for each itemset in L2  4. for item in itemset  5. if item not in supportData.keys  or itemset-item not in supportData.keys  continue  6. confidence←supportData[item]/supportData[itemset-item]  7. if confidence >= minConfident  8. insert (itemset-item, item, confidence) in ruleList  9.return ruleList |

## 4.2独立级联模型在商品网络的应用

独立级联模型是社交网络影响力传播模型的一种经典模型，该模型的特点是认为个体间的影响是相互独立，且只使用l-1轮激活的节点尝试一次激活l轮未被激活的节点。每个节点的激活可能性是(3.4)。

在逛超市的过程中，购物者看到将一件商品加入购物篮子，则认为该商品处于激活状态。刚放入购物篮中的商品会让购物者有可能性p想去购买其他相关的商品（该商品在商品网络中的邻居商品)，而apriori关联规则算法求得的商品关联性则反映了购买欲p。这样的购买行为与独立级联模型存在一定的相似性，消费者购买一件商品后考虑相关商品的过程与独立级联模型的用刚激活的节点尝试激活新节点类似。而商品的影响力可以定义为，如果消费者在超市购买第一件商品后，在最终购买的商品作为该商品的影响力，商品影响力不代表实际会购买量，只是作为该商品带动销售量的能力。

在进行商品促销决策时候，往往会想选择何种商品进行促销能够带来更多的销量。本文基于这样的应用场景采用基于独立级联模型的影响力最大化算法尝试寻找出影响力最大的商品组合，为零售商家选择促销商品提供信息参考。

基于3.3.2节所述2009年W. Chen等人[23]的独立级联模型模拟方法，求商品影响力增益的伪代码如表4-6。ICM算法第1~2步骤是初始化，5~7步骤是对独立级联模型的模拟，节点不能通过边成功影响,则移除边。在新的图中，节点可达的节点均可认为是被激活的节点，所以在求影响力前用步骤8求出图的强连通组件。9~14步骤是用于求节点的影响力增益，步骤10中的 就是防止与S节点重叠影响的节点算作节点的影响增益。15~16是将节点平均的影响力增益作为节点的最终影响力。最后返回所有节点的影响增益f。

表4-6 ICM算法伪代码

|  |
| --- |
| Algorithm ICM(G)  输入：有向图G  输出：影响力f  1.N ←2000 # 模拟影响力传播次数  2.f←zeros array # f初始是一个0数组  3.for r in [1,2,...,N]  4. ←复制G  5. for each 中的边  6. if  7. 从移除  8. C←求的强连通组件  9. For each in V\S |

续表4-6

|  |
| --- |
| 10. if  11.  13. end if  14. end for  15.for in f  16.  17.输出f |

## 4.3基于热量传播模型的影响力最大化算法设计

热量在介质上的复杂传播过程，高温物体与低温物体接触过程中，高温物体的温度会逐渐减低，而低温物体的温度会逐渐的升高，并且接触时间的增加温度会渐渐趋于相同。为了对社交网络影响力传播建模，研究者对热量传播进行了简化和调整。在社交网络的热量传播模型下，每个个体v会按相同的比例的将影响平均传播给他的邻居，个体v的邻居越多那么每个邻居受到个体v影响就越弱，同时个体v受到邻居的影响是可以累积的。在不同时刻，活跃的节点数也有所不同，且随着影响时间的增加，每个个体的温度会趋于一个稳定值。2008年Hao Ma等人[19]使用公式(3.9)来计算带权有向网络中节点的被影响程度。

热量传播模型与独立级联的不同之处是，热量传播模型不仅仅是使用上次刚激活的节点来激活新节点，它考虑的是所有节点间的影响力在彼此之间的传播，一个节点的活跃程度可用温度来衡量，活跃的节点会让其他节点也变得活跃，而其他商品活跃度提高的同时自身的活跃度可能会有一定的降低。这可以用来模拟购物者在看到一些促销品的时候，突然想购买便顺手将其放入购物篮子中，然后又想购买相关的一些商品，在商品对比过程中，对相关商品的购买欲望可能会增加或降低，然后又把想购买的商品放入购物篮中。促销过程中促销品可以看做是热源，热量传播时间可看做是购物者购买促销品后考虑和比对其他商品的时间，最终购买的商品看做是这个促销品带来的影响。

为方便叙述先给出变量的含义表，如表4-2。

表4-2 热量传播模型算法伪代码重要变量的含义

|  |  |
| --- | --- |
| S | 种子集合 |
| h0 | 初始温度 |
| a | 传导率 |
| t | 传播时间 |
|  | 激活阈值 |
|  | 节点影响力列表 |
|  | 节点影响增益列表 |
|  | 节点的影响增益 |

基于3.3.3所述2008年Hao Ma等人[19]的热量传播模型思路，使用热量传播模型求节点影响力的伪代码如表4-7所示。算法HeatDiffusion中，第1~4步骤是初始化，步骤5求解影响力传播矩阵，计算,步骤6求解式3.9得到每个节点经过时间t后的变化，第7~9步是计算当前被激活的节点数量。最后返回种子集合的影响力。

表4-7 HeatDiffusion算法为伪代码

|  |
| --- |
| Algorithm HeatDiffusion  input：G, S, h0, a, t,  output：  1.←0 # 初始化影响力为0  2.f←zeros array # f初始是一个0数组  3.for each in S  4. ←h0 # 初始化种子节点的温度为h0  5.求解公式3.9得到H  6.计算(t) =  7.for each in G.V  8. if  9. += 1  10.return |

使用热量传播模型求节点影响力增益的伪代码如表4-8所示。算法HeatDiffusionGain中第1步是求当前种子集合的影响力，步骤2~3初始化每个节点影响力增益为0，步骤4~9是计算每个节点加入种子集会带来的影响力增益。最后返回每个节点的影响增益f。

表4-8 HeatDiffusionGain算法伪代码

|  |
| --- |
| Algorithm HeatDiffusionGain(G, S, h0, a, t, ):  input：图G，种子集合S，初始温度h0，传导率a，时间t, 激活阈值  output：影响力增益  1.cur\_incluence←HeatDiffusion(G, S, h0, a, t, )  2.num = g.number\_of\_node  3.f←zeros(num) # 0 数组，元素个数是num个  4.for each in G.V  5. if node in S  6. continue  7. new\_seed\_set =  8. influence = HeatDiffusion(g, new\_seed\_set, h0, a, t, )  9. = cur\_incluence – influence  10.return f |

## 4.4基于贪心思想的影响力最大化算法

求影响力最大化的商品集合是一个组合最优化问题，本文采用效仿2013年hui li等人[18]的CINEMA算法的贪心思想来进行影响力最大化种子集合的选取。因为本文要做的是应用探究实验，算法在时间复杂度上要求不那么严格，所以使用省略了基于子模性质的延迟更新影响的步骤。

在此基础上，为了使种子节点能更均匀的分布在各个社区，本文会给每个社区进行打分（score），当从两个社区的选取的影响力增益最大的节点，如果影响增益差别不大，则选取分数较高的那个社区的节点。

