# 基于商品影响力分析的购物篮方法研究

计算机科学与技术（软件方向）2012级 袁国文

指导老师 杨书新（副教授）

**摘要**： 选择促销商品的组合的问题是购物篮分析中的重要研究，通过了解购物篮分析的国内外研究现状，发现现有的方法不能很好的求解促销商品选择这一组合最优化问题。因为商品网络中商品影响传播与独立级联模型存在一定相似性，本文提出在商品网络使用基于独立级联模型影响力最大化算法。算法首先通过关联规则算法构建商品网络，然后进行社团划分，对每个社团求取影响增益最大的商品，每次从所有社区选择一个影响增益最大的商品作为促销商品，直到选择了指定个数的促销品。如果一个社区中商品被选取作为促销品，则从该社区从新选择新的影响增益最大的商品。最后在真实数据中，使用基于独立级联模型的影响力最大化算法求解了选择促销商品的问题，促销商品呈现出销量低但是对邻居节点影响强的特点。

**关键词：**购物篮分析；关联规则；商品网络；社团划分；影响力最大化率。

超市经常会有多种商品都在进行促销，促销商品需要尽可能带动其他商品的销售，基于这样的场景，选取促销商品的组合是一个组合最优化的问题。

购物篮分析（Market Basket Analysis，MBA）可分为两类，非商品网络和基于商品网络的研究，非商品网络的研究的主要是关联规则的发现等研究，基于商品网络的研究主要是商品社区划分和度数分布等研究。通过在DBLP数据库中查询sigkdd收录的购物篮分析的文献，发现现有的购物篮算法并不能很好的解决选择促销商品的组合的问题。

在商品网络分析中，未找到关于商品影响传播的研究。本文将在商品网络引入商品影响传播的概念，通过分析商品影响传播特点发现其与社交网络中的基于独立级联模型的影响力传播具有相似性，并且社交网络影响力最大化问题与选择促销商品的组合均是一个组合最优化的问题。所以在本文提出在商品网络中使用基于独立级联模型的影响力最大化算法来求取促销商品的组合。

在本文中将使用其中的两种算法，基于独立级联模型的影响力最大化算法和基于热量模型的影响力最大化算法求解促销商品选择问题。等论文基本写完在写章节介绍

1相关工作

购物篮分析可分为两类，非商品网络的购物篮分析和基于商品网络的购物篮分析。

在非商品网络的购物篮分析中主要是指商品关联性分析。1993年Agrawal等人在购物篮分析提出了关联规则算法，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。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，餐馆菜单的选择（tang等人2008，Haughton 等人2003等，Ting和Steve等人2010）。2011年Bogdan Hoanca等人用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。2013年Wiley Interdisc等人使用MapReduce来提高购物篮分析效率。

在基于商品网络的购物篮研究中，主要是进行商品社团的发现、度数分布及商品重要性研究。2012年Hyea Kyeong Kim等人为了能从整体的视角上来分析商品的特点，从客户-商品网络中抽取出一起购买网络和购物篮网络，并分析商品网络中节点的degree centrality与商品热销程度的关系等。2014年戚戚将根据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，并使用层次聚类进行了商品社团研究。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。

综合上述购物篮分析的方法，发现现有的购物篮分析方法并不能很好的解决选择促销商品的组合这一组合最优化的问题。

通过购物者的购物行为，可发现在逛超市的过程中，购物者见到一件商品并加入购物篮中，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，消费者在超市购买第一件商品后，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。

基于这样的商品影响传播行为，本文提出了基于独立级联模型的商品影响传播的购物篮分析方法，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

2 独立级联模型

对于一个给定的有向图，*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。让表示在第*t*轮被激活的种子集合。表示初始的种子集合。在*t>0*轮，一个给定的节点，被它在*t-1*步骤激活的邻居节点影响的概率是独立的，其中。在这样的模型下节点被激活的概率可以用(3.4)来表示，按照上面的迭代步骤，直到为空。

(3.4)

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。如果则边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。对于一个节点，将可到达的节点计做，节点数量是，节点的影响力是。如果想要求取影响力增益，则需要先从中删除,及它们的边得到。如果节点加入S会带来影响增益计作，那么计算如公式（3.5）所示：

（3.5）

以图3-1为例，图中有4个节点，7条边，假设每条边的权值均是0.5，那么该图的边的权值矩阵如表3-2所示，第1行和第1列代表节点的编号。假设当前的种子节点只有2号节点，现在需要求节点0，1，3其中一个节点作为新种子节点会带来的影响力增益。

