文章编号：2096-1472(2022)-04-17-04
DOI:10.19644/j.cnki.issn2096-1472.2022.004.004

基于关联规则对马铃薯及其制品在电商平台的销售分析

黄宇承[1,2]，吴丽丽[1]

(1.甘肃农业大学信息科学技术学院，甘肃 兰州 730070；
2.湖南都市职业学院，湖南 长沙 410137)
✉huangyu7630@sina.com; wull@gsau.edu.cn

摘 要：为充分挖掘电商平台中马铃薯及其制品的销售情况，在综合分析Apriori、DHP和FP-Growth算法原理的前提下，用Python语言实现上述三种算法，结果显示FP-Growth算法性能更佳。再通过FP-Growth算法挖掘出马铃薯及其制品在电商平台中销售情况的关联规则，得到一组强关联规则记录：马铃薯及其制品的月销售量和产品品种、品牌、产地、销售价格之间存在关联规则。根据强关联规则记录分析得出消费者在电商平台中对不同产地的马铃薯及其制品的购买趋势及兴趣度，为指导马铃薯及其制品的进一步销售和种植生产提供理论依据。

关键词：关联规则；电商平台；马铃薯及其制品；销售记录；兴趣度
中图分类号：TP391 **文献标识码**：A

Sales Analysis of Potatoes and Their Products on E-Commerce Platform based on Association Rules

HUANG Yucheng[1,2], WU Lili[1]

(1.*College of Information Science and Technology*, *Gansu Agricultural University*, *Lanzhou* 730070 *China*;
2.*Hunan Urban Professional College*, *Changsha* 410137 *China*)
✉huangyu7630@sina.com; wull@gsau.edu.cn

Abstract: In order to fully mine the sales of potatoes and their products on E-commerce platform, this paper proposes to comprehensively analyze and implement Apriori, DHP and FP-Growth algorithms. Results show that FP-Growth algorithm has a better performance. FP-Growth algorithm is used to explore the sales association rules of the sales of potatoes and their products on the E-commerce platform, and a set of strong association rule records are obtained: the association rules between the monthly sales volume of potatoes and their products, and the product variety, brand, origin and sales price. Based on the strong association rules, consumers' purchase trend and interest in potatoes and their products from different origins on the e-commerce platform are analyzed, which provides theoretical basis for directing further sales and planting of potatoes and their products.

Keywords: association rules; E-commerce platform; potatoes and their products; sales records; interest

1 引言(Introduction)

在电商平台销售农产品成为当前农产品流通的一种新型电子商务模式[1]。商务部的数据显示，2020 年上半年全国农产品网络零售额达1,937.7亿元，同比增长39.7%，比2019 年上半年增速高了6 个百分点。特别是近两年，更多的人愿意通过电商平台购买农产品。甘肃省是马铃薯及其制品的主要产地[2]，马铃薯及其制品是典型的特色农产品，通过电商平台销售是其主要的营销渠道之一。马铃薯及其制品在电商平台的销售使传统销售中受种植环境、保存条件、南北差异等因素影响而导致农民受损的情况得到改善，不仅使特色农产品的销售具有及时性和准确性，同时降低了销售成本和风险[3]。关联规则分析能挖掘出销售记录中与销售量相关联的属性和强关联规则记录，对指导马铃薯及其制品的种植和加工具有实际意义。

2 关联规则分析(Association rule analysis)

关联规则分析是发现大数据对象之间隐含的关联关系、相互影响，以及根据一(多)个事件的发生对另一(多)个事件所产生的反应[4]，通过现象发现本质，以便更好地为决

基金项目：民生科技专项基金项目；甘肃省科技计划项目(20CX9NA95).