本文设计的计分函数如（4.1）所示，其中num\_node是社区的节点个数，seed\_num是当前选取的种子数量：

(4.1)

这个贪心算法的框架就能适用于独立级联模型和热量传播模型，所以不按模型分开赘述，算法的流程如图3-9，其中处理1是指将图划分成多个子图，记作subg。处理2是指从子图中寻找影响增益最大的节点更新mag列表，当mag相应位置的节点被选为种子节点，处理3是指从所有社区中选择影响力增益最大的节点，如果节点影响增益相等，则选取所在的社区score值最大的那个。

图4-9 贪心算法流程



为方便理解伪代码，先给出重要变量及其含义，如表4-9所示。

表4-9 贪心算法伪代码重要变量的含义

|  |  |
| --- | --- |
| k | 种子节点数 |
| seed\_set | 种子集合 |

续表4-9

|  |  |
| --- | --- |
| dfm\_gain | HeatDiffusionGain或ICM算法 |
| mag | 是一个列表，mag[i][0]存放i社区当前影响增益最大的节点, mag[i][1]存放该节点的的增益 |
| subg | 是一个社区列表，subg[i]的当前影响增益存放在mag[i][0] |
| fhc | 是3.4.2节介绍的有向图社区划分算法 |

根据表4-10设计选择k节点影响力最大化种子集合的算法（get\_seed\_set）。get\_seed\_set算法步骤1是使用3.4.3节fnc算法进行社区划分，第2~5步是构建mag列表，初始化各个社区种子集合为空。步骤6的循环指的是每次循环获得一个种子节点。步骤7~9是从所有社区中选择影响增益最大的种子。步骤10~11是从获取了最大影响增益的种子节点的社区重新选取一个影响增益最大的种子来更新mag列表。最后返回k节点的影响力最大种子集合。

表4-10 get\_seed\_set算法伪代码

|  |
| --- |
| algorithm: get\_seed\_set (G, k):  input: G, k  output: seed\_set  1. subgfhc(g) # 使用3.4.2的有向图社区划分  2. maginit\_mag(subg) # 初始化mag列表  3. seed\_setempty list # 空种子列表  4. for each i in [0,1,2…,len(subg)]  5. seed\_set.append([]) # 初始化每个社区种子集合为空  6. for i in [0,1,…,k-1]  7. idxarg\_max\_mag(mag, seed\_set, subg)  8. seed\_set[idx][0].append(mag[idx][0]) # mag[idx][0]是影响增益最大种子  9. seed\_set[idx][1]seed\_set[idx][1] +mag[idx][1] # mag[idx][1] # 影响增益  10. subgthsubg[idx]  11. mag[idx]get\_node\_with\_max\_gain (subgth, seed\_set[idx][0])  12. return seed\_set |

从一个社区中选择一个影响力最大的节点的伪代码如表4-11。步骤1是计算传入的图中每个节点的影响增益，步骤2~3是获得和返回影响增益最大的节点和该节点的增益。

表4-11 get\_node\_with\_max\_gain算法伪代码

|  |
| --- |
| algorithm: get\_node\_with\_max\_gain  input: G, seed\_set, dfm\_gain  output: 节点node,收益gain  1. f = dfm\_gain (g, seed\_set)  2. node,gainthe node with max gain from f  3. return node,gain |

初始化mag列表的伪代码如表4-12。初始化mag列表，mag[i]是一个元组，元组内容是[社区i的影响增益最大的节点，该节点的增益]。步骤1构建mag列表，步骤2~4构建mag列表，步骤5返回mag列表。

表4-12 init\_mag算法伪代码

|  |
| --- |
| algorithm init\_mag  input: subg  output: mag  1. mag = []  2. for each subgth in subg:  3. node, gainget\_node\_with\_max\_gain(subgth, [])  4. mag.append([node, gain])  5. return mag |

从mag获取一个影响力增益最大的节点所在mag列表的索引的伪代码如表4-13。步骤1是构造新的列表记为new\_mag，new\_mag存放的每个元素会是一个元祖，元组内容是(节点在mag的位置， 取整后的种子节点影响增益，社区的score值）。步骤3~8是新建new\_mag。步骤9~10是获取并返回影响增益最大的节点在mag 的位置索引。

图4-1表 arg\_max\_mag算法

|  |
| --- |
| Algorithm arg\_max\_mag  input: mag, seed\_set, subg  output: idx  1.new\_magempty list  2.for idx in [0,1,2,…, len(mag)  3. nodemag[idx] # node[0]是节点，node[1]是节点的影响力  4. numsubg[idx].number\_of\_node # 社区节点个数  6. score = num / (len(seed\_set[idx]) + 1) # 使用公式xxx计算score值  7. influ = int(node [0]+0.3) # 取整后的种子节点影响增益  8. new\_mag.append([idx, influ, score])  9.idx, influ, scoremax(new\_mag, key=lambda x : [x[1],x[2]])  10.return idx |

## 4.5本章小结

本章首先对商品网络的构建进行伪代码设计，因为构建商品网络只需求商品两两之间的关联性，所以简化了apriori算法，求出1项集和2项集的支持度，然后求出商品间的关联即可。本章还分析了独立级联模型和热量传播模型在商品网络中的适用性及设计基于这两个模型的影响增益算法伪代码，最后根据2013年hui li等人[18]的贪心算法设计了本文求解影响力最大化的贪心算法的伪代码。

# 第五章 实验

## 5.1引言

该章主要探究上一章的设计的基于独立模型和基于热量传播模的影响力最大化算法在实际数据集的变现。在该章会介绍实验的背景，实验数据集，实验结果评估方式，然后展示和分析实验结果。

## 5.2实验背景

### 5.3.1实验环境

本文所有实验均是在一台计算机上完成，计算机主要配置如下：

处理器：Intel(R) Celeron(R) CPU 1000m @ 1.80GHz 1.80GHz

安装内存：4.00 GB(3.85 GB 可用)

硬盘内存：1000G；

操作系统：Windows 8.1，64位系统，基于x64的处理器

实验所用的算法均采用python语言在ipython notebook工具下实现。

### 5.3.2实验数据集

本文使用的数据集来自一个真实的超市pos机记录的交易记录，取了其中一个月作为本实验的数据集。该数据集含有的字段信息如表5-1所示。

表5-1 实际数据集的含有的字段

|  |  |  |  |
| --- | --- | --- | --- |
| 日期 | 部门编号 | 销售数量 | 经营小类 |
| 销售流水号 | 商品小类 | 销售额 | 经营小类名称 |
| 商品名称 | 商品小类名称 | 销售时间 | 品牌 |
| 项次 | 商品中类 | 品牌名称 | 毛利 |
| 商品ID | 商品中类名称 | 商品种数 | 会员编号 |

