文章编号:1007-6735(2012)01-0098-05

# 基于关联规则的房地产广告效果反馈分析

王正友[1], 刘 倩[2]
(1.上海出版印刷高等专科学校 出版与传播系,上海 200093;2.上海理工大学 出版印刷与艺术设计学院,上海 200093)

**摘要**:提出了将数据挖掘技术运用于广告媒体选择的观点,并运用关联规则对房地产公司决策型关系数据库进行广告效果的数据分析,从而获得有价值的信息.为解决现阶段房地产广告效果的定量分析和寻找高性价比的广告模式提供一定的理论基础和现实指导.

**关键词**:数据挖掘;关联规则;媒体选择
**中图分类号**:TP 391 **文献标志码**:A

# Feedback Analysis of Real Estate Advertising Effectiveness Based on Association Rules

WANG Zheng-you[1], LIU Qian[2]
(1. *Department of Publication and Communication, Shanghai Publishing and Printing College, Shanghai* 200093, *China*;
2. *College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai* 200093, *China*)

**Abstract**: The idea of using data mining for advertising media selection was brought forward and the association rules were adopted in data analysis of relational database of real estate company to obtain valuable information for advertising performance inspection. A theoretical basis and practical guidance were provided to address the quantitative analysis on present real estate advertising effects and to find cost-effective advertising model.

**Key words**: *data mining*; *association rules*; *media selection*

目前安排房地产项目的广告预算仍只能凭借经验,缺乏广告投入与广告效果之间的定量分析,而这样的方式存在很大的风险隐患.应该认清广告成本与广告效果之间的关系,使广告投放有效性更强,以此来节约成本.

广告作为营销的一种方式,一直受到广泛的关注,但是,相对于广告传播媒介及其方式的快速转变和发展,广告营销渠道的研究就显得有些落后于时代变化.广告行业有这样一句名言"我知道我的广告费至少浪费了一半以上,但我不知道究竟浪费在哪里".房地产企业在投放广告时,由于没有根据自身项目性质和目标客户特征来选择适当的媒体,使得房地产有效信息的传播渠道过窄.只有找出房地产广告营销的合理取舍因素、揭示媒体选择"合理性"及未来走向,才能为房地产广告的进一步合理选择找到切实可行的路径.

广告媒体选择方面的研究一直引起国内外广泛的关注[1-4].传统的房地产广告媒体投放与客户反馈效果分析是通过建立表格分析广告的到达率.这种分析的工作量非常巨大,但却不精确,也很难找出数据之间的关联关系.在数字媒体时代,可以通过建立房地产广告投放数据库,应用数据关联规则,对房

**收稿日期**:2010-05-28
**作者简介**:王正友(1959-),男,副教授.研究方向:计算机软件应用、多媒体设计与制作等.E-mail:wangzy59@163.com

地产广告数据进行科学和定量的分析. 借助计算机程序，分析广告媒体与客户之间的对应关系，提出有效而精准的营销模式方案，为解决现阶段房地产营销提供一定的理论基础和现实指导.

## 1　数据库建立

### 1.1　数据挖掘

数据关联是数据中存在的一类重要的可被发现的知识. 若两个或多个变量的取值之间存在某种规律性，就称为关联. 关联规则是数据挖掘中一种主要的挖掘技术，最近几年被业界广泛研究. 数据挖掘就是从大型数据库的数据中提取人们感兴趣的知识. 这些知识是隐含的、事先未知的潜在有用信息，提取的知识表示为概念、规则、规律及模式等形式[5].

关联规则的挖掘问题可形式化描述如下：

设 $I=\{i_1,i_2,\cdots,i_m\}$是二进制文字的集合，其中的元素称为项(item). 记 $D$ 为交易 $T$ 的集合，这里交易 $T$ 是项的集合，并且 $T\subseteq I$. 对应每一个交易有唯一的标识，如交易号，记作 TID. 设 $X$ 是一个 $I$ 中项的集合，如果 $X\subseteq I$，那么称交易 $T$ 包含 $X$. 一个关联规则是形如 $X\Rightarrow Y$ 的蕴涵式，这里 $X\subset I$，$Y\subset I$，并且 $X\Rightarrow Y=\Phi$。规则 $X\Rightarrow Y$ 在交易数据库 $D$ 中的支持度(support)是交易集中包含 $X$ 和 $Y$ 的交易数与所有交易数之比，记为

$$\text{support}\left(\bigcup\nolimits_k L_k\right)=|\{T:X\cup Y\subseteq T,T\in D\}|/|D|$$

规则$X\Rightarrow Y$ 在交易集中的可信度(confidence)是指包含 $X$ 和 $Y$ 的交易数与包含 $X$ 的交易数之比，confidence($X\Rightarrow Y$)，即

$$\text{confidence}(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|\{T:X\subseteq I,T\in D\}|$$

给定一个交易集 $D$，挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度(min-supp)和最小可信度(min-conf)的关联规则[6].

