基于 RFA 模型和聚类分析的百度外卖客户细分

包志强[1] 赵媛媛[1] 赵 研[1] 胡啸天[1] 高 帆[2]
(西安邮电大学通信与信息工程学院 西安 710121)[1]
(航天科工集团第四研究院第九总体部 武汉 430040)[2]

摘 要 针对百度外卖行业具有的客户数量大、消费数据多、维度多等特点,提出一种基于客户消费行为视角的改进 RFM 模型。采用层次分析算法确定模型中各个变量的权重,并在此基础上采用 K-Means 聚类算法进行客户细分,计算确定客户对于商家的个人价值。数据分析结果表明,基于改进 RFM 模型的客户细分方法可以使商家对不同价值的客户采取针对性的策略。

关键词 百度外卖,改进 RFM 模型,K-Means 聚类,客户细分
中图法分类号 F270 **文献标识码** A

Segmentation of Baidu Takeaway Customer Based on RFA Model and Cluster Analysis

BAO Zhi-qiang[1] ZHAO Yuan-yuan[1] ZHAO Yan[1] HU Xiao-tian[1] GAO Fan[2]
(Department of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)[1]
(The Ninth Research Institute of the General Department of the Fourth, Aerospace Science and Technology Group, Wuhan 430040, China)[2]

Abstract In view of the characteristics of Baidu Take-out industry, such as large number of customers, large consumption data, high dimensions and so on, this paper proposed an improved RFM model based on perspective of customer consumption behavior, and uses the AHP algorithm to determine the weight of each variable in the model. K-Means clustering algorithm is used for customer segmentation, and the customer's personal value for the business is computed and determined. The results of data analysis show that the customer segmentation method based on the improved RFM model can make merchants adopt targeted strategies for customers with different values.

Keywords Baidu takeaway, Improved RFM model, K-Means clustering, Client subdivision

随着互联网技术的不断成熟,中国 O2O 行业迅速发展,互联网餐饮尤其是互联网外卖具有巨大的发展潜力。2010 年前后,第三方外卖平台如雨后春笋般纷纷涌现[1]。当前的外卖市场呈现出饿了么、美团外卖、百度外卖三足鼎立的局面,在不同城市、差异化细分市场的条件下,各平台都将面临着激烈的市场竞争。在这种条件下,外卖领域需要像传统的市场营销一样进行客户细分及客户价值分析。客户关系管理(Customer Relationship Management, CRM)[2]通过探索客户和商家之间潜在的关系来评估和维护客户关系,鉴于外卖服务竞争激烈以及发展速度过快,商家和平台尽可能地满足客户不断增长的个性化需求,针对不同价值的客户实行差异化服务;同时采取针对性的策略吸引客户,使其形成长期的购买行为,提升客户的忠诚度和持续购买能力,从而使平台或商家在激烈的市场竞争中立于不败之地。徐翔斌等[3]通过引入总利润属性,建立 RFP 模型,对电子商务客户做了细分分析;吴晓雪[4]将用户赎回行为考虑在 RFM 模型之内,对互联网金融平台做了用户细分研究;王召义等[5]在传统 RFM 模型的基础上引入顾客持续购买力,建立 RFT 模型,并制定了产品推荐算法。研究表明,RFM 模型及改进的 RFM 模型在不同领域均取得了不错的进展。

本文针对百度外卖行业具有的客户数量大、消费数据多、维度多等特点,提出一种基于客户消费行为视角的改进 RFM 模型,同时采用层次分析算法确定模型中各个变量的权重,并在此基础上采用 K-Means 聚类算法进行客户细分,计算确定客户的个人价值,最后达到对不同价值的客户分别采取针对性策略的效果。

1 传统 RFM 模型

在众多客户关系管理的客户分析模式中,RFM 分析是比较受欢迎的分析方法,是衡量客户价值的重要评价指标。传统的 RFM 模型最初由 Hughes 于 1994 提出,曾被广泛应用于直销领域,它包括 R(Recency),F(Frequency),M(Monetary) 3 个变量[6]。同时,基于 RFM 模型的客户关系管理分析也已经在市场中得到了广泛的应用[2]。R(Recency)表示最近一次购买时间,理论上最近一次购买时间越近的用户对提供即时商品或服务也最可能有反应,因此 R 越小越好。F(Frequency)表示消费者在某个时间段中的购买次数,经常购买的消费者越有意向再次购买,客户忠诚度高,因此 F 越大越好。M(Monetary)表示某个时间段中客户购买的总金额,购买金额越大,给企业带来的价值越大,因此 M 越大越好。但 F 和 M 之间存在强线性关系,影响最终客户价值分析的准确性,因此,按照传统 RFM 模型对百度外卖数据进行价值分