构建商品网络前，首先要构建商品的组合数据。从原始数据集合筛选出销售流水号和商品小类这两列集，同一销售流水号码标识商品小类在统一购物篮中。然后根据销售流水号筛选出每次购买的商品组合。部分商品组合如表5-2所示所示，每一行代表一次交易，一次交易的购买的商品用空格隔开。

表5-2 商品组合

|  |
| --- |
| 礼盒 橄榄菜 |

续表5-2

|  |
| --- |
| 成人牙膏 洗护发套装 桶装面 益生菌乳酸奶 专柜其它面包 超市塑料袋 |
| 散装糕点 含片类糖果 |
| 其它保健品 |
| 散装糖果 纯牛奶 散装巧克力 |
| 其它餐具 散装糕点 成人牙膏 |
| 酸奶 超市购物袋/篮 其它零食 |
| 其它茶饮料 海产制品 洗衣粉 含乳饮料 散装糖果 橙类 |
| 花果类 超市塑料袋 进口糖果 叶菜类 其它浆果 |

构建商品网络，使用4.2节的generateRules方法。实验构建网络选择的最小置信度1/tn，支持度均是0.001,其中tn是商品组合数。以用节点表示商品，边表示商品间的关联性，边的权值表示关联性的强度，构建的网络节点数是1039，边的数量是282192。

### 5.3.2商品特征的选择

选出的影响力最大化商品组合，因为没有实际促销活动的结果供参考和验证，所以在本文使用商品的基本特征和在网络中重要性度量的特征来分析实验结果。本文选取了三个特征来衡量影响力最大化商品的特点，第一个是反映销售量情况的特征（销售量排名），第二个和第三个是反映商品重要性的特征（带权出度排名和pagerank值排名）。

1、商品销售量排名。用来衡量影响力最大化商品的销量与畅销商品的关系

2、带权出度排名。用来衡量影响力最大化商品直接影响邻居节点的强度。

3、pagerank值排名。在2014年戚戚[14]认为商品间的联系与网页间的联系相似，为了挖掘出一些重要的商品，即潜在的可能带动销售的商品，使用pagerank算法求出商品的pagerank值，排名靠前的则认为更可能是重要的商品。本文将使用该方法查看影响力最大化商品的pagerank值的排名情况。pagerank的两个参数阻尼系数（gehpi软件中称为概率p），收敛阈值（gehpi软件中称为误差epsilon），分别设置为0.85和0.0001。

根据商品网络的特征，以商品销量排名为横坐标，以带权值出度排名为纵坐标绘制散点图如图5-1所示。以商品销量排名为横坐标，以pagerank值排名为纵坐标绘制散点图如图5-2所示。从图可以看出带权出度排名与销量排名并没有明显的相关性，而pagerank排名与销量排名呈现正相关性，所以在分析商品特征时，只用销量排名和带权出度排名分析商品的特征。

图5-1 带权出度排名与销量排名的关系图

图5-2 pagerank排名与销量排名的关系图

## 5.3实验结果

### 5.3.1社区划分结果

网络被划分成7个社区，特征如表5-3所示。

表5-3 社区特征

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 社区 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 节点数 | 258 | 150 | 176 | 208 | 235 | 8 | 4 |
| 边数 | 42090 | 913 | 9613 | 14389 | 25964 | 20 | 12 |
| 平均入度 | 163.1395 | 6.0867 | 54.6193 | 69.1779 | 110.4851 | 2.5000 | 3.0000 |
| 平均出度 | 163.1395 | 6.0867 | 54.6193 | 69.1779 | 110.4851 | 2.5000 | 3.0000 |
| 密度 | 0.6347 | 0.0408 | 0.3121 | 0.3341 | 0.4721 | 0.3571 | 1.0 |

每个社区中的商品小类所属的经营小类如表5-4。

表5-4 社区中的经营小类

|  |  |  |
| --- | --- | --- |
| 社区 | 数量 | 经营小类的名称 |
| 1 | 32 | 饼干糕点 禽蛋类 厨房用品 快餐 手机通讯 蔬菜 母婴用品 即食食品 畜肉类 消杀用品 水果 腌菜 日用小件 粮油 休闲食品 酒 水产 综合生鲜 奶制品 豆制品 休闲小吃 女性护理用品 中式面制品 饮料 家居清洁护理 副食品 应节食品 清洁工具 禽肉类 冲饮品 熟肉制品 冷冻冷藏食品 |
| 2 | 34 | 运动服装 皮具箱包 内衣 数码商品 彩电 家居用品 手机通讯 配饰 童装 冲饮品 婴幼产品 袜子 婴幼食品 生活电器 日用小件 休闲装 体育器材 男士服装 少淑女装 黄金珠宝 厨卫 碟机音响 办公用品 化妆品 儿童用品 洗衣机 功能服 厨房用品 (停用)生活配套 钟表 皮鞋 保健滋补品/药品 淑女装 冰箱 |
| 3 | 17 | 生活用纸 母婴用品 文化用品 数码商品 家居用品 消杀用品 家纺用品 汽车用品 眼镜 家居清洁护理 生活电器 厨房用品 日用小件 冲饮品 文化音像制品 清洁工具 床上用品 |
| 4 | 29 | 家纺用品 生活电器 面部护理用品 厨房用品 消杀用品 美容工具 母婴用品 成人用品 婴幼产品 婴幼食品 床用品 日用小件 生活用纸 宠物食品用品 女性护理用品 洗浴用品 男士洗护用品 洗发美发用品 床上用品 保健滋补品/药品 化妆品 口腔清洁用品 粮油 家居清洁护理 衣物清洁用品 身体护理用品 冲饮品 文化音像制品 糖果巧克力 |
| 5 | 27 | 饼干糕点 生活电器 面部护理用品 厨房用品 消杀用品 烘焙面制品 母婴用品 即食食品 饮料 婴幼食品 禽蛋类 日用小件 应节食品 酒 奶制品 男士洗护用品 文化音像制品 家纺用品 休闲食品 快餐 粮油 副食品 水产 冲饮品 保健滋补品/药品 糖果巧克力 冷冻冷藏食品 |
| 6 | 1 | 酒 |
| 7 | 1 | 空调 |

根据社区划分结果可以直观的估计商品小类间影响关系（关联性），以社区3为例，可以从商品小类图中抽取出经营小类间的关系，如图5-1所示。从这个图可以直观的看出各个商品小类之间的关系，如厨房用品，日用小件，清洁工具这三者在这个社区中是关联较为紧密的，说明用户喜欢同时购买这些商品小类中的商品，如果其中一个小类的商品促销，可能会潜在性的带动其他商品小类商品的销售。

图5-1 社区3的商品网络



### 5.3.2基于独立级联模型的影响力最大化算法实验结果

实验设置独立级联模型的k为15，选出影响力最大化的商品组合中的15商品所属的社区如表5-5所示，每行中商品的顺序是商品假如到该社区影响力最大化商品集合顺序。

表5-5 独立级联模型最大化算法选取的商品组合

|  |  |
| --- | --- |
| 社区编号 | 影响最大商品组合 |
| 1 | 冰鲜鱼类 |
| 2 | 其它照相器材，玉兰油 OLAY I |
| 3 | 筷子(竹/木)，汤勺(不锈钢) ，洁厕剂，牙签/牙签盒，陶瓷面碗（5-6寸），其它餐具 |
| 4 | 成人牙刷，杀虫喷雾，有芯卷纸，男士洗面奶，香皂 |
| 5 | 夹心饼干 |
| 6 |  |
| 7 |  |