### 1.2　数据库构建

决策型关系数据库中的属性分为条件属性和决策属性. 每条记录本身就是一条规则，但每条规则中并非所有的条件属性都是必要的，有些是多余的. 对数据库的挖掘实际上就是要去掉多余的没必要的规则，得到更简洁直观的关联规则. 现以某房地产开发企业作为研究对象，问卷中涉及属性包括：日期、客户姓名、联系电话、年龄、居住区域、年收入、第几次置业、购房原因、认为单价是否合理、了解渠道、吸引原因、工作地点、工作单位、职业、需求面积、需求房型、关注焦点、购房意向. 本文主要研究媒体渠道，假定决策型关系数据库中有 3 个条件属性，分别为购房客户的年龄 $C_1$、居住地 $C_2$、工作单位 $C_3$，因为这 3 个条件属性都和精准广告的选择比较相关，可以直接为精准广告提供数据支持. 决策属性 $D$ 为购房者获取本企业楼盘信息的渠道(即不同的广告媒体). 本文将对决策型关系数据库进行关联规则挖掘，找出条件属性与决策属性之间的关联规则.

通过对来人来电、网络信息反馈、客户反馈表等原始数据的分析，可以对广告受众有一个基础的了解. 而对受众各项单一特征的分析，更有助于房地产商针对楼盘定位、楼盘风格及楼盘卖点等进行决策. 但是，这样的分析相对是单一特征的，并且这样的数据保留和处理都不是最佳的选择. 所以，本文研究是在以上原始客户反馈信息的基础上，构建网络数据库，对原始数据进行有效的维护，便于数据保留和处理. 并且在此基础上建模，使用 Apriori 算法[7]进行数据挖掘，用以寻求更有价值的双项甚至多项特征之间的关联，从而获得更多更有价值的信息，为房地产广告营销提供有效的资讯.

关系数据库一般可分为无决策型和决策型. 无决策型数据库是指所建数据库中数据价值平等，不含决策属性. 决策型关系数据库则分为条件属性和决策属性.

现以某楼盘广告作为研究对象，其企业楼盘信息渠道(广告媒体选择)就是其决策属性 $D$，而其余的各属性可以作为条件属性. 本文选取其中比较重要的 3 个条件属性，分别为客户年龄段 $C_1$、居住区域 $C_2$、工作单位 $C_3$，并将广告商选择的投放媒体和客户反馈所得到的客户信息存储到数据表 ASSAY_DATA. mdb 中，图 1 是数据库存放数据表 ASSAY_DATA 的数据库结构.

| | 字段名称 | 数据类型 | |
|---|---|---|---|
| ⚷ | 编号 | 自动编号 | |
| | RECORD_DATE | 文本 | 日期 |
| | CUTOMER_NAME | 文本 | 客户姓名 |
| | CUSTOMER_TEL | 文本 | 联系电话 |
| | CUSTOMER_AGE | 文本 | 年龄 |
| | CUSTOMER_AREA | 文本 | 居住区域 |
| | HOUSE_TIMES | 文本 | 第几次置业 |
| | HOUSE_PURPOSE | 文本 | 购房原因 |
| | PRISE_REASON | 文本 | 认为单价是否合理 |
| | KNOW_WAYS | 文本 | 了解渠道 |
| | INTREST_REASON | 文本 | 吸引原因 |
| | JOB_AREA | 文本 | 工作地点 |
| | JOB_COMPANY | 文本 | 工作单位 |
| | JOB_POSTION | 文本 | 职位 |
| | HOUSE_AREA | 文本 | 需求面积 |
| | HOUSE_TYPE | 文本 | 需求房型 |
| | HOUSE_FOCUS | 文本 | 关注焦点 |
| | BUY_ORDER | 文本 | 购房意向 |

**图 1　ASSAY_DATA 的数据库结构**

**Fig.1　ASSAY_DATA's database structure**

## 2 数据关联程序设计

### 2.1 语言选择及数据库选择

程序的开发工具:Microsoft Visual Studio 2008是一套完整的开发工具集,用于生成 ASP. NET Web 应用程序、XML Web Services、桌面应用程序和移动应用程序. Visual Basic、Visual C++、Visual C#和 Visual J#全都使用相同的集成开发环境. 选用 Visual Studio 2008 的优势在于其能够快速创建用户体验丰富而又紧密联系的应用程序,这使得采集和分析信息变得简单便捷,业务决策也更为有效.