本文受陕西省教育厅专项科研计划项目(17JK0703)资助。
包志强(1978—),男,博士,副教授,主要研究方向为数据挖掘、大数据分析、导航抗干扰;**赵媛媛**(1996—),女,硕士生,主要研究方向为数据挖掘、大数据分析,E-mail:1732055344@qq.com。

析并不能达到理想的效果。鉴于此，本文从商家角度出发建立百度外卖 RFA(Recency Frequency Average_Monetary)模型。

2　RFA 模型

2.1　RFA 模型与 RFM 模型

国外学者认为客户细分模型的构建直接影响到数据挖掘技术[7]的准确性。模型描述得越准确，数据挖掘的效果就越好。本文基于百度外卖行业客户数量大、消费数据多、覆盖范围广的特点对传统 RFM 模型进行了适应性的改变，提出 RFA 模型，并将 RFA 模型与传统 RFM 模型做了比较，如表 1 所列。

表 1　传统 RFM 模型与百度外卖 RFA 模型中各指标含义的比较

模型	R(近度)	F(频度)	M/A(值度)
传统 RFM 模型	客户最近一次购买时间距离分析点的时间	客户一定时期内购买企业产品的次数	客户一定时期内购买企业产品的总金额
RFA 模型	客户最近一次订单时间距离分析点的时间间隔	客户在一定时间内的订单次数	客户在一段时间内的平均单次订单消费金额

R，F，A 是分析百度外卖客户订单交易数据所生成的二次特征，从不同角度反映了客户的购买行为，可被用来判断客户为百度外卖企业带来的价值，是衡量百度外卖客户价值的充分变量。

在 RFA 模型中，将 RFM 模型中的 M 指标替换为 A 指标，可以消除 RFM 模型中频度 F 与值度 M 之间存在的共线性问题，从而提了高模型的可靠性以及企业对客户价值判断的准确性。

2.2　RFA 权重分析

对于 RFA 各变量的指标权重问题，Hughes 于 1994 年提出应该同等看待 3 个指标，为其赋予相同的权重。Stone 于 1995 年对客户信用卡相关信息进行研究分析时，结合行业特殊性，认为 RFM 模型中的消费频率最为重要，其次是最近消费时间，最后是消费金额。本文认为针对不同的领域和行业，各个指标的权重应该存在差异，结合百度外卖行业的特殊性，采用层次分析法[8]来解决 RFA 的权重分析问题。应用层次分析法计算指标权重系数，是通过指标之间的两两比较对系统中各指标予以优劣评判，并利用这种评判结果综合计算各指标的权重系数。

首先，用 1－9 及其倒数作为标度对 3 个变量的相对重要性进行两两比较，以判断矩阵[9] $A=(a_{ij})_{n\times n}$。表 2 给出一个判断矩阵的例子[10]。

表 2　标度法判断矩阵的示例

	R	F	M
R	1	5	7
F	1/5	1	3
M	1/7	1/3	1

其次，进行一致性检验。

(1)计算一致性指标 CI(Consistency Index)

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{1}$$

其中，$\lambda_{\max}$为判断矩阵的最大特征值。

(2)查找一致性指标 RI(见表 3)

表 3　一致性指标

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14
RI	0	0	0.52	0.89	1.12	1.24	1.36	1.41	1.46	1.49	1.52	1.54	1.56	1.58

(3)计算一致性比例 CR(Consistency Ratio)

$$CR=\frac{CI}{RI} \tag{2}$$

当 $CR<0.1$ 时，认为判断矩阵是可接受的；若 $CR>0.1$，则应重新设计判断矩阵。

(4)计算 RFA 指标的相对权重

若$[W_R,W_F,W_A]=[0.122,0.346,0.532]$，则认为 A 的权重最大，即认为客户交易金额的大小是衡量客户价值高低最重要的因素。

基于以上对指标权重的判断，可以将客户价值定义为各指标与其相对应权重的乘积求和，即：

$$RFA=w_R\times R+w_F\times F+w_A\times A \tag{3}$$

3　数据的采集与整理

3.1　数据采集

从百度外卖某市肯德基商家清理数据，得到 2600 个客户半个月(2017 年 7 月 1 日至 2017 年 7 月 14)的 2601 个订单交易数据，从中统计出每个客户最近一次发生订单行为距离分析点(2017 年 7 月 15 日)的时间间隔(R)、时间段内的有效订单交易次数(F)和每个客户平均单次订单交易金额(A)。某客户的订单交易样本数据如表 4 所列。

