

# 基于内容偏好的移动客户互联网访问行为分析

宫婧<sup>1,2</sup>, 周飞飞<sup>1</sup>, 吕佳<sup>1</sup>, 刘陈伟<sup>1</sup>

(1. 南京邮电大学理学院, 江苏南京 210046;

2. 南京邮电大学物联网学院, 江苏南京 210046)

**摘要:**文中主要研究基于内容偏好的移动客户互联网访问行为。首先,进行移动客户互联网访问偏好内容的细分,对原始数据进行剔除噪声数据及转化数据为布尔类型的预处理。其次,基于改进的ORAR关联规则算法分析布尔类型的数据,挖掘出各偏好间的关联程度。最后,根据移动公司提供的4万条移动客户互联网访问的随机数据对改进的ORAR关联规则算法进行验证,得到各个子偏好间的关联程度且改进的算法使结果更加准确。结果表明,根据关联程度能够有效命中目标客户群的偏好访问内容,借此向特定用户推销相关的业务,从而达到精确营销、获得最大客户满意度的要求。

**关键词:**内容偏好;访问行为分析;改进ORAR关联规则算法

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2012)12-0149-04

## Analysis of Clients Internet Access Behavior of China Mobile Based on Content Preference

GONG Jing<sup>1,2</sup>, ZHOU Fei-fei<sup>1</sup>, LÜ Jia<sup>1</sup>, LIU Chen-wei<sup>1</sup>

(1. College of Science of Nanjing University of Posts and Telecommunications, Nanjing 210046, China;

2. College of Internet of Things of Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

**Abstract:** It is studying the analysis of the clients internet access behavior of China mobile based on the content preference. Firstly, subdivide the content preferences of mobile clients internet access behavior, removing the noise data from original data and converting the original data into Boolean type. Secondly, analyze the Boolean type data based on the improved ORAR association rules algorithm and dig out the connection degree between individual preferences. Lastly, give the validation of the improved ORAR association rules algorithm according to 40,000 random data of surfing the internet which mobile company provided of, achieve a better conclusion of the connection degree between individual preferences. The result shows that the accuracy clients content preferences will come true when using connection degree, taking the opportunity to sell to a specific user-related business, then achieving precise marketing and obtaining customer satisfaction requirements.

**Key words:** content preference; visit behavior analysis; improved ORAR association rules algorithm

## 0 引言

当今信息社会,随着移动互联网规模的日益庞大,手机阅读、支付、游戏和导航等丰富多彩的移动互联网应用正逐渐渗透到人们生活、工作等领域。我国移动通信营销策略逐渐由功能型业务的营销转向内容型业务的营销,因此基于内容偏好<sup>[1]</sup>的营销正成为增值业务的重点,分析客户访问移动互联网行为显得尤为重要。客户由于不同的兴趣爱好而表现出不同的互联网

访问行为,该访问行为里面蕴含着丰富的客户信息,通过这些信息进行客户内容偏好分析,从而更好地把握客户需求,支撑和促进移动互联网业务发展。因此基于内容偏好的移动互联网访问行为已经成为一个研究的热点问题。

## 1 移动客户互联网访问内容数据预处理

数据预处理的过程主要分为数据的采集过程、提取过程以及预处理<sup>[2,3]</sup>。由于数据的采集过程及提取过程都是采用通常的算法,在此不做赘述,文中主要给出数据预处理的具体过程。

### 1.1 客户偏好内容模块划分

手机浏览器上各种分类模块方便客户浏览丰富网络资源的同时也为文中客户偏好的细分提供了思路。将客户访问的内容划分模块会提高数据的分析效率。

收稿日期:2012-04-06;修回日期:2012-07-10

基金项目:国家自然科学基金资助项目(60973140/F0208);南京邮电大学科技创新训练计划项目(STTP)(Y2011117)

作者简介:宫婧(1977-),女,江苏南京人,副教授,硕士生导师,研究方向为计算机网络与安全;周飞飞(1990-),女,江苏南通人,研究方向为计算机网络。

将大模块进行细分成子模块能够更加精确地分析客户的访问行为<sup>[4,5]</sup>。文中将客户手机上网偏好内容大体细分如图 1:

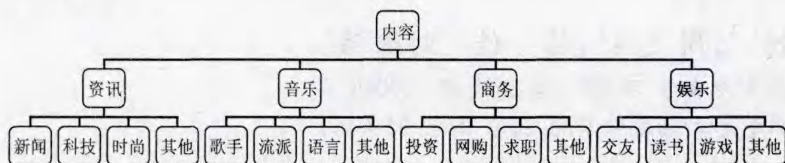


图 1 移动客户上网访问偏好内容的细分

### 1.2 客户数据预处理

移动客户偏好内容的细分模块已经建立,基于准确分析客户偏好内容的需要,下面对客户互联网访问行为<sup>[4~6]</sup>的源数据进行预处理。

#### ①数据的定义。

定义 1 分析数据:在实际数据分析过程中,有使用价值的数,文中定义它为对客户行为分析有实际价值的数。

定义 2 噪音数据:在实际数据分析过程中,对实际分析结果产生干扰的数,文中主要指客户的错误访问数据。

#### ②数据的获得与处理。

原始数据主要通过 WAP 网关访问日志、网络爬虫等手段获取。由于错误操作数据的存在使原始数据不能精确反应客户的偏好内容,因此需要剔除噪声数据。

文中规定,客户访问当下网站的时间低于 5s 就默认为错误数据。

下面以“娱乐偏好”为例,讨论数据的获取和预处理过程(见图 2):

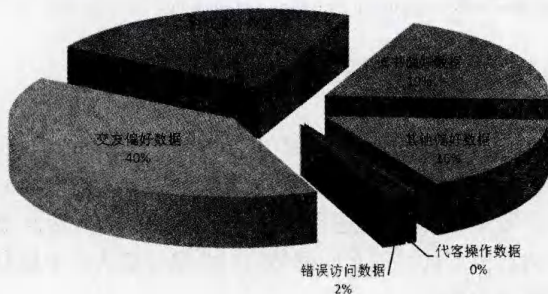


图 2 娱乐偏好中数据的预处理及娱乐偏好内各种子偏好所占的比例

从图 2 中可以看出噪声数据所占的比例很小,对于代客数据和错误操作数据完全可以忽略不计。舍弃错误数据,对分析数据进行研究。所以说,文中提出剔除噪声数据方法可行。

### 1.3 客户数据类型转化

得到分析数据后,对其进一步处理从而确定客户的内容偏好,为此定义了内容偏好阈值<sup>[6]</sup>作为客户对浏览内容偏好与否的标准。考虑一般情况下一个用户

对应一个移动号码,将此号码作为该用户的身份标识。

#### ①内容偏好阈值的定义。

定义 3 内容偏好的阈值  $\delta$ :移动用户对互联网访问内容偏好与否的分界值。文中给出两个评价标准,一是:客户每天访问某网站的总时间,二是:每天访问某网站的次数。

#### ②内容偏好阈值的确定。

文中通过对移动客户互联网访问时间段及访问频数进行分析来给出各种内容偏好的判定阈值。通过统计分析给出各种内容偏好的判定阈值,如表 1:

表 1 移动客户访问互联网阈值的判定

内容偏好	监测时间	平均访问 时间/天	访问频数/ 天
音乐偏好	9:00-23:00	1h	5
资讯偏好	11:00-13:00 和 17:00-20:00	3h	10
商务偏好	9:00-11:00 和 3:00-17:00	3h	10
娱乐偏好	11:00-13:00 和 17:00-23:00	4h	15

#### ③内容偏好数据类型转化。

由于量纲的不同不能将数据逐一比较并分析,文中对数据做了以下规定:达到阈值则归为该客户的偏好内容,否则不属于偏好内容,从而确定用户的偏好内容 Pref,

$$\text{Pref} = \begin{cases} 1 & \text{偏好内容} \\ 0 & \text{不偏好内容} \end{cases} \quad (1)$$

通过公式(1)将分析数据转化为易于处理的布尔类型的数据,以“娱乐偏好”模块为例,得到如下的关系矩阵 A。

$$A = \begin{matrix} & \begin{matrix} \text{交友} & \text{读书} & \text{游戏} & \text{其他} \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \vdots \\ n \end{matrix} & \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ a & b & c & d \end{bmatrix} \end{matrix} \quad (2)$$

注:  $n$  为总的客户数,  $a, b, c, d$  为 0 或 1

式(2)关系矩阵较为直观地表述出特定用户与各内容偏好的关系。将上面得到的关系矩阵作为输入数据,成为以下对各个内容偏好间关联程度研究的基础数据<sup>[7]</sup>。

## 2 内容偏好行为的关联规则算法研究

关联规则<sup>[8,9]</sup>是用来找出事务中频繁发生的项或

属性的所有子集,以及项目之间的相互关联性。中国拥有着庞大的移动客户群,可以使用关联规则进行业务数据挖掘。对于一给定客户,利用其各个偏好内容空间的关联度来确定偏好内容的关联组合。从而根据此关联组合对该客户进行业务套餐的推销。

### 2.1 ORAR 算法改进

目前关联规则算法有多种<sup>[10]</sup>,但有许多都不适合文中数据分析。ORAR 算法是针对布尔类型的数据进行研究,且只扫描一遍数据库,因此 ORAR 算法会更加适合该项研究的数据关联分析。