表3-2 独立级联模型中有向图1的权值矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** |
| **0** | 0 | 0.5 | 0 | 0.5 |
| **1** | 0 | 0 | 0.5 | 0.5 |
| **2** | 0 | 0 | 0 | 0.5 |
| **3** | 0 | 0.5 | 0 | 0 |

通过对每条存在的边，随机给一个0到1之间的随机数字，可得到随机数矩阵，如表3-3所示。

表3-3 随机数矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **0** | **1** | **2** | **3** |
| **0** | 0 | 0.3 | 0 | 0.35 |
| **1** | 0 | 0 | 0.55 | 0.25 |
| **2** | 0 | 0 | 0 | 0.15 |
| **3** | 0 | 0.15 | 0 | 0 |

如果对于，如果则移除该边。最终可得到新图，如图3-2所示。

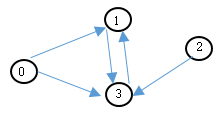


图3-2 随机级联有向图

先删除种子节点2及其可达的节点，可得到新图，如图3-3。



图3-3 影响增益图

根据图3-3，由公式（3.3）可求得节点的影响力增益如表3-4所示。

表3-4 独立级联模型节点的影响力增益

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **节点编号** | 0 | 1 | 2 | 3 |
| **影响力增益** | 1 | 0 | 0 | 0 |

**3 基于商品影响力最大化的购物篮分析**

本文的商品影响力最大化的购物篮分析方法为3个阶段，第一阶段是构建商品网络，第二阶段是使用基于社区划分的影响力最大化算法求解促销商品的选择的问题。

3.1 商品网络构建

根据传统的apriori算法，可以求出商品之间的关联性。用节点表示支持度不小于最小小支持度的商品。如果两个商品之间的关联性强度不小于最小置信度，就认为这两个商品节点存在有向边，边的权值是关联性强度。

3.2基于社区划分的影响力最大化算法

在使用有向图快速层次聚类的社区划分基础上，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，如果影响增益相等，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



1、设计选择k节点影响力最大化种子集合的算法（*get\_seed\_set*）如表4-10所示。

表4-10 *get\_seed\_set*算法伪代码

|  |
| --- |
| algorithm: *get\_seed\_set*  input: *G*, *k*  output: *seed\_set*  1. *subgFHC(G)*  2. *maginit\_mag(subg)*  3. *seed\_setempty list*  4. for each *i* in *[0,1,2…,len(subg)]*  5. *seed\_set.append([])*  6. for *i* in *[0,1,…,k-1]*  7. *idxarg\_max\_mag(mag, seed\_set, subg)*  8. s*eed\_set[idx][0].append(*  *mag[idx][0])*  9. *seed\_set[idx][1]seed\_set[idx][1] +mag[idx][1]*  10. *subgthsubg[idx]*  11. *mag[idx]get\_node\_with\_max\_gain (subgth, seed\_set[idx][0])*  12.return *seed\_set* |

*get\_seed\_set*算法输入有向图*G,*可种子个数 *k*，最后输出影响力最大的*k*节点种子集合。第1行是使用有向图层级聚类算法对图*G*进行社区划分得到社区列表*subg*。第2行是初始化*mag*列表，该列表存储每个社区当前会带来影响力增益最大的节点，及其增益。第3行是创建一个空的列表*seed\_set*，用于存储每个社区的种子节点。第4行至第5行是对于每个子图，给*seed\_set*添加一个空列表，表示该社区仍未选出种子节点。第6行至第11行每次从*mag*列表中选取一个影响力增益最大的节点加入到相应社区的种子集合中，并从相应社区寻找新的影响力增益最大的节点更新*mag*。第12行是返回影响力最大的种子集合。

2、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。

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: *node*, *gain*  1. *f = dfm\_gain (g, seed\_set)*  2. *node,gainthe node with max gain from f*  3. return *node,gain* |

*get\_node\_with\_max\_gain*算法输入有向图*G,* 种子集合*seed\_set,* 计算影响力增益的函数*dfm\_gain，*该函数是基于独立级联模型的求节点影响增益的函数。第一行是计算所有节点的影响力增益。第2行是获得影响力增益最大的节点及其影响增益。第3行是返回影响力增益最大的节点及其影响增益。