策提供理论依据。针对马铃薯及其制品在电商平台的销售记录，一方面品种、品种规格、品牌、产地和单价、月销售量可以反映消费者的购买意向；另一方面可以反映马铃薯及其制品的生产地所产出的不同产品的销售量，可推断出不同产地的何种马铃薯有利于销售或指导种植生产。因此，本文采用Python语言运行关联规则算法Apriori算法、DHP算法和FP-Growth算法，从而比较三种算法中哪种算法运算时间最短；将运算时间最短的FP-Growth算法在马铃薯及其制品的销售数据集中运行得到频繁项集，并找出其月销售量和其他因素间的关联关系，以期促进特色农产品在电商平台中销售的良性发展，同时指导特色农产品的正确种植和加工。

特色农产品在电商平台销售得好坏与产品的品种、规格、品牌建设与推广、生产地及售价相关[5]。搜集淘宝、拼多多等常用电商平台中的马铃薯及其制品的销售数据，结合甘肃省农业科学院马铃薯研究所对马铃薯品种的研究，经过数据清洗后的马铃薯及其制品的部分销售数据如表1所示，通过关联规则分析得出月销售量与品种、规格、品牌、产地、销售价格之间的关系[6-7]。

表1 马铃薯及其制品在电商平台的销售记录(部分)

Tab.1 Sales records of potatoes and their products on E-commerce platform (part)

序号	品种	果皮/果肉	规格	品牌	产地	单价/元/斤	月销售量/份
1	陇薯5号	白肉	大果	禾果小镇	甘肃省定西市	1.98	1,027
2	陇薯5号	白肉	中果	无	甘肃省定西市	2.58	28
3	大西洋(ck)	黄皮/白肉	大果	无	甘肃省陇南市	4.36	587
4	陇薯7号	黄肉	小果	禾果小镇	甘肃省定西市	2.36	652
5	陇薯7号	黄肉	小果	鹿西小镇	甘肃省平凉市	2.32	3,000
6	陇薯15号	黄皮	中果	无	甘肃省金昌市	2.96	83
7	陇薯15号	黄皮	中果	无	甘肃省兰州市	3.98	568
8	L1192-4	红皮	中果	妃芸	甘肃省定西市	2.28	605

要进行关联规则分析，须经过两个过程：第一步，发现频繁项集；第二步，生成关联规则。首先找到频繁项集考虑的度量标准，主要有支持度、置信度。设马铃薯及其制品在电商平台的销售记录为数据集D，项集$S=\{S_1,S_2,S_3,\cdots,S_i\}$是马铃薯及其制品在电商平台销售记录属性的集合，包括品种、规格、品牌、产地、单价(元/斤)等，并且$S\subseteq D$。X是马铃薯及其制品在电商平台中的每条销售记录，$X\subseteq S$。假设S中有两个项集A和B，$A\subseteq X$时，销售记录X中必定包含A，则马铃薯及其制品在电商平台的销售有关联的规则可描述成如下公式：

$$A\Rightarrow B[\text{support}=s\%,\text{confidence}=c\%] \tag{1}$$

支持度计算公式如下：

$$\text{support}(A\Rightarrow B)=\frac{\text{count}(A\cup B)}{|D|}\times 100\% \tag{2}$$

置信度计算公式如下：

$$\text{confidence}(A\Rightarrow B)=\frac{\text{support}(A\cup B)}{\text{support}(A)}\times 100\% \tag{3}$$

关联规则$A\Rightarrow B$，要判断是否为强关联规则，必须同时满足两个条件：$\text{support}(A\Rightarrow B)\geqslant \min_\text{sup}$和$\text{confidence}(A\Rightarrow B)\geqslant \min_\text{conf}$，否则称之为弱关联规则。关联规则的阈值越大，表示关联性越强；反之，则关联性越弱。最小支持度和最小置信度可以根据实验需要进行设置。

3 不同的关联规则算法(Different association rule algorithms)

3.1 Apriori算法

Apriori算法采用逐层搜索的迭代方法[8]，对数据集D多次遍历，并且将每次遍历所得的频繁项集作为搜索项集，产生新的候选项集，对候选项集进行筛选，找到频繁项集，依次循环，直到没有找到更长的频繁项集为止。要找到最终频繁项集需完成两个步骤，即连接步和剪枝步，在频繁项集中找出强关联规则。

3.2 DHP算法

DHP算法是Apriori算法的优化，基本过程与Apriori相同，生效于Apriori算法的剪枝步过程中[9]。在第k次扫描时，生成每个事务的$k+1$项集，代入一个Hash函数中，生成一个Hash表，建立k项集的Hash表，同时记录每个桶中的元素个数。