开发语言:C#. C#是微软公司发布的一种面向对象的、运行于. NET Framework 之上的高级程序设计语言. C#的门槛要比 VC 和 Java 的低,相对比较容易学习操作,其兼容性也比较强.

数据库:Microsoft Office Access. Access 是由微软发布的关联式数据库管理系统,是 Microsoft Office的成员之一. 另外,Access 还是 C 语言的一个函数名和一种交换机的主干道模式,其优势包括存储方式单一、面向对象、界面友好、易学习、集成环境、可处理多种数据信息. Access 还可以将程序应用于网络,并与网络上的动态数据相联接,利用数据库访问页对象生成 HTML 文件,轻松构建 Internet/Intranet 的应用.

挖掘关联规则算法:Apriori 算法. Apriori 最早由 Agrawal 等人提出,是一种宽度优先算法,也是目前一种最有影响的挖掘布尔关联规则频繁项集的算法,其基本思想是通过对数据库的多次扫描来计算项集的支持度,发现所有的频繁项集,从而生成关联规则,其核心是基于两阶段频集思想的递推算法. 该关联规则在分类上属于单维、单层、布尔关联规则. 在这里,所有支持度大于最小支持度的项集称为频繁项集,简称频集.

利用 Apriori 算法的原理,首先找出所有的频集,这些项集出现的频繁性至少和预定义的最小支持度一样;然后由频集产生强关联规则,这些规则必须满足最小支持度和最小可信度;最后使用第一步找到的频集产生的期望规则,产生只包含集合的项的所有规则,其中,每一条规则的右部只有一项. 一旦这些规则被生成,只有那些大于用户给定的最小可信度的规则被留下来. 使用递推方法,生成所有频集.

### 2.2 程序框架

数据关联分析程序流程图如图 2 所示.

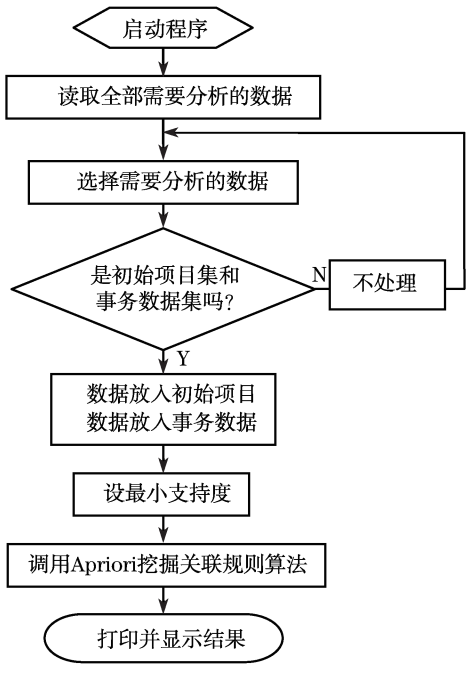


**图 2 关联分析程序框架**

**Fig. 2 Framework of association rules program**

**a.** 框图中读取全部需要分析的数据:本文中 ASSAY_DATA 数据库中 280 条来人客户信息表数据.

**b.** 选择需要分析的数据:是需要比较的重点数据列,问卷中涉及属性包括日期、客户姓名、联系电话、年龄、居住区域、第几次置业、购房原因、认为单价是否合理、了解渠道、吸引原因、工作地点、工作单位、职业、需求面积、需求房型、关注焦点和购房意向. 本文主要研究媒体渠道,假定决策型关系数据库中有 3 个条件属性,分别为购房客户的年龄 $C_1$、居住地 $C_2$、工作单位 $C_3$. 这 3 个条件属性都和精准广告的选择比较相关,可以直接为精准广告提供数据支持.