表 4　某客户订单交易数据

ID	R	F	A
102505872	26.05	2	70.5

3.2　数据整理

数据规范化[11](归一化)处理是数据挖掘的一项基础工作。不同指标之间往往具有不同的量纲，数值间的差别也可能很大，不进行处理可能会影响到数据分析的结果。为了消除指标之间的量纲和取值范围差异的影响，需要对数据进行标准化处理，将数据指标按照比例进行缩放，使之落入一个特定的区域，以便于综合分析。

本文采用最小-最大规范化[11]方法对数据进行处理。最小-最大规范化也称为离差标准化，是对原始数据的线性变换，将数据映射到[0,1]。转换公式如下：

$$x^*=\frac{x-min}{max-min} \tag{4}$$

其中，max 为样本数据的最大值，min 为样本数据的最小值，$max-min$ 为极差。离差标准化保留了原样本数据中存在的关系，是消除量纲和数据取值范围的简单方法。

应用到 RFA 模型中即为：

$$R^*=\frac{R-\min(R)}{\max(R)-\min(R)} \tag{5}$$

$$F^*=\frac{F-\min(F)}{\max(F)-\min(F)} \tag{6}$$

$$A^*=\frac{A-\min(A)}{\max(A)-\min(A)} \tag{7}$$

4　数据分析与结果

采用 R，F，A 变量作为聚类变量，基于 R 语言，采用 K-

Means 聚类方法对数据进行聚类分析，将顾客细分为 5 种不同价值的群体。本文通过层次分析法将 R，F，A 的权重定义为 $w_R=0.072$，$w_F=0.279$，$w_A=0.649$，利用 $RFA=w_R\times R+w_F\times F+w_A\times A$，经过一系列 R 语言统计分析算法，得出具体的部分顾客价值分析结果，如表 5 所列。

表 5 RFA 模型的顾客价值分析结果

ID	R	F	A	W_RFA	LB	价值级别	N
1001	0.4523	0.4	0.4214	0.4392	1	1	755
1004	0.4660	0.4	0.3375	0.3642	4	2	111
1006	0.4037	0.2	0.2098	0.2211	3	3	802
1007	0.4080	0.2	0.1929	0.2031	2	4	480
1008	0.0331	0.2	0.1482	0.1544	5	5	453

从表 5 可以看出，价值最大的客户群体是第一类客户，包括 755 名顾客，占肯德基所有参与者的 29%。这类客户的订单频率高，单次订单消费金额大，可将其定义为肯德基商家的铂金顾客群，商家可以重点保持这类客户。

第二类最有价值的客户群体是第四类客户，包括 111 名客户，占肯德基所有参与者的 4%。这类客户的订单交易频繁，但平均单次订单金额不高，可将这类客户定义为肯德基商家的黄金客户群，将重点发展这类客户。

第三类客户群包括 802 名客户，占肯德基所有参与者的 31%。这类客户订单交易不太频繁，将其定义为肯德基商家的银质客户群，商家应重点培养这类客户，实施针对性策略，尽可能提升这类客户群的价值。

第四类客户群包括 480 名客户，占肯德基所有参与者的 18%。这类客户在 R 和 F 方面类似于第三类客户，区别在于这类顾客的单次消费金额较低，可将这类客户定义为肯德基商家的铜质客户群。

价值最低的客户群是第五类客户群，包括 453 名客户，占肯德基所有参与者的 17%。这类客户订单次数少，单次订单金额低，对商家价值低，可将其定义为肯德基商家的铁质客户群。

结束语 通过以上分析，综合 R，F，A 值对客户进行分类。以肯德基商家为例，从对商家的价值角度分析客户，能更全面地反映客户对商家的重要程度，可以辅助商家为不同价值的客户群体制定相对应的营销策略，例如商家可以对铂金价值的客户群体采取频繁下单打折、优惠券发放等行为，以提高客户满意度以及商家的业务盈利水平。

参 考 文 献

[1] 李雪苑. 浅谈第三方外卖平台的运营管理——以百度外卖为例[J]. 经贸实践，2016(8)：208.

[2] SONG M，ZHAO X，HAIHONG E，et al. Statistics-based CRM approach via time series segmenting RFM on large scale data[C]// International Conference on Utility and Cloud Computing. IEEE，2017：282-291.

