# 基于商品影响力分析的促销商品组合问题研究

袁国文，杨书新

（江西理工大学信息工程学院，江西 赣州 341000）

**摘 要**：选择促销商品的组合是购物篮分析中重要的内容。本文考虑到商品网络中商品影响传播与社交网络的独立级联模型存在一定相似性，提出基于独立级联模型影响力最大化的购物篮分析方法。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。最后我们将本文的算法用在真实数据中，实验结果验证了算法的合理性。

**关键词：**购物篮分析；商品网络；影响最大化

# Research on the combination of promotion product commodity based on the analysis of the influence of commodity

Yuan Guowen,Yang Shuxin

（School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China）

**Abstract:** Determining the combination of promotional merchandise is an important research topic in the market basket analysis. Because some similarities between commodities diffuse influence in the commodity network and the independent cascade model in social network, so this paper puts forward a market basket analysis method based on influence maximization in independent cascade model. Firstly, we construct a commodity network by using association rules. Secondly, we divide the network into many clusters. Thirdly, we get the commodity with most increase of influence from each clusters, and then, each time we select a commodity with most increase of influence from all clusters until special number of commodities have been selected. The algorithm we proposed are applied on real data, the results of experiment show the rationality of our algorithm.

**Key words**: market basket analysis; commodity network; influence maximization

超市经常会同时促销多种商品，这些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。

本课题的研究属于购物篮分析的范畴。现有的购物篮分析（Market Basket Analysis，MBA）可分为两类，基于非商品网络和基于商品网络的研究。基于非商品网络的研究的主要是关联规则的发现等，基于商品网络的研究主要是商品社区划分和度数分布等。通过查询国际会议收录的购物篮分析的相关文献，发现很少文献研究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。本文利用社交网络影响力最大化算法来求解选择促销商品的组合的问题，提出了一种基于商品影响力分析的促销商品组合问题的求解算法。

本文第1部分介绍购物篮分析中的相关工作。第2部分本文提出的基于商品影响力分析的促销商品组合问题的求解算法。第3部分是实验与分析。第4部分介绍基于社区划分的影响力最大化算法。第5部分介绍本文的实验。第6部分是总结。

1相关工作

现有的购物篮分析方法可分为两类，即非商品网络的购物篮分析和基于商品网络的购物篮分析。

基于非商品网络的购物篮分析主要是指商品关联性分析。1993年Agrawal等人[1]在购物篮分析提出了关联规则算法，该算法可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[3]，顾客行为分析[2]和餐馆菜单的选择[4]等。2013年Wiley Interdisc等人[5]使用MapReduce来提高购物篮分析效率。

在基于商品网络的购物篮分析方法中，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[6]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中心度与商品热销程度的关系等。2014年戚戚[7]在商品网络中使用层次聚类进行商品社团发现，并使用pagerank算法寻找重要商品。2015年Zhu Z等人[8]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。

目前购物篮分析领域仍较少文献研究选择促销商品的组合。通过分析购物者的购物行为，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，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*。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。商品的影响力可以定义为，一件商品能够影响购物者购买其他商品的能力，在独立级联模型中使用激活的商品数量来表示。

基于这样的商品影响传播行为，本文利用独立级联模型模拟商品影响传播，提出基于独立级联模型影响力最大化的购物篮分析方法，尝试寻找出影响力最大的商品组合，为零售商选择促销商品提供信息参考。

2 基于商品影响力分析的促销商品组合问题的求解算法

本算法包含三个重要组成部分。第1部分是使用apriori算法求取商品间的关联性并构建商品网络。第2部分是独立级联模型[9]下商品影响增益的计算。第3部分是使用基于社区划分的影响力最大化算法求解影响力最大的*k*件促销商品。

2.1 构建商品网络

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相互影响的关系，然后在网络中用有向边表示这个影响关系，边的权值表示商品间的影响强度。

其中apriori算法计算商品的支持度如式1所示，其中*T*是交易记录数，*num*(*A*)表示*A*商品在交易记录中出现的次数。计算商品间的关联性如式（2）所示。