3、初始化*mag*列表的伪代码如表4-12。

表4-12 init\_mag算法伪代码

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

*init\_mag*算法输入子图列表*subg，*最后输出*mag*列表*。*第*1*行新建一个空列表*mag*。第2行至第3行对于每个子图在*mag*中添加一个元组，*mag[i]*元组用于存储[社区i的影响增益最大的节点，该节点的增益]。第5行返回mag列表。

4、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。

表4-13 arg\_max\_mag算法

|  |
| --- |
| Algorithm: *arg\_max\_mag*  input: *mag, seed\_set, subg*  output: *idx*  1.*new\_magempty list*  2.for each *idx* in *[0,1,2,…, len(mag)*  3. *nodemag[idx]*  4. *numsubg[idx].number\_of\_node*  6. *score = num / (len(seed\_set[idx])+1)*  7. *influ = int(node [0]+0.3)*  8. *new\_mag.append([idx, influ, score])*  9.*idx, influ, scoremax(new\_mag, key=lambda x : [x[1],x[2]])*  10.return *idx* |

*arg\_max\_mag*算法输入存放每个社区影响增益最大的节点及其增益的*mag*，当前的种子集合*seed\_set*，社区列表*subg，*最后输出*mag*中影响力增益最大的节点所在的索引*。*第1行是新建空列表*new\_mag*，该列表将会存放多个3元组，每个3元组的内容是[节点在*mag*的位置， 取整后的种子节点影响增益，社区的*score*值]。第2行至第8行对于*mag*的第*idx*个元素，根据公式(4.1)计算对应社区的得分，然后将[*idx*，*mag*[*idx*]存放的影响力增益，社区的*score*值]存入到*new\_mag*中。第9行获取new\_mag中影响一增益最大的索引，如果有多个，则选择*score*最大对应的那个索引。第10行返回索引。

4实验

4.1实验背景

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：处理器是Intel(R) Celeron(R) CPU 1000m @ 1.80GHz 1.80GHz。安装内存是4.00 GB。实验所用的算法均采用python语言实现。

为了验证算法的有效性，本文使用一个超市的一个月的记录的交易作为本实验的数据集。基于这个数据集使用apriori算法构建商品网络，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。

实验选择了带权出度排名和商品销量排名来衡量影响力最大化算法选取的商品的特点。

4.2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 | 肉松, 进口食醋 |
| 2 | 单反相机, 其它办公用品, 电热杯 |
| 3 | 砍骨刀, 漏勺/滤勺, 其它家居, 其它刀具 |
| 4 | 脱毛工具（美体/脱毛）, 须后水, 干燥剂, 雨披 |
| 5 | 婴儿牙胶, 其它身体护理用品 |
| 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** |
| 肉松 | 983 | 9 |
| 进口食醋 | 954 | 411 |
| 单反相机 | 1020 | 845 |
| 其它办公用品 | 1021 | 869 |
| 电热杯 | 1018 | 843 |
| 砍骨刀 | 904 | 366 |
| 漏勺/滤勺 | 883 | 10 |
| 其它家居 | 684 | 123 |
| 其它刀具 | 1022 | 523 |
| 脱毛工具  （美体/脱毛） | 1023 | 680 |
| 须后水 | 969 | 705 |
| 干燥剂 | 1004 | 2 |
| 雨披 | 1027 | 817 |
| 婴儿牙胶 | 922 | 180 |
| 其它身体护理用品 | 1030 | 847 |
| **平均排名** | **964** | **482** |

表5-7 独立级联模型中选取促销商品未给社区打分

|  |  |
| --- | --- |
| **社区编号** | **影响最大商品组合** |
| 1 | 烧烤用具 |
| 2 | 单反相机,电热杯,其它办公用品 |
| 3 | 其它家居,剪刀,其它刀具,西装衣架,米桶,漏勺/滤勺 |
| 4 | 干燥剂,脱毛工具（美体/脱毛）,雨披,免洗润发乳 |
| 5 | 烘焙咖啡 |
| 6 |  |
| 7 |  |

对比表5-5、表5-6和表5-7，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。

从表5-6知道影响力最大的商品的平均销量排名是很靠后，即选择促的促销商品是销量较少的商品。

（2）由表5-5和表5-6可知，每个社区先选择的商品都是出度排名靠前的商品，即对邻居的影响较强。随后为了考虑影响增益的问题，会选择出度排名较靠后的商品作为种子节点。

（3）对照表5-5和表5-7可知道，在基于独立级联模型的影响最大化算法中，根据每个社区选择种子个数情况对社区打分，可以让选取的种子节点在各个社区分布更均匀。

参考文献

后面补充