[3] 徐翔斌，王佳强，涂欢，等. 基于改进 RFM 模型的电子商务客户细分[J]. 计算机应用，2012，32(5)：1439-1442.

[4] 吴晓雪. 基于 RFM 改进模型的互联网金融平台用户细分研究[D]. 北京：北京交通大学，2016.

[5] 王召义，汪琪. 基于改进 RFM 模型的产品推荐算法[J]. 宿州学院学报，2016，31(11)：101-104.

[6] 何敏，张洪伟，张波. 模糊 ISODATA 及在 CRM 中的应用[J]. 计算机应用，2005，25(6)：1455-1457.

[7] HAN J W，KAMBER M. Data mining：Concepts and techniques[M]. 北京：机械工业出版社，2002.

[8] 耿俊成，袁少光，万迪明，等. 基于改进 RFM 模型的电力客户缴费渠道分析预测[J]. 电力信息与通信技术，2017，15(8)：55-59.

[9] 邓雪，李家铭，曾浩健，等. 层次分析法权重计算方法分析及其应用研究[J]. 数学的实践与认识，2012，42(7)：93-100.

[10] 刘朝华. 基于客户价值的客户分类模型研究[D]. 武汉：华中科技大学，2008.

[11] 张良均，云伟标，王路，等. R 语言数据分析与挖掘实战[M]. 北京：机械工业出版社，2015：46-47.

(上接第 421 页)

参 考 文 献

[1] 文继军，王珊. SEEKER：基于关键词的关系数据库信息检索[J]. 软件学报，2005，16(7)：1270-1281.

[2] 张阔，李涓子，吴刚，等. 基于关键词元的话题内事件检测[J]. 计算机研究与发展，2009，46(2)：245-252.

[3] 李峰，黄金柱，李舟军，等. 使用关键词扩展的新闻文本自动摘要方法[J]. 计算机科学与探索，2016，10(3)：373-380.

[4] 吴舜尧，邵峰晶，王金龙，等. 融合语义资源和关键词的文本聚类[J]. 计算机工程，2014，40(4)：223-227.

[5] VIDAL M，MENEZES G V，BERLT K，et al. Selecting Keywords to Represent Web Page Using Wikipedia Information[J]. WebMedia，2012，4(10)：15-18.

[6] TURNEY P D. Learning Algorithms for Keyphrase Extraction[J]. Information Retrieval，2000，2(4)：303-336.

[7] BELLAACHIA A. NE-Rank：A Novel Graph-based Keyphrase Extraction in Twitter[J]. Web Intelligence and Intelligent Agent Technology，2013，1(12)：372-379.

[8] 李然，张华平，赵燕平，等. 基于主题模型与信息熵的中文文档自动摘要技术研究[J]. 计算机科学，2014，41(S2)：298-300.

[9] 刘通. 基于复杂网络的文本关键词提取算法研究[J]. 计算机应用研究，2016，33(2)：365-369.

[10] 陈伟鹤，刘云. 基于词或词组长度和频数的短中文文本关键词提取算法[J]. 计算机科学，2016，43(12)：50-57.

[11] 王立霞，淮晓永. 基于语义的中文文本关键词提取算法[J]. 计算机工程，2012，38(1)：1-4.

[12] 李鹏，王斌，石志伟，等. Tag-TextRank：一种基于 Tag 的网页关键词抽取方法[J]. 计算机研究与发展，2012，49(11)：2344-2351.

[13] 罗燕，赵书良，李晓超，等. 基于词频统计的文本关键词提取方法[J]. 计算机应用，2016，36(3)：718-725.

[14] 李晓超，赵书良，罗燕，等. 中文文本同频词统计规律及在关键词提取中的应用[J]. 计算机应用研究，2016，33(4)：1007-1012.

[15] 潘虹，徐朝军. LCS 算法在术语抽取中的应用研究[J]. 情报学报，2010，29(5)：853-857.

[16] 车海燕，冯铁，张家晨，等. 面向中文自然语言文档的自动知识抽取方法[J]. 计算机研究与发展，2013，50(4)：834-842.

[17] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术，2013，29(9)：30-34.

[18] 方康，韩立新. 基于 HMM 的加权 TextRank 单文档的关键词抽取算法[J]. 信息技术，2015，4(4)：114-116.

[19] 顾益军. 融合 LDA 与 TextRank 的关键词抽取研究[J]. 现代图书情报技术，2014，30(7)：41-47.

[20] BENGIO Y，DUCHARME R，VINCENT P，et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research，2003，3(6)：1137-1155.