当生成$Ck+1$时，将$Lk\times Lk$自连接产生的结果先代入上述Hash函数，若所落入该桶的计数小于最小支持阈值，则该元素必定不为频繁项集，故可以过滤掉，不放入$Ck+1$中。所有具有相同Hash值的项的总个数小于最小支持阈值，如：$\text{Hash}(A,B)=4$，$\text{Hash}(X,Y)=4$，不妨假设4号桶的元素个数小于最小支持阈值，则单个的(A,B)个数也必定小于最小支持阈值，故可排除。

3.3 FP-Growth算法

FP-Growth算法[10-11]巧妙地将树型结构引入算法中，它采取如下分治策略：提供频繁项集的数据库压缩到一棵频繁模式树(FP-Tree)，但仍保留项集关联信息。该算法和Apriori算法最大的不同有两点：

第一，不产生候选集。

第二，只需要两次遍历数据集，大大提高了效率。

现在对马铃薯在电商平台的部分销售情况通过FP-Growth算法进行关联规则分析，分析流程用表2的事务数据集D举例说明，用代号I描述马铃薯及其制品品种、规格、品牌、产地、价格、月销售量，假设最小支持度计数为2。

表2 事务数据集D

Tab.2 Transaction dataset D

编号	销售事务项	编号	销售事务项
S001	I_1,I_2,I_5	S006	I_2,I_3
S002	I_2,I_4	S007	I_1,I_3
S003	I_2,I_3	S008	I_1,I_2,I_3,I_5
S004	I_1,I_2,I_4	S009	I_1,I_2,I_3
S005	I_1,I_3		

FP-Growth算法对数据集*D*只需要扫描两次：

第一次扫描，先对事务数据集*D*的所有项进行支持度计数，若有最小支持度小于2的项集则删除。以支持度计数进行降序排序，得到频繁1-项集，如表3所示。

表3 频繁1-项集

Tab.3 Frequent 1-itemsets

1-项集	I_2	I_1	I_3	I_4	I_5
支持度计数	7	6	6	5	2

第二次扫描数据集*D*，构建FP树，如图1所示。

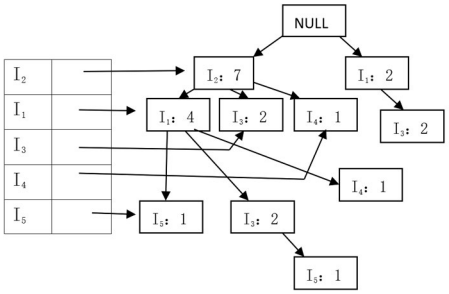

图1 构建FP树

Fig.1 Building FP tree

然后挖掘频繁项集：按照从下往上的顺序，首先考虑I_5，得到条件模式基<(I_2,I_1: 1)>，<(I_2,I_1,I_3: 1)>构造FP树，删除小于支持度的节点，形成单条路径后进行组合，得到I_5的频繁项集：{{I_2,I_5: 2}，{I_1,I_5: 2}，{I_2,I_1,I_5: 2}}。其次考虑I_4，得到条件模式基<(I_2,I_1: 1)>,<(I_2: 1)>构造条件FP树，得到I_4的频繁项集：{{I_2,I_4: 2}}。第三考虑I_3，得到条件模式基<(I_2,I_1: 2)>,<(I_2: 2)>,<(I_1: 2)>构造条件FP树，由于此树不是单一路径，需要递归挖掘I_3，从而得到I_1的条件模式基<(I_2: 2)>，I_1和I_3的条件模式基为<(I_2: 2)>构造条件FP树，得到I_3的频繁项集{{I_2,I_3: 4}，{I_1,I_3: 4}，{I_2,I_1,I_3: 2}}。最后考虑I_1，得到条件模式基<(I_2: 4)>构造条件FP树，得到I_1的频繁项集{I_2,I_1: 4}。

4 实验与结果分析(Experiment and results analysis)