**c.** 初始项目集:为选择的数据列的全部组合. 本文列举的是年龄、居住区域和工作单位的全部组合(如短消息与 30～34 岁、短消息与杨浦区、短信息与国营/集体企业等).

**d.** 事务数据集:所列项的不重复集合.

**e.** 设定最小支持度:最小支持度可选范围为 $0<n<280$,当总份数为 280 份,且运算结果大于 $n$ 时显示,本文为了最大限度地显示所有可能性,设定 $n=0.1$.

**f.** Apriori 挖掘关联规则算法.

输入：数据库 ASSAY_DATA，最小支持度阈值 $n$.

输出：ASSAY_DATA 中的频繁集 $L$.

```
L1 = find_frequent_1 - itemsets(D);
                                        //频繁项集
for(k = 2;Lk-1≠Φ;k ++ ){
    Ck = Apriori_gen(Lk - 1);    //调用函数
         Apriori_gen(Lk - 1)通过频繁(k - 1)
         项集产生候选 k 项集
    for each transaction t ∈D { //所有数据集
        Ct = subset(Ck,t);// t 包含的候选集
        for each candidate c ∈ Ct{
                                        //所有候选集
            c.count ++ ;
            Lk = {C ∈ Ck | c.count>= n}
        }
    }
    return L1∪L2∪Lk…∪Lm;
}
```

## 3　数据关联分析

在数据库中构建数据表 ASSAY_DATA 用以存放记录，总数 280 条，运用 Apriori 挖掘关联规则算法寻找决策属性 $D$ 与条件属性 $C$ 之间的关联关系，通过数据挖掘获得年龄 $C_1$ 与决策属性 $D$ 为购房者获取本企业楼盘信息的渠道关系数据、居住地 $C_2$ 与决策属性 $D$ 为购房者获取本企业楼盘信息的渠道关系数据、职业 $C_3$ 与决策属性 $D$ 为购房者获取本企业楼盘信息的渠道关系数据.

**a.** 年龄 $C_1$ 与决策属性 $D$ 为购房者获取本企业楼盘信息的渠道关系，如表 1 所示. 表 1 表明，年龄在 35～44 岁的受众通过《新闻晨报》前来的比较多，搜房网对年龄在 30～44 岁的受众有明显的优势，手机短消息对各个年龄段的受众都比较有效，《上海楼市》这种专业性的杂志对于 30～49 岁的受众更有优势.

**b.** 居住地 $C_2$ 与决策属性 $D$ 为购房者获取本企业楼盘信息的渠道关系，如表 2 所示. 表 2 表明，各类媒体主要能吸引到的受众多为杨浦、虹口、宝山和浦东这 4 个区. 本文的楼盘位于杨浦区，关联关系的分析符合实际的情况. 对于报纸、杂志这类大众媒体，我们并不能选择发送的范围，而反馈率显示主要人群却在这 4 个区中，报纸、杂志在其它区域的很大一部分投放就比较无效，成本也有所浪费. 而对于手机短信彩信或者 DM 广告（直接邮送广告），则可以更好地选择投放的区域.

**表 1　年龄 $C_1$ 与决策属性 $D$ 的反馈人数比**

**Tab.1　Feedback population ratio of age and decision parameter**

| 年龄 | 媒　体 | | | | | |
|---|---|---|---|---|---|---|
| | 《新闻晨报》 | 搜房网 | 短消息 | 《上海楼市》 | 《新民晚报》 | 邮寄广告 |
| 30 岁以下 | 4/280 | 5/280 | 0/280 | 1/280 | 0/280 | 0/280 |
| 30～34 岁 | 2/280 | 14/280 | 7/280 | 5/280 | 3/280 | 3/280 |
| 35～49 岁 | 15/280 | 6/280 | 7/280 | 7/280 | 2/280 | 3/280 |
| 40～44 岁 | 15/280 | 6/280 | 5/280 | 0/280 | 2/280 | 1/280 |
| 45～49 岁 | 6/280 | 1/280 | 3/280 | 0/280 | 0/280 | 1/280 |
| 50～54 岁 | 2/280 | 1/280 | 5/280 | 0/280 | 0/280 | 0/280 |
| 55～59 岁 | 0/280 | 0/280 | 0/280 | 0/280 | 0/280 | 0/280 |