在整个网络中的商品特征降序得出的排名，筛选出影响力最大组合中的15个商品，如表5-6所示：

表5-6 独立级联模型最大化算法选取的商品特征

|  |  |  |  |
| --- | --- | --- | --- |
| 商品 | 销量排名 | 带权出度排名 | pagerank排名 |
| 冰鲜鱼类 | 65 | 280 | 54 |
| 其它照相器材 | 968 | 866 | 218 |

续表5-6

|  |  |  |  |
| --- | --- | --- | --- |
| 玉兰油 OLAY I | 484 | 858 | 394 |
| 筷子(竹/木) | 310 | 187 | 284 |
| 汤勺(不锈钢) | 593 | 58 | 579 |
| 洁厕剂 | 191 | 139 | 149 |
| 牙签/牙签盒 | 311 | 284 | 311 |
| 陶瓷面碗（5-6寸） | 491 | 142 | 495 |
| 其它餐具 | 260 | 240 | 229 |
| 成人牙刷 | 38 | 480 | 31 |
| 杀虫喷雾 | 675 | 462 | 715 |
| 有芯卷纸 | 24 | 485 | 20 |
| 男士洗面奶 | 328 | 490 | 249 |
| 香皂 | 197 | 451 | 195 |
| 夹心饼干 | 84 | 388 | 70 |
| 平均排名 | 334.6 | 387.3333 | 266.2 |

对比表5-5和表5-6，对热量传播模型的影响力最大化商品组合中商品的特征的分析可得出如下的结论：

1. 由表5-5和表5-6可知，在6社区和7社区过小，在这两个社区并未选取出种子节点，因为在小社区影响力不能像大社区扩展的那么大。
2. 从表5-6知道影响力最大的商品的销量排名范围是65~968，带权出度排名范围是58~866，因为组合最优化的限制，影响力最大的商品并未集中在某一排名段，而是尝试在整个网络中寻找影响力最大节点。
3. 从商品在各个社区的分布情况来看，实验中设置的社区的打分没有起到预期的作用，原因是选取影响力最大化种子过程中，很少出现商品影响相等的情况。
4. 从表5-6中得到商品的销量平均排名和商品的带权值出度的平均排名均靠前，说明该模型选取商品的特点是综合考虑销量较高及与对其他商品有较强影响的商品。
5. 观察同一社区的影响力最大化商品组合，可以发现这些商品有销售量排名靠前和销量靠后的商品，说明算法考虑了销量较高的和影响力增益大的商品后，如果发现有销量较低的但影响力增益大的商品，也会将其加入到影响力最大化商品集合中。如社区4选取的影响力增益最大商品，按选取顺序，销量排名分别是38、675、24、328、197，销量波动刚开始比较大。社区3选取的影响力增益最大商品，销量波动较小。

### 5.3.3基于热量传播模型的影响力最大化算法实验结果

实验中热量传播模型的参数设置如表5-7所示，其中n是图节点个数，k是种子数：

表5-7 热量传播模型算法参数表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 参数组 | k | t | a | threshold | H0 |
| 1 | 15 | 0.6 | 1.0 | 0.35 | 500 |
| 2 | 15 | 0.6 | 1.0 | 0.35 | 200 |
| 3 | 15 | 0.6 | 1.0 | 0.35 | n/k |

参数组2设置下，选出影响力最大化的商品组合中的15商品所属的社区如表5-8所示,每行中商品的顺序是商品假如到该社区影响力最大化商品集合顺序：

表5-8 热量传播模型最大化算法选取的商品组合

|  |  |
| --- | --- |
| 社区编号 | 影响最大商品组合 |
| 1 | 肉松，料酒，鲜鸡副产品，进口调味粉 |
| 2 | 镜头 |
| 3 | 磨刀石，桌布，其它保鲜盒/饭盒，其它家居 |
| 4 | 干燥剂，脱毛膏/蜡，婴儿护肤用品，童伞 |
| 5 | 烘焙咖啡，五谷/蔬菜饮料 |
| 6 |  |
| 7 |  |

在整个网络中的商品特征降序得出的排名，筛选出影响力最大组合中的15个商品，如表5-9所示：

表5-9 热量传播模型最大化算法选取的商品特征

|  |  |  |  |
| --- | --- | --- | --- |
| 商品 | 销量排名 | 带权出度排名 | pagerank排名 |
| 肉松 | 983 | 9 | 995 |
| 料酒 | 156 | 85 | 119 |
| 鲜鸡副产品 | 984 | 12 | 1021 |
| 进口调味粉 | 1015 | 246 | 1014 |
| 镜头 | 894 | 884 | 169 |
| 磨刀石 | 940 | 4 | 926 |
| 桌布 | 494 | 148 | 503 |
| 其它保鲜盒/饭盒 | 976 | 3 | 963 |
| 其它家居 | 684 | 123 | 667 |
| 干燥剂 | 1004 | 2 | 991 |
| 脱毛膏/蜡 | 936 | 615 | 985 |
| 婴儿护肤用品 | 224 | 714 | 261 |
| 童伞 | 979 | 1 | 957 |
| 烘焙咖啡 | 964 | 430 | 945 |
| 五谷/蔬菜饮料 | 388 | 549 | 421 |
| 平均排名 | 774.7333 | 255 | 729.1333 |

对比表5-8和表5-9，对热量传播模型的影响力最大化商品组合中商品的特征的分析可得出如下的结论：

1. 跟独立级联模型一样，社区6和社区7过小，影响力扩展不开，算法没在这两个社区中选择商品加入到影响最大化商品组合中。
2. 从销量排名来看， 商品的平均销量排名靠前，而商品的带权出度平均排名靠后，这说明该模型的特点是综合考虑选取销量较少，但是对其他商品影响较强的商品。
3. 观察同一社区的影响力最大化商品组合，也可以发现每个社区先选出的影响力增益最大商品是销量排名靠后的，即销量较低的。

### 5.3.4对比简析

独立级联模型和热量传播模型的影响力最大化算法的相同点是均挑选对其他商品的有较强的直接影响强度（带权出度和）的商品，不同点是独立级联模型选取的商品销量上较热量传播模型的选取的商品销量高。独立级联模型直观简单，如果基于先验知识给出了边的权值，那么所选取的影响力最大化商品集合基本是可以确定的，而热量传播模型因为有多个可控参数，如时间，传导率，初始温度等，可以根据网络特征进行相应的进行相应的调整，可以得到多个影响力最大化商品组合供选择。热量传播模型比热量传播模型更具有灵活性，但是也带来一定的操作复杂性。