（1）

（2）

* 1. 基于独立级联模型的商品影响增益计算

对于一个给定的有向的带权商品网络，*V*是商品节点集，*E*是边集合，是边的权值集合，表示从商品节点到商品节点的边的权重。用用于表示商品节点的入邻居节点，用用于表示商品节点的出邻居节点。假设预先选好的初始商品节点集合记为，商品节点的影响力记作。用表示在*t*轮被激活的商品节点。在独立级联模型中，第*t*轮()的每个节点对中未被激活的商品节点尝试以的概率激活一次。每个未被激活的商品节点在第*t*轮被激活的概率可用式（3）来计算。算法不断使用*t*-1轮的激活的节点尝试去激活未被激活的商品节点得到*t*轮激活的商品节点，直到没有新的节点被激活。初始的商品节点集合的影响力计算如式（4）。商品节点的影响增益的计算如式（5）。

(3)

(4)

(5)

2.3 基于独立级联模型影响力最大化的购物篮分析方法

本算法在使用有向图快速层次聚类的社区划分[11][12]基础上，给每个社区进行计分（*score*），score值的计算如式（5）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。社区计分基于这样的一个概念，社区的每个种子节点的影响的节点数量是相同的，且社区中的种子节点会影响所有的节点，如果社区选择的种子数越多那么平均每个节点影响的节点数就会变少，即社区的计分会变小。如果两个节点的影响增益相同，则选择社区计分大的种子节点，因为该种子节点会比社区计分小的种子节点在实际中潜在的影响力可能会更大。

(6)

本文所提出的影响力最大化算是一个贪心算法，算法的流程如图1所示。

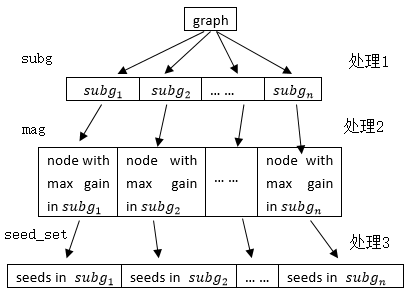


图1 贪心算法流程

处理1是使用有向图快速层次聚类算法将图划分成多个子图，记作*subg*。第*i*个子图用表示。处理2是指当*mag*的第*i*个位置的节点被选为种子节点，从子图中寻找影响增益最大的节点更新*mag*列表第*i*个位置存放的信息。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。

假设有一个7个节点11条边的有向带权商品网络如图2所示。以该商品网络为例子，寻找两个促销商品，这两个商品能尽可能的带动商品的销售。



图2 商品网络

假设图2的商品网络经过社区划分的结果如图3所示。



图3 社区划分结果

假设每个社区的影响增益最大的节点如表1所示。此时节点1和节点2拥有相同的影响增益，但节点1所属社区的社区计分更高，所以选择1号节点为种子节点。

表1 社区中影响增益最大节点

|  |  |  |  |
| --- | --- | --- | --- |
| **社区编号** | **节点编号** | **最大影响增益** | **社区****计分** |
| 1 | 1 | 2 | 4 |
| 2 | 4 | 2 | 3 |

重新从节点1所属的社区选择新的影响增益节点，可得到表2。此时4号节点影响增益最大，所以选择4号节点作为种子节点。此时已经选取了两个种子节点，即两个促销商品，流程结束。

表2 社区中新的影响增益最大节点

|  |  |  |  |
| --- | --- | --- | --- |
| **社区编号** | **节点编号** | **最大影响增益** | **社区计分** |
| 1 | 2 | 1 | 2 |
| 2 | 4 | 2 | 3 |

基于社区划分的影响最大化算法是*get\_seed\_set*，算法伪代码如表4-11所示。这个算法还包含三个其他算法，分别是*FHC*算法*、mag*算法*、arg\_max\_mag*算法和*dfm\_gain*算法。其中*FHC*算法是有向图快速层次聚类算法。*init\_mag*算法是初始化社区影响增益最大节点列表的算法。*arg\_max\_mag*算法是获取*mag*中影响增益最大节点在*mag*中的索引编号的算法。