**表 2　居住地 $C_2$ 与决策属性 $D$ 的反馈人数比**

**Tab.2　Feedback population ratio of area and decision parameter**

| 居住地 | 媒　体 | | | | | |
|---|---|---|---|---|---|---|
| | 《新闻晨报》 | 搜房网 | 短消息 | 《上海楼市》 | 《新民晚报》 | 邮寄广告 |
| 杨浦 | 31/280 | 19/280 | 14/280 | 7/280 | 5/280 | 6/280 |
| 虹口 | 1/280 | 4/280 | 5/280 | 1/280 | 1/280 | 1/280 |
| 宝山 | 7/280 | 3/280 | 2/280 | 3/280 | 0/280 | 1/280 |
| 长宁 | 0/280 | 0/280 | 1/280 | 0/280 | 0/280 | 0/280 |
| 浦东 | 5/280 | 5/280 | 0/280 | 2/280 | 2/280 | 0/280 |
| 闸北 | 0/280 | 1/280 | 0/280 | 0/280 | 0/280 | 0/280 |

**c.** 工作单位 $C_3$ 与决策属性 $D$ 为购房者获取本企业楼盘信息的渠道关系，如表 3 所示. 表 3 表明，各种职业类型的受众对于手机短信彩信之类的媒体都比较能够接受，其反馈人数相对比较均匀. 而对于《新闻晨报》、《新民晚报》、《上海楼市》这种有一定阅读性和专业性的报纸杂志，其受众就会有所差别，比如私营/民营企业职员相对比较少.

**表 3　工作单位 $C_3$ 与决策属性 $D$ 的反馈人数比**

**Tab.3　Feedback population ratio of work-unit and decision parameter**

| 工作单位 | 媒　体 | | | | | |
|---|---|---|---|---|---|---|
| | 《新闻晨报》 | 搜房网 | 短消息 | 《上海楼市》 | 《新民晚报》 | 邮寄广告 |
| 国营/集体企业 | 2/280 | 6/280 | 1/280 | 0/280 | 1/280 | 1/280 |
| 事业单位 | 5/280 | 4/280 | 2/280 | 3/280 | 1/280 | 0/280 |
| 私营/民营企业 | 1/280 | 2/280 | 2/280 | 0/280 | 0/280 | 2/280 |
| 外资/合资企业 | 4/280 | 5/280 | 1/280 | 4/280 | 3/280 | 1/280 |
| 政府机关 | 5/280 | 3/280 | 1/280 | 1/280 | 0/280 | 0/280 |

## 4 结 论

通过对Apriori算法的优化能有效挖掘出人们真正感兴趣的、有价值的信息.但在处理大数据集时,Apriori算法由频繁 $k-1$ 项集进行自连接生成的候选频繁 $k$ 项集数量巨大;在验证候选频繁 $k$ 项集时需要对整个事物数据库进行扫描,非常耗时[8].如何有效地提高算法执行的效率是值得进一步研究的问题.

数据库构建在房地产广告渠道精准选择中起到了非常重要的作用,它提供了海量数据存放、数据分析、目标客户整合定位,为最终的广告投放提供了强大支持.

虽然数据表ASSAY_DATA存放的记录总数只有280条,但在广告媒体投放的受众反馈中,由于用到了数据关联技术,得到的数据挖掘分析相对于人工分析,工作量大大减轻,使数据分析精确和效果反馈定量化,尤其在数据量信息增多时,优势也更明显.

本文的数据关联原理、数据库构建和计算机程序设计方法同样也可用于其它类型的广告效果反馈分析.

**参考文献:**


[1] 雷小平,蔡楚纯.基于消费者偏好的媒体选择策略市场研究[J].市场研究,2009,57(1):33-37.

[2] 吴建伟,唐万生,郝占刚.广告媒体选择的DEA模型研究[J].西北农林科技大学学报(社会科学版),2005,5(1):89-92.

[3] 王雪梅,江文斌,冯源源.基于目标规划的广告媒体选择问题研究[J].商场现代化,2009,38(7):87.

[4] 陈立.试论广告策略及其在营销中的运用[J].中国电力教育,2009(8):250-251.

[5] Fayyad U M, Piatetsky-Shapiro G, Smyth P, et al. Advances in knowledge discovery and data mining[M]. Cambridge:MIT Press,1996.

[6] 杨洪涛,李桂君.关联规则在房地产广告媒体选择中的应用[J].计算机工程与运用,2006,42(5):230-232.

[7] Agrawal R, Imielinski T, Swami A N. Mining association rules between sets of items in large databases[C]// Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. Washington: ACM Press,1993.

[8] 黄进,尹治本.关联规则挖掘的Apriori算法的改进[J].电子科技大学学报,2003,32(1):76-79.