## 5.5本章小结

本文选取了三个特征来衡量影响力最大化商品的特点，第一个是反映销售量的特征（销量排名），第二个和第三个是反映商品重要性的特征（pagerank值排名和带权出度排名）。但是发现商品销量排名与pagerank值排名具正相关性，所以只选择了销量排名和带权出度排名可度量商品的特点。

在实验中分析选取影响力最大化商品集合前社区的划分情况，可以初步的直观的了解到商品所属的经营小类间的直接影响关系的强度。在社区划分的基础上，分析这个两个影响力最大化算法求取的商品的一些特点，发现基于独立级联模型的影响力最大化算法选出的商品具有销量较高（销量排名靠前）且对邻居商品的直接影响（带权出度）也较高的特点，而基于热量传播模型的影响力最大化算法选出的商品对邻居商品的直接影响较高，但是商品的销量较低（销量靠后）的特点。观察每个社区选取的影响增益最大的商品，发现本文设置的score分值并未起到预期的效果，很可能的原因很少出现社区节点的最大影响增益相同的情况。

# 第六章 总结与展望

本文旨在为零售商选择超市促销品提供商品组合建议，促销商品的选择是一个组合最优化的问题，即所选的商品组合要尽可能的带动其他商品的销售。为解决这样的一个问题，笔者了解了购物篮分析的研究现状并学习了相关的理论知识。在购物篮分析方面，很多的研究方法是使用关联规则算法分析商品的关联性，也有的方法通过频繁项集发现或者使用聚类分析等数据挖掘的方法挖掘商品的组合情况。基于商品网络的购物篮方法仍比较少，且大多是进行比较简单的分析，如第一章商品网络研究现状讲述的基于商品网络的中心度分析，聚类分析等。因此笔者想在商品网络上求解促销商品组合的选取问题，同时也可以为商品网络的研究贡献一份力量。

为了寻找促销商品组合的选取问题，笔者通过查阅文献发现在社交网络影响力传播研究中有一个研究热点是影响力最大化算法研究，该研究是寻找一部分人，通过这些人可以尽可能的影响到更多人去做某件事，如购买某件商品，接受某个决定，给某人投票等。影响力最大化问题和促销商品组合的选取问题均是组合最优化问题，并且促销商品带动其他商品的销售的能力可理解为促销商品对其他商品的影响强度。

基于这样的启发，笔者想将社交网络影响力最大化算法应用到商品网络中，求取影响力最大化的商品组合（促销商品组合）。本文在第四章回答了影响力最大化问题在购物篮分析中的三个问题，商品网络的如何构建，选取怎样的模型来对商品影响力传播进行建模，在给定的影响力传播模型下使用何种方法来求解组合最优化的问题。

在了解基于商品网络的购物篮分析的研究现状后，撰写了本文的第二章基于商品网络的购物篮分析概述。了解了常用的商品网络构建方法后选取了apriori算法进行求取商品的相关性，然后以商品作为节点，在统一频繁项集中的商品认为存在边的关系，商品关联性作为边的权重。

本文的第四章是笔者学习社交网络影响力最大化方法的研究撰写的概述，为商品影响力建模提供了多种模型作为选择，并且明确了影响力最大化算法的求解方式。最终，经过第四章的分析选择了独立级联模型和热量传播模型来对商品影响力传播建模，影响力最大化的问题采用hui li等人[18]的思路进行求解。

在真实的交易数据中对本文设计的算法进行了实验探究，发现基于独立级联模型的影响力最大化算法寻找的商品呈现出对邻居节点的影响较强且销量较高的特点，而基于热量传播模型的影响力最大化算法寻找的商品同样呈现出对邻居节点的影响较强，但是商品的销量较低的特点。

本文的实验主要是一个探究性的实验，选出的商品集合是否在实际应用中起到良好的效果需要实践进行验证，但是可为为零售商选择促销商品提供多一些信息参考。本文提出的社区打分问题未能起到使种子节点均衡分布在各个社区的效果，原因是各个社区的节点最大影响增益很少出现相同的情况。在未来笔者会尝试寻找有效的方法均衡的从各个社区选择影响力最大的影响力节点，从而使影响力更好的传播网络的各个部分。因为在不同的时间（如不同的季节）消费者的行为会呈现不同的特点，因此影响力最大的商品组合可能也会有所变化，笔者在未来可能会给购物篮数据加上时间窗口，尝试分析商品间的影响程度与时间的关系。

# 附录

## 独立级联模型的主要代码

基于独立级联模型的影响力最大化算法的主代码：

|  |
| --- |
| def max\_influ\_using\_ica\_diffu(subg, k):  mag = init\_mag(subg)  seed\_set = []  for i in range(len(subg)):  seed\_set.append([[],0])  for i in range(k):  print(i," seed")  idx = arg\_max\_mag(mag, seed\_set, subg)  seed\_set[idx][0].append(mag[idx][0])  seed\_set[idx][1] += mag[idx][1]  print(mag)  subgth = subg[idx]  mag[idx] = get\_node\_with\_max\_gain(subgth,seed\_set[idx][0])  repeat\_time = 1000  influ\_sum = 0  for i in range(len(subg)):  print(i)  t1 = time.time()  seed\_set[i][1] = avgiac(subg[i], seed\_set[i][0], repeat\_time)  influ\_sum += seed\_set[i][1]  print(time.time() - t1)  return seed\_set |

独立级联模型的主要代码

|  |
| --- |
| from \_\_future\_\_ import division  from copy import deepcopy # copy graph object  import random  # from priorityQueue import PriorityQueue as PQ  import networkx as nx  import networkx as nx  import time  def findccs\_in\_digraph(di\_g):  # remove blocked edges from directed graph G  E = deepcopy(di\_g)  edge\_rem = [e for e in E.edges() if random.random() > E[e[0]][e[1]]['weight']]  E.remove\_edges\_from(edge\_rem)  ccs = list(nx.strongly\_connected\_components(E))  return ccs  def runiac (di\_graph, seeds):  ''' Runs independent arbitrary cascade model.  Input: G -- networkx graph object  seed -- initial set of vertices  Output: T -- resulted influenced set of vertices (including S)  '''  actived\_nodes = deepcopy(seeds) # copy already selected nodes  # ugly C++ version  i = 0  while i < len(actived\_nodes):  for v in di\_graph[actived\_nodes[i]]: # for neighbors of a selected node  if v not in actived\_nodes: # if it wasn't selected yet  w = di\_graph[actived\_nodes[i]][v]['weight']  if random.random() <= w:  actived\_nodes.append(v)  i += 1  return actived\_nodes  def avgiac(di\_graph, seeds, repeat\_time):  '''  Input:  G -- undirected graph  S -- seed set  Ep -- propagation probabilities  I -- number of iterations  Output:  avg -- average size of coverage  '''  avg = 0  for i in range(repeat\_time):  avg += float(len(runiac(di\_graph, seeds)))  avg /= repeat\_time  return avg  def greedy\_ic\_digraph(digraph, k, r=20):  seed = []  for i in range(k):  print(i)  time2k = time.time()  scores = {v: 0 for v in digraph}  for j in range(r):  print (j,)  ccs = findccs\_in\_digraph(digraph)  for cc in ccs:  for v in seed:  if v in cc:  break  else: # in case CC doesn't have node from S  for u in cc:  scores[u] += float(len(cc))/r  max\_v, max\_score = max(scores.items(), key=lambda key\_value: key\_value[1])  seed.append(max\_v)  print()  print(time.time() - time2k)  return seed |