表3 *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\_set*  4. for each *i* in[0,1,2…,*len*(*subg*)]  5. *seed\_set.append*([])  6.*i*0  7. while *i* < *k*  8. *idxarg\_max\_mag*(*mag, seed\_set, subg*)  9. s*eed\_set*[*idx*][0]*.append*(  *mag*[*idx*][0])  10. *seed\_set*[*idx*][1]*seed\_set*[*idx*][1] *+mag*[*idx*][1]  11. *subgthsubg*[*idx*]  12. *f = dfm\_gain* (*subgth,*  *seed\_set*[*idx*][0])  13. *mag*[*idx*] *select the node*  *with max gain from f*  14. *ii+*1  15.return *seed\_set* |

*get\_seed\_set*算法输入的参数有两个，有向图*G*和种子个数 *k*。最后输出影响力最大的*k*节点种子集合。第1行是使用有向图层次聚类算法对图*G*进行社区划分得到社区列表*subg*。第2行至第6行是初始化影响增益最大节点列表*mag*、种子集合*seed\_set*和记录循环次数的变量*i*。第7行至第12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*，其中第11行是计算独立级联模型中所有节点的影响力增益。第14行是获得影响力增益最大的节点及其影响增益，并存放至*mag*列表的第*idx*个位置。第15行是返回影响力最大的种子集合。

*init\_mag*算法用于初始化*mag*列，该算法的伪代码如表4。

表4 *init\_mag*算法伪代码

|  |
| --- |
| Algorithm: *init\_mag*  input: *subg*  output: *mag*  1. *mag*  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*中添加一个包含节点编号和最大影响增益的列表，*mag*[*i*]是第*i*个列表，第一个元素是社区*i*的影响增益最大的节点，第二个元素是该节点的影响增益。第5行返回*mag*列表。

*arg\_max\_mag*算法是获取*mag*中影响增益最大的节点在*mag*的索引编号，该算法的伪代码如表5。

表5 *arg\_max\_mag*算法

|  |
| --- |
| Algorithm: *arg\_max\_mag*  input: *mag, seed\_set, subg*  output: *idx*  1.*new\_mag* |

续表5 *arg\_max\_mag*算法

|  |
| --- |
| 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. *key=lambda x :* [*x*[1]*,x*[2]]  10.*idx, influ, scoremax*(*new\_mag, key*)  11.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行是创建一个新的存放影响增益最大节点的列表*new\_mag*，该列表将会存放多个3元组，每个3元组的内容包含节点在*mag*的位置，取整后的种子节点影响增益以及社区的*score*值。第2行至第8行对于*mag*的第*idx*个元素，根据式(6)计算相应的社区计分，然后将*idx*，*mag*[*idx*]存放的影响力增益和社区的*score*值存入到*new\_mag*中。其中第7行是将浮点型的影响增益强制转换为整型。第9行至第10行是优先选择*new\_mag*中影响增益最大的那个三元组，影响增益相同则选择社区打分最大的那个三元组。其中*key*是创建一个匿名函数，该函数获取输入的三元组的最后两个元素的内容，即影响增益和社区打分。第11行返回索引编号*idx*。

3实验及分析

3.1实验背景

实验所用的计算机配置如下：

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。

本文本实验的数据集是超市提供的一个月的交易记录。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1/179164，支持度均是0.001。求出商品的关联性后，构建网络。该商品网络含有的节点数是1039，边的数量是282192。

实验为了方便的对影响力最大化算法选取的促销商品的合理性进行分析，选择了带权出度排名、带权入度排名和商品销量排名来描述促销商品的特点。

3.2结果及分析

社区划分的结果如表6所示。

表6 社区划分结果

|  |  |  |
| --- | --- | --- |
| **社区** | **节点数** | **边数** |
| **1** | 258 | 42090 |
| **2** | 150 | 913 |
| **3** | 176 | 9613 |
| **4** | 208 | 14389 |
| **5** | 235 | 25964 |
| **6** | 8 | 20 |
| **7** | 4 | 12 |

算法选出影响力最大的15个商品所属的社区如表7所示，每行中商品的顺序是商品加入到该社区的影响力最大化商品集合的顺序。从表6和表7可知社区6和社区7的规模很小，商品的影响也因此变得比较小，所以没有从社区6和社区7选择促销商品。