4.1 三种算法性能比较

Apriori算法、DHP算法和FP-Growth算法的性能在数据集记录数固定的情况下与其运算速度息息相关[12]。在进行马铃薯及其制品在电商平台的销售数据的关联规则实验时，在1,000余条销售数据固定的情况下，设置置信度固定为80%，支持度有变化，分别设置为2%、4%、6%、8%、12%、14%、16%、18%、20%、22%，使三种算法在对马铃薯及其制品在电商平台的销售数据进行挖掘时，不同支持度下的运行时间发生变化，所花时间越少，则证明该种算法的效率越高，性能也越高。图2是Apriori、DHP和FP-Growth算法的运行时间，通过实验得出FP-Growth算法更优于Apriori算法和DHP算法。

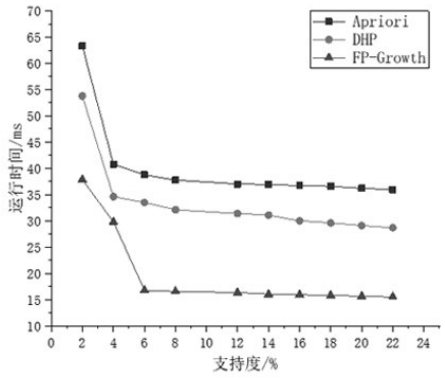

图2 三种算法运行时间比较

Fig.2 Comparison of running time of three algorithms

4.2 销售记录数据挖掘

采用三种算法中运行时间最短的FP-Growth算法对马铃薯及其制品在电商平台的销售数据集进行数据挖掘[13]，设置最小置信度为80%，最小支持度为20%，得到一组强关联规则记录：{品种,规格,品牌,产地,单价}，挖掘结果中部分置信度相对较大的记录如表4所示。

表4 马铃薯及其制品在电商平台的销售关联规则挖掘结果(部分)

Tab.4 Association rule mining results of potatoes and their products sold on E-commerce platform (part)

编号	关联规则	置信度/%
001	{品种=陇薯5号,规格=大果,品牌=禾果小镇,产地=甘肃省定西市,单价=1.98}⇒{月销售量=1027}	99.01
002	{品种=陇薯7号,规格=大果,品牌=乡隆,产地=甘肃省武威市,单价=1.98}⇒{月销售量=2249}	99.62
003	{品种=陇薯15号,规格=中果,品牌=无,产地=甘肃省兰州市,单价=3.98}⇒{月销售量=568}	94.60
004	{品种=L1192-4,规格=中果,品牌=妃芸,产地=甘肃省定西市,单价=2.28}⇒{月销售量=605}	93.87
005	{品种=陇薯7号,规格=小果,品牌=鹿西小镇,产地=甘肃省平凉市,单价=2.32}⇒{月销售量=3000}	99.43
006	{品种=陇薯7号,规格=小果,品牌=禾果小镇,产地=甘肃省定西市,单价=2.36}⇒{月销售量=652}	99.45
007	{品种=土豆粉,规格=宽粉,品牌=薯之梦,产地=甘肃省定西市,单价=10.64}⇒{月销售量=602}	93.85
008	{品种=大西洋(ck),规格=大果,品牌=无,产地=甘肃省陇南市,单价=4.36}⇒{月销售量=587}	94.23
009	{品种=土豆粉,规格=韭叶粉,品牌=无,产地=甘肃省兰州市,单价=6.60}⇒{月销售量=800}	95.31
010	{品种=土豆粉,规格=宽粉,品牌=无,产地=甘肃省临夏回族自治州,单价=5.80}⇒{月销售量=300}	94.36