## 热量传播模型的主要代码

基于热量传播模型的影响力最大化算法的主代码：

|  |
| --- |
| def max\_influ\_using\_heat\_diffu(k, heat\_function, threshold = 0.6, t = 0.3, alpha=1.0, num = 1):  """  Input：  k ： 需要选取的种子节点个数  t ： 与邻居节点的交互时间  threshold : 激活阈值  num : 影响力传播的层数，总花费时间是 t \* num  output:  seed\_set : 种子集合  """  mag = init\_mag(subg\_H, k, heat\_function, threshold, t, alpha, num)  seed\_set = []  for i in range(len(subg\_H)):  seed\_set.append([[],0])  for i in range(k):  idx = arg\_max\_mag(mag, seed\_set, subg\_H)  seed\_set[idx][0].append(mag[idx][0])  seed\_set[idx][1] += mag[idx][1]  H = subg\_H[idx]  h0 = heat\_function(H.shape[0])  mag[idx] = get\_new\_one\_of\_mag\_value(H, mag[idx], seed\_set[idx], h0, threshold, t, alpha, num)  return seed\_set |

获取影响增益最大的节点的代码

|  |
| --- |
| def get\_node\_with\_max\_gain(H, seed\_set, h0, threshold = 0.6, t = 0.1, alpha = 1.0, num = 1):  """  挑选一个种子， 该的种子 与 种子集合，一起传播的影响力最大  input:  H : 一个子图的影响力传播矩阵  seed\_set : 当前子图已经选择的种子集合， 一个集合{节点1，节点2}  h0 : 种子的初始温度  threshold ： 激活阈值  t : 与相邻节点的交互时间, 可以理解为单位时间  alpha : 传导系数  num : 影响力传播的层数，总花费时间是 t \* num  return :  (idx, node\_influ[idx]) ： (节点， 节点的影响力)  """  number\_node = H.shape[0]  c = np.matrix( np.zeros((number\_node, num)) )  node\_influ = []  for node in range(number\_node):  new\_seed\_set = seed\_set.copy()  new\_seed\_set.append(node)  # get submatrix  select\_index = new\_seed\_set.copy()  for i in range(num):  select\_index = get\_out\_nei\_index\_also\_own(H, select\_index)  submatrix\_size = len(select\_index)  select\_index = sorted(select\_index)  submatrix\_H = get\_sub\_matrix(H, select\_index)  map\_submatrix\_index = { index:i for i,index in enumerate(select\_index)}  # 初始化温度  f0 = np.zeros(submatrix\_size)  for node in new\_seed\_set:  f0[map\_submatrix\_index[node]] = h0  ft = f0.copy()  ft = ft[:, np.newaxis]  ll = ft  hh = submatrix\_H  for i in range(num):  ft = h\_pass\_t(submatrix\_H, ft, t, alpha)  sum\_influ = 0  for heat in ft:  if heat > threshold:  sum\_influ += 1  node\_influ.append([node, sum\_influ])  node, influ = max(node\_influ, key = lambda x: x[1])  return (node, influ) |

计算H矩阵的代码

|  |
| --- |
| def get\_subg\_sm(subg):  """  Input  subg :子图列表  return :  subg\_sm ： 每个子图的邻接矩阵，有weight则是weight的值矩阵，否则关联则为1  """  par\_num = len(subg)  # 子图的相似度矩阵  subg\_sm = []  for i in range(par\_num):  subg\_sm.append(nx.to\_numpy\_matrix(subg[i]))  return subg\_sm  def get\_subg\_outd(subg):  """  input:  subg : 子图列表  return:  subg\_d : 每个子图节点的出度数，就是有多少条出边  """  par\_num = len(subg)  # 子图的相似度矩阵  subg\_outd = []  for i in range(par\_num):  d\_dict = subg[i].out\_degree()  nodes = subg[i].nodes()  d = []  for i in range( len(nodes) ):  d.append(d\_dict[nodes[i]])  subg\_outd.append(d)  return subg\_outd  def get\_subg\_H(subg\_sm, subg\_outd):  """  Input:  subg\_sm : 每个子图的相似度矩阵  subg\_outd : 每个子图的节点的出度数  return :  subg\_H : 每个子图的影响力传播矩阵    Hij = wj \* pij / dj when (vj, vi) belong to E,  Hij = -(Ti \* wi / di) \* sum(pik) when i == j and (vi, vk) belong to E , and Tii = 1 if degree of vi more than 0, otherwise Ti = 0  Hij = 0 otherwise  """  par\_num = len(subg\_sm)  # 求影响力传播矩阵  subg\_H = []  for i in range(par\_num):  nodes\_num = subg\_sm[i].shape[0]  ci = []  for d in subg\_outd[i]:  if d > 0:  ci.append(1.0/d)  else :  ci.append(0)  H = np.matrix( np.zeros((nodes\_num, nodes\_num)) )  # H = np.zeros((nodes\_num, nodes\_num))  for nodei in range(nodes\_num):  # (n x 1) .x (n x 1) => (n x 1)  H.A[nodei,:][:,np.newaxis] = np.multiply( subg\_sm[i][:,nodei] , np.matrix( np.array(ci)[:,np.newaxis] ) )  # H[nodei,:][:,np.newaxis] = np.multiply( subg\_sm[i][:,nodei] , np.array(ci)[:,np.newaxis] )  for nodei in range(nodes\_num):  H[nodei, nodei] = 0  for nodei in range(nodes\_num):  if subg\_d[i][nodei] != 0:  H[nodei, nodei] = -np.sum(subg\_sm[i][nodei, :]) \* 1.0 \* ci[nodei]  subg\_H.append(H)  return subg\_H |

在mag列表获取取影响增益最大节点在mag列表的索引的代码

|  |
| --- |
| def arg\_max\_mag(mag, seed\_set, subg\_H):  """  Input:  mag : [[子图1中节点k，在该子图的影响力增益], [子图2中节点m，在该子图的影响力增益处]... ]  return :  返回索引  """  influ = [ [i,int(item[1]+0.3), len(subg\_H[i]) / (len(seed\_set[i][0]) + 1)] for i, item in enumerate(mag)]  print(influ)  idx, influ, priori= max(influ, key=lambda x : [x[1],x[2]])  return idx |