表7 选取的商品组合

|  |  |
| --- | --- |
| **社区编号** | **影响最大商品组合** |
| 1 | 0501, 0513 |
| 2 | 0683, 0761, 0769 |
| 3 | 0987, 0917, 0317, 0911 |
| 4 | 0683, 0713, 0643, 0683 |
| 5 | 0318, 0389 |
| 6 |  |
| 7 |  |

选取的商品具有的特征如表8所示。从表8可以知道影响力最大的商品的平均销量排名和平局带权入度是很靠后的，而平均带权出度排名在中上位置的特点。结合表7和表8可以发现，每个社区最先选择的影响最大商品，均是该社区的影响最大商品中带权出度最靠前的，但带权入度排名是很靠后的。如商品0501带权出度排名是9，在社区1的促销商品中，带权出度排名最靠前，但带权入度排名963，带权入度排名很靠后。

表8 选取的商品特征

|  |  |  |  |
| --- | --- | --- | --- |
| **商品** | **销量排名** | **带权出度排名** | **带权入度排名** |
| 0501 | 983 | 9 | 963 |
| 0513 | 954 | 411 | 893 |
| 0683 | 1020 | 845 | 784 |
| 0761 | 1021 | 869 | 892 |
| 0769 | 1018 | 843 | 650 |
| 0987 | 904 | 366 | 748 |
| 0917 | 883 | 10 | 688 |
| 0317 | 684 | 123 | 939 |
| 0911 | 1022 | 523 | 925 |
| 0643 | 1004 | 2 | 950 |
| 0683 | 1023 | 680 | 938 |
| 0713 | 969 | 705 | 933 |
| 0683 | 1027 | 817 | 991 |
| 0318 | 922 | 180 | 826 |
| 0389 | 1030 | 847 | 1003 |
| **平均排名** | **964** | **482** | **875** |

本文提出的算法寻找出的商品具有平均销量排名和平均带权出度高，但平均带权出度排名低的特点，反映了选择的影响力最大的商品的销量较低和受到邻居的直接影响虽然较低，但是对邻居商品的直接影响强度大。

因为选取的商品受到邻居商品的直接影响强度低，即邻居商品的销售无法推动这些商品的销售，所以这些商品的销量低是合理的。销量低反映了促销的需要，销量低有可能会导致库存的囤积，以致库存成本的提高。如果将这些商品作为促销商品，可能可以帮助减少库存的压力。

选取的商品对邻居商品的直接影响强度大，即能够有效的影响其邻居商品，并通过邻居商品将影响扩散出去。促销这些商品很有可能会让购物者大量购买相关商品，进而带动其他商品的销售。因此本文提出的算法所选择的促销商品具有一定的合理性。

4 总结

本文采用独立级联模型对商品的影响传播进行建模，提出了基于商品影响力分析的促销商品组合问题的求解算法。在真实数据集中得到的实验结果验证了本文提出的算法的合理性。

在未来工作中会尝试使用其他的影响力传播模型对商品影响进行建模，并且探究不同时间下的商品影响的变化。

参考文献

[] Agrawal R. Mining association rules betw-een sets of items in large databases [J]. Acm Sigmod Record, 1993, 22(2):207-216.

[2] 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.

[3] Haughton, D., J. Deichmann, A. Eshghi, and S. Sayek, A review of software pac-kages for data mining[C]. The American Statistician, 2003 57 (4): 290-309.

[4] 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.

[5] Woo J. Market Basket Analysis algorithms with MapReduce[J]. Wiley Interdisciplin-ary Reviews Data Mining & Knowledge Discovery, 2013, 3(6):445–452.

[6] 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.

[7] 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.

[8] Zinoviev D, Zhu Z, Li K. Building Mini-Categories in Product Networks[J]. Studies in Computational Intelligence, 2015, 597(2):179-190.

[9] Kempe, David, Kleinberg, Jon, Tardos.et al. Maximizing the spread of influence throu-gh a social network[C] ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2003, 16(3):137-146.

[0] Chen W, Wang Y, Yang S. Efficient influ-ence maximization in social networks. In: KDD[J]. Proc of Acm Kdd, 2009, 61(2):

199-208.

[1] Leicht E A, Newman M E J. Community Structure in Directed Networks[J]. Physi-cal Review Letters, 2008, 100(11):2339-2340.

[2] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics Theory & Experiment, 2008, 30(2):155-168.