根据挖掘结果得到甘肃省各地在电商平台中销售较好的马铃薯及其制品，如表5所示。通过表5间接反映各地销售较好的品种是该地大面积种植和加工的农产品，又表明

各地销售较好的品种受到消费者的喜爱。(1)陇薯15号以产地兰州市和定西市销售较好，但均无品牌，兰州市的价格在4.0 元/斤以下，定西市为1.0—4.0 元/斤。针对陇薯15号，可以加大品牌建设，定西市的价格跨度较大，兰州市的价格在4.0 元/斤以下，在保证马铃薯品质和低价稳定不变的同时将高价降低0.5—1.0 元/斤，从而提高市场竞争力。(2)L1192-4销量较好的是定西市，价格适中，但只有少量是有品牌的，可以加大品牌建设和推广。(3)陇薯5号销量较好的有定西市、平凉市、武威市，定西市的价格比平凉市和武威市低，但只有少量有品牌，平凉市有品牌，武威市无品牌，可提升定西市和武威市无品牌土豆的品牌机制，将价格调整至定西市的价格水平，同时保证产品品质。(4)陇薯7号是强关联规则记录中最多的，以定西市、平凉市、武威市销量较佳，定西市大部分有品牌，平凉市、武威市有品牌，价格属武威市最低，三市的价格差别不大，将该品种的土豆种植推广至周边其他市更有利于销售。(5)大西洋(ck)以定西市、平凉市、陇南市销量较好，定西市少数有品牌、平凉市有品牌，陇南市无品牌，其中定西市的价格最低，陇南市的价格最高，价格差最高达3.5 元/斤，在定西市对大西洋(ck)品种加大品牌推广力度的同时保持价格稳定，而在陇南市加大品牌推广力度的同时则适当降低价格。(6)土豆粉销售较好的有兰州市、定西市、临夏回族自治州、天水市，价格差较大，仅定西市大部分有品牌，质量难以把控，可在对其价格进行监督的同时对加工质量进行管控。(7)土豆片(薯片)仅兰州市销量较好，且建立有品牌机制，可在定西市等土豆产出较多的市增设加工厂，同时大力建立品牌机制。

表5 数据挖掘结果中各电商平台月销售较好的马铃薯及其制品

Tab.5 Potatoes and their products with better monthly sales on E-commerce platforms from the results of data mining

产地	品种	品牌	价格/元/斤
兰州市	陇薯15号	无	<4.0
兰州市	土豆粉	无	<7.0
兰州市	土豆片	有	<25
定西市	陇薯5号	少量有	<2.0
定西市	L1192-4	少量有	<2.5
定西市	陇薯7号	大部分有	<2.5
定西市	土豆粉	大部分有	<15.0
定西市	大西洋(ck)	少数有	1.0—2.5
定西市	陇薯15号	无	1.0—4.0
平凉市	陇薯7号	有	<2.7
平凉市	陇薯5号	有	<3.0
平凉市	大西洋(ck)	有	<3.0
临夏回族自治州	土豆粉	无	<6.0
陇南市	大西洋(ck)	无	<4.5
天水市	土豆粉	无	<16.5
武威市	陇薯7号	有	<2.0
武威市	陇薯5号	无	<3.0

4.3 关联规则兴趣度评估

消费者对农产品的购买意向和兴趣度相关，关联规则的兴趣度有正关联规则兴趣度和负关联规则兴趣度。判断消费者对购买马铃薯及其制品的兴趣度，求正关联规则即可，即马铃薯及其制品关联规则本身的置信度与它所包含的月销售量的交易支持度的差，其公式是：

规则的兴趣度=规则的置信度-月销售量的支持度

由表4挖掘结果中所得的置信度和所设置的月销售量的支持度之差，可以得出其兴趣度，结果如表6所示。通过兴趣度计算结果得出，强关联规则的销售记录消费者购买的兴趣度在70%以上，推断出消费者在后期购买马铃薯及其制品时大部分人会选择再次购买。

表6 消费者的购买兴趣度(部分)

Tab.6 Consumers' purchase interest (part)

编号	置信度/%	兴趣度/%
001	99.01	79.01
002	99.62	79.62
003	94.60	74.60
004	93.87	73.87
005	99.43	79.43
006	99.45	79.45
007	93.85	73.85
008	94.23	74.23
009	95.31	75.31
010	94.36	74.36

5 结论(Conclusion)

本文通过采用Python语言实现Apriori、DHP、FP-Growth三种算法，比较得出FP-Growth算法性能更优于另外两种算法。同时，采用性能更优的FP-Growth算法对马铃薯及其制品在电商平台的销售数据集进行关联规则分析，得到马铃薯及其制品在电商平台销售的强关联规则记录，将甘肃省各地销售较好的品种进行分析，以指导马铃薯及其制品的种植和加工，同时分析得出消费者购买的兴趣度在70%以上，由此可以推断大多数消费者的再次购买意向。

参考文献(References)

[1] 高会生,秦家瑞.互联网环境下的农产品营销模式探讨[J].中国集体经济,2021(23):62-63.

(下转第16页)