初始化mag列表

|  |
| --- |
| def init\_mag(subg\_H, k, heat\_function, threshold = 0.6, t = 0.1, alpha = 1.0, num = 1):  """  Input  subg\_H : 子图的传播矩阵列表  k : 需要选取的种子个数  threshold ： 激活阈值  t : 与相邻节点的交互时间  alpha : 传导系数  num : 影响力传播的层数，总花费时间是 t \* num  return：  mag : [[子图1中节点k，在该子图的影响力增益], [子图2中节点m，在该子图的影响力增益处]... ]  存放的每个子图带来最大影响力的节点及其影响力增益  """  mag = []  for subgth in range(len(subg\_H)):  H = subg\_H[subgth]  number\_node = H.shape[0]  h0 = heat\_function(number\_node)  node\_with\_max\_gain = get\_node\_with\_max\_gain(H, [], h0, threshold, t, alpha, num)  mag.append(node\_with\_max\_gain)  return mag |

## 社区划分的主要代码

快速层次聚类主函数代码代码：

|  |
| --- |
| def generate\_di\_dendrogram(di\_graph, part\_init=None, weight='weight', resolution=1., randomized = False):  """Find communities in the graph and return the associated dendrogram  A dendrogram is a tree and each level is a partition of the graph nodes.  Level 0 is the first partition, which contains the smallest communities,  and the best is len(dendrogram) - 1. The higher the level is, the bigger  are the communities  Parameters  ----------  graph : networkx.DiGraph  the networkx graph which will be decomposed  part\_init : dict, optional  the algorithm will start using this partition of the nodes. It's a  dictionary where keys are their nodes and values the communities  weight : str, optional  the key in graph to use as weight. Default to 'weight'  resolution : double, optional  Will change the size of the communities, default to 1.  represents the time described in  "Laplacian Dynamics and Multiscale Modular Structure in Networks",  R. Lambiotte, J.-C. Delvenne, M. Barahona  Returns  -------  dendrogram : list of dictionaries  a list of partitions, ie dictionnaries where keys of the i+1 are the  values of the i. and where keys of the first are the nodes of graph  Raises  ------  TypeError  If the graph is not a networkx.Graph  See Also  --------  best\_partition  Notes  -----  Uses Louvain algorithm  References  ----------  .. 1. Blondel, V.D. et al. Fast unfolding of communities in large  networks. J. Stat. Mech 10008, 1-12(2008).  .. 2. Community structure in directed networks E. A. Leicht1 and M. E. J. Newman1, 2  Examples  --------  >>> G=nx.erdos\_renyi\_graph(100, 0.01, directed=True)  >>> dendo = generate\_di\_dendrogram(G)  >>> for level in range(len(dendo) - 1) :  >>> print("partition at level", level,  >>> "is", partition\_at\_level(dendo, level))  """  if type(di\_graph) != nx.DiGraph:  raise TypeError("Bad graph type, use only directed graph")  # special case, when there is no link  # the best partition is everyone in its community  if di\_graph.number\_of\_edges() == 0:  part = dict([])  for node in di\_graph.nodes():  part[node] = node  return part  current\_graph = di\_graph.copy()  di\_status = DiStatus()  di\_status.init(current\_graph, weight, part\_init)  di\_status\_list = list()  \_\_one\_level(current\_graph, di\_status, weight, resolution, randomized)  new\_mod = \_\_di\_modularity(di\_status)  partition = \_\_renumber(di\_status.node2com)  di\_status\_list.append(partition)  mod = new\_mod  current\_graph = induced\_di\_graph(partition, current\_graph, weight)  di\_status.init(current\_graph, weight)  while True:  \_\_one\_level(current\_graph, di\_status, weight, resolution, randomized)  new\_mod = \_\_di\_modularity(di\_status)  if new\_mod - mod < \_\_MIN:  break  partition = \_\_renumber(di\_status.node2com)  di\_status\_list.append(partition)  mod = new\_mod  current\_graph = induced\_di\_graph(partition, current\_graph, weight)  di\_status.init(current\_graph, weight)  return di\_status\_list[:] |

计算有向图Q值的python 代码

|  |
| --- |
| def \_\_di\_modularity(di\_status):  """  Fast compute the modularity of the partition of the graph using  status precomputed  """  links = float(di\_status.total\_weight)  result = 0.  for community in set(di\_status.node2com.values()):  internals = di\_status.internals.get(community, 0.)  in\_degree = di\_status.in\_degrees.get(community, 0.)  out\_degree = di\_status.out\_degrees.get(community, 0.)  if links > 0:  # eii - ai\_out \* ai\_in  result += internals / links - in\_degree \* out\_degree / (links \*\* 2)  return result |

将每个节点到调整至一个相邻的社区，加入这个社区可达模块值增益最大。

|  |
| --- |
| def \_\_one\_level(di\_graph, di\_status, weight\_key, resolution, randomized):  """Compute one level of communities  """  modified = True  nb\_pass\_done = 0  # 计算当前的模块值  # https://en.wikipedia.org/wiki/Modularity\_(networks) 公式（4）  cur\_mod = \_\_di\_modularity(di\_status)  new\_mod = cur\_mod  while modified and nb\_pass\_done != \_\_PASS\_MAX:  cur\_mod = new\_mod  modified = False  nb\_pass\_done += 1  nodes = di\_graph.nodes()  if randomized:  random.shuffle(nodes)  for node in nodes:  # 节点所属的社区  com\_node = di\_status.node2com[node]  degc\_totw\_in = di\_status.in\_gdegrees.get(node, 0.) / di\_status.total\_weight degc\_totw\_out = di\_status.out\_gdegrees.get(node, 0.) / di\_status.total\_weight  neigh\_communities = \_\_di\_neighcom(node, di\_graph, di\_status, weight\_key)  \_\_di\_remove(node, com\_node,  neigh\_communities.get(com\_node, 0.), di\_status)  best\_com = com\_node  best\_increase = 0  # key  for com, dnc in neigh\_communities.items():  incr = resolution \* dnc - \  di\_status.in\_degrees.get(com, 0.) \* degc\_totw\_out - \  di\_status.out\_degrees.get(com, 0.) \* degc\_totw\_in  if incr > best\_increase:  best\_increase = incr  best\_com = com  \_\_di\_insert(node, best\_com,  neigh\_communities.get(best\_com, 0.), di\_status)  if best\_com != com\_node:  modified = True  new\_mod = \_\_di\_modularity(di\_status)  if new\_mod - cur\_mod < \_\_MIN:  break |

# 参考文献

1. [] Julander C. BASKET ANALYSIS: A NEW WAY OF ANALYSING SCANNER DATA [J]. International Journal of Retail & Distribution Management, 1992, 20(7). [↑](#endnote-ref-1)
2. [] Agrawal R. Mining association rules between sets of items in large databases [J]. Acm Sigmod Record, 1993, 22(2):207-216. [↑](#endnote-ref-2)
3. [] Chena, Y. L., Tang, K., Shena, R. J., & Hua, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems,40* (2), 339-354. [↑](#endnote-ref-3)
4. [] Berry, Michael J A, Linoff, et al. Data mining techniques for marketing, sales and customer relationship[J]. 2004. [↑](#endnote-ref-4)
5. [] Tang K, Chen Y L, Hu H W. Context-based market basket analysis in a multiple-store environment[J]. Decision Support Systems, 2008, 45(1):150-163. [↑](#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. [] Pingho T, Pan S, Shuoshiung C. Finding ideal menu items assortments: an empirical application of market basket analysis.[J]. Cornell Hospitality Quarterly, 2010, 51(4):492-501. [↑](#endnote-ref-7)
8. [] Bell D R, Corsten D, Knox G. Unplanned Category Purchase Incidence: Who Does It, How Often, and Why[J]. 2009. [↑](#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. A new approach for generating efficient sample from market basket data[J]. Expert Systems with Applications, 2011, 38(3):1321-1325. [↑](#endnote-ref-10)
11. [] Woo J. Market Basket Analysis algorithms with MapReduce[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2013, 3(6):445–452. [↑](#endnote-ref-11)
12. [] Kim H K, Kim J K, Chen Q Y. A product network analysis for extending the market basket analysis[J]. Expert Systems with Applications, 2012, 39(8):7403–7410. [↑](#endnote-ref-12)
13. [] 戚威. 基于复杂网络的购物篮商品网络分析研究[D]. 江苏科技大学, 2014. [↑](#endnote-ref-13)
14. [] Zinoviev D, Zhu Z, Li K. Building Mini-Categories in Product Networks[J]. Studies in Computational Intelligence, 2015, 597:179-190. [↑](#endnote-ref-14)
15. [] Kempe, David, Kleinberg, Jon, Tardos, &#. Maximizing the spread of influence through a social network[C]// ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2003:137--146. [↑](#endnote-ref-15)
16. [] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007:420-429. [↑](#endnote-ref-16)
17. [] B Borgs C, Brautbar M, Chayes J, et al. Maximizing Social Influence in Nearly Optimal Time[J]. Computer Science, 2012. [↑](#endnote-ref-17)
18. [] Li H, Bhowmick S S, Sun A. CINEMA: Conformity-Aware Greedy Algorithm for Influence Maximization in Online Social Networks[C]// International Conference on Extending Database Technology. 2013:323-334. [↑](#endnote-ref-18)
19. [] Ma H, Yang H, Lyu M R, et al. Mining social networks using heat diffusion processes for marketing candidates selection[C]// International Conference on Information and Knowledge Management. 2008:233-242. [↑](#endnote-ref-19)
20. [] Chen Y C, Zhu W Y, Peng W C, et al. CIM: Community-Based Influence Maximization in Social Networks[J]. Acm Transactions on Intelligent Systems & Technology, 2014, 5(2):529-544. [↑](#endnote-ref-20)
21. [] 余颖. 购物篮分析在网络零售业中的应用研究[D]. 天津大学, 2007. [↑](#endnote-ref-21)
22. []Efficient Influence Maximization in Social Networks [↑](#endnote-ref-22)
23. [] Ma H, Yang H, Lyu M R, et al. Mining social networks using heat diffusion processes for marketing candidates selection[C]// International Conference on Information and Knowledge Management. 2008:233-242. [↑](#endnote-ref-23)
24. [] Li H, Bhowmick S S, Sun A. CINEMA: Conformity-Aware Greedy Algorithm for Influence Maximization in Online Social Networks[C]// International Conference on Extending Database Technology. 2013:323-334. [↑](#endnote-ref-24)
25. [] Jiang Q, Song G, Cong G, et al. Simulated Annealing Based Influence Maximization in Social Networks.[C]// AAAI Conference on Artificial Intelligence. 2011. [↑](#endnote-ref-25)
26. [] Wang C, Chen W, Wang Y. Scalable influence maximization for independent cascade model in large-scale social networks[J]. Data Mining & Knowledge Discovery, 2012, 25(3):545-576. [↑](#endnote-ref-26)
27. [] Leicht E A, Newman M E J. Community Structure in Directed Networks[J]. Physical Review Letters, 2008, 100(11):2339-2340. [↑](#endnote-ref-27)
28. [] Blondel, V.D. et al. Fast unfolding of communities in large networks. J. Stat. Mech 10008, 1-12(2008).

    ## 致谢

    论文写至此也意味着即将毕业，回想大学的生活基本是工作室，食堂和寝室3点一线，收获不算颇丰，相对于刚入学的茫然少年已然成长不少。

    记得刚上大一的时候，对于计算机这个名词的直觉反应是看电影，网游戏，写写word文档等。初入大学心中不禁有些茫然和强烈的新鲜感，有幸遇到了acm训练招小伙伴就一头扎入编程的世界。在acm训练过程中，我知道了玩电脑除了看电影和玩游戏，通过编程解决实际问题也是一件有趣的事情，并且在这里认识了一路支持我和鼓励我的朋友们，还有一位将我领进科研这个大门的充满热情的杨老师。

    杨老师让我有机会跟王师姐学习社交网络影响力最大化理论，在这个过程中学习了如何去阅读一篇论文和通过编程解决实际的问题，现在依然能清晰的记得与王师姐讨论问题，杨老师指导我书写邮件与论文作者交流等情景。这个过程无疑是我最充实和难忘的经历，这也是我选择这个论文题目的原因。一来是我想对所学知识进行一个总结，二来是解决购物篮分析这个我没接触过的问题，可以让我从查找文献，做实验，到最终的书写论文的每一步骤实践一遍。

    我衷心的感谢杨书新老师一直鼓励着我前进，培养了我对科研的兴趣，耐心的指导我查阅文献，书写论文，还给予我机会参加学术的活动等等。如果不是杨老师及时的指导，恐怕我会在大学的学习和生活中走很多的弯路。在此，向老师表示由衷的敬意和真诚的感谢！

    感谢吴雅林老师在大三和大四期间指导我学习图像处理的知识，让我了解了正在兴起的虚拟现实研究，并给我们分享了在韩国的科研经历。与老师交流的过程中，不管是在生活还是科研都教了我们很多。

    感谢陈小军老师给予我去深大做毕设的机会和为我提供购物篮这个应用场景，并及时的指出我的实验存在的问题并给予我建议。在这个过程我培养了主动学习的习惯，参加了学术活动并开阔了科研的视野。

    感谢江西理工大学的领导和老师们，是你们的不辞辛劳让我们享受到这么好的学习和生活环境，感谢你们在我们的学业倾注了大量的心血。

    感谢工作室的师兄师姐在我大学成长过程中给予我很多的建议，帮助我解开读研的一些疑惑。感谢一直督促和鼓励我备战考研和推免的游捷同学，陪我度过愉快大学生活的室友，陪我奋战多天完成论文的丁美坤同学等等，大学遇见你们真好。

    最后，我想感谢我的家人，特别是我的父母，一直给予我无微不至的关怀和支持。父母是我在成长路上的坚强的后盾,给予了我源源不断的动力，没有你们的陪伴就没有我现在的成就。

    感谢一路陪伴我的所有亲人和朋友，在此致以诚挚的谢意！ [↑](#endnote-ref-28)