基于 ORAR 算法内容分析后,可知该算法只能在给出最小支持度阈值基础上求解频繁项集,通过判定支持度是否大于最小支持度来确定该频繁项集的保留与丢弃,但并没有将该频繁项集与其对应的支持度存储在同一名词空间内。因此难以查找各频繁项集对应的支持度,无法求解各关联规则对应的置信度及强规则。

针对以上不足,对 ORAR 算法进行了如下改进:

- 将大于支持度阈值的频繁项集的各个偏好标签与其对应的支持度存储在一个名字空间矩阵内。
- 在求解关联规则的置信度时,编写求解函数。根据偏好标签查找名字空间矩阵,再找出与偏好标签相匹配的支持度,代入置信度公式求解。

### 2.2 改进 ORAR 算法的内容偏好关联规则的实现

改进 ORAR 算法寻找内容偏好强关联的具体步骤如下:

step1: 客户手机访问互联网的行为,形成  $m$  条记录、 $n$  个内容偏好标签,将数据转换成布尔类型(即 0, 1 序列),存储成一个  $m \times n$  阶的矩阵。

step2: 根据算法设定支持度阈值,求解 1-项集的支持度。对于大于支持度阈值的项目集,将偏好标签与其对应的支持度存入矩阵  $Aorarl = (a_{ij})_{p \times q}$  ( $p = x$ ,  $q = 2$ ) 中,其中  $x$  为筛选结果行数。

step3: 根据求解的 1-项集,将 1-项集中的偏好标签两两组合(排除重复),求解 2-项集的支持度。对于大于支持度阈值的项集,将组合偏好标签与其对应的支持度存入  $Aorar2 = (a_{ij})_{p \times q}$  ( $p = y, q = 3$ ),其中  $y$  为筛选结果行数。

step4: 在求解  $k$ -项集时,根据求解的 1-项集和  $(k-1)$ -项集进行  $k$  阶组合,排除两种重复:项目集内部的偏好标签重复和项目集间的组合重复。对于大于支持度阈值的项目集,将组合偏好标签与其对应的支持度存入  $Aorark = (a_{ij})_{p \times q}$  ( $p = z, q = k+1$ ),其中  $z$  为筛选结果行数。

step5: 在置信度求解函数中,对形如  $X \Rightarrow Y$  的置信度求解时,从矩阵  $Aorarl = (a_{ij})_{p \times q}$  找出  $X$  的支持度,

从  $Aorar2 = (a_{ij})_{p \times q}$  中找出  $XY$  的支持度,从而求解出 confidence ( $X \Rightarrow Y$ ); 对形如  $XY \Rightarrow Z$  的置信度求解时,从矩阵  $Aorar2 = (a_{ij})_{p \times q}$  中找出  $XY$  的支持度,从  $Aorar3 = (a_{ij})_{p \times q}$  中找出  $XYZ$  的支持度,从而求解出 confidence ( $XY \Rightarrow Z$ )。依次类推,从而求解出形如  $A_0 \cdots A_m \Rightarrow B_0 \cdots B_n$  的置信度 confidence ( $A_0 \cdots A_m \Rightarrow B_0 \cdots B_n$ )。

## 3 基于改进 ORAR 互联网访问内容偏好行为算法验证

下文将根据移动公司提供的从 2010 年 5 月至 2011 年 10 月中随机抽取的 4 万条移动客户互联网访问情况的数据<sup>[11]</sup>,用 MATLAB 软件对上面的理论分析进行验证。

### 3.1 大模块内容偏好的强关联

改进 ORAR 算法对关系矩阵数据进行频繁项集和置信度求解,得出形如  $X \Rightarrow Y$  (即:偏好 1  $\Rightarrow$  偏好 2) 的强关联规则<sup>[12]</sup>。在支持度阈值为 0.2 的情况下得到结论。图 3 给出了直观图,其中  $S$  代表支持度,  $C$  代表置信度。

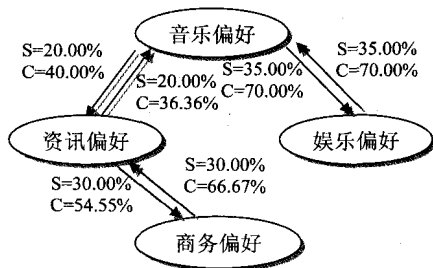


图3 内容偏好关联关系直观

由图 3 可知,这四大模块的内容偏好标签具有很强的相关性,为了更加精确地研究内容偏好的关联行为,需要研究子模块内容偏好的关联度。

对于形如  $XY \Rightarrow Z$  的强关联规则,由于其求得的支持度与置信度会更加小,因此不再求解其频繁项集和置信度。

### 3.2 子模块内容偏好的强关联

以“资讯偏好”和“商务偏好”两大模块为例,合并两者后研究两者子模块内容偏好之间的关联关系。通过改进算法得到各个关联规则的置信度。设置置信度阈值为 0.5,筛选得到的强关联规则直观图见图 4:

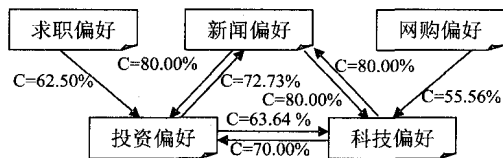


图4 形如  $X \Rightarrow Y$  的内容偏好强关联结果直观图

图 4 中关联组合 {新闻、投资}, {科技、投资} 和 {网购、科技} 都跨越大模块内容偏好分类,因此可以

组合这些子模块内容偏好业务进行推广销售。基于改进算法,可以求出更多内容偏好的关联组合,对于  $XY \Rightarrow Z$  形式的强关联规则,其结果见表 2。

表 2 形如  $XY \Rightarrow Z$  的内容偏好强关联结果

偏好 1	偏好 2	偏好 3	置信度
新闻	求职	投资	100.00%
投资	科技	新闻	100.00%
新闻	投资	科技	87.50%
新闻	科技	投资	87.50%
求职	投资	新闻	80.00%
新闻	投资	求职	50.00%

注:表中关联规则为:偏好 1  $\cup$  偏好 2  $\Rightarrow$  偏好 3

表 2 中组合{新闻、求职、投资}和{投资、科技、新闻}也都跨越大模块内容偏好分类。由此可见,不仅可以组合各相关内容偏好进行销售,而且可以根据客户已经具有的内容偏好类别,推荐该客户另一由这两个偏好推导出的偏好业务。

通过对大模块内容偏好强关联分析可知,客户互联网访问内容偏好模块,即音乐、资讯、商务、娱乐之间存在着很强的关联性,可以在大方向上进行业务组合的销售;通过进一步对子模块内容偏好强关联分析,得到更细化关联分析结果,这不仅可以进行业务的组合销售,还可以挖掘客户的潜在内容偏好,进行业务的推荐销售。

#### 4 结束语

综上所述,文中是基于关联规则挖掘移动客户互联网访问内容数据间的关联,在已细分的各个偏好内容模块中寻找内在的关联。从而针对特定的用户提供相应的精确业务,赢得营销机会。该方法相对于以往

凭借客户基础资料和客户通信行为的分析有较大进步,更能有效命中目标客户群的偏好访问内容,提高营销效率,增加客户满意度。

#### 参考文献:

- [1] 中国移动通信集团福建有限公司. 基于客户偏好研究,开展数据及信息业务内容深度营运[EB/OL]. [2010-09-20]. <http://www.doc88.com/p-703879803187.html>.
- [2] Witten L H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques[M]. 3rd ed. [s. l.]: Morgan Kaufmann, 2011.
- [3] 安淑芝. 数据仓库与数据挖掘[M]. 北京:清华大学出版社, 2005.
- [4] 储文静, 奉国和. 基于 Weka 读者借阅行为分析[J]. 情报科学, 2010, 28(3): 424-429.
- [5] 付 锋. 移动互联网访问行为分析研究[EB/OL]. [2011-09-06]. <http://labs.chinamobile.com>.
- [6] 王 静, 张春海. 基于布尔关联规则挖掘的加权阈值分析[J]. 计算机技术与发展, 2009, 19(12): 13-16.
- [7] 丁艳辉, 王洪国, 高 明, 等. 一种基于矩阵的关联规则挖掘新算法[J]. 计算机科学, 2006, 33(4): 188-189.
- [8] Lu Juemin, Ma Guodong, Zheng Yu. The Association Analysis for Library Circulation Data Based Mining Technique[J]. Journal of Modern Information, 2009, 29(9): 108-110.
- [9] 邵峰晶, 于忠清. 数据挖掘原理与算法[M]. 第 2 版. 北京: 科学出版社, 2009.
- [10] 王爱平, 王占凤, 陶嗣干, 等. 数据挖掘中常用关联规则算法[J]. 计算机技术与发展, 2010, 20(4): 105-108.
- [11] 中国互联网监测研究机构 & 数据平台. DCCI 2011 中国移动互联网用户调查报告[EB/OL]. [2011-06-10]. <http://www.dcci.com.cn>.
- [12] Han Jiawei, Kamber M. Data Mining Concepts and Techniques[M]. San Francisco: Morgan Kaufmann Publishers, 2001.

(上接第 148 页)

- [3] 陈大峰, 万洛楷. 移动 BOSS 系统中客户信用度综合评定的研究[J]. 南京审计学院学报, 2006(4): 99-103.
- [4] 王丽平, 李多全. 基于 AHP 方法计算电信用户信用度[J]. 计算机工程与应用, 2008(32): 232-239.
- [5] 李祚泳, 彭荔红. BP 网络学习能力与泛化能力满足的不确定关系式[J]. 中国科学( E 辑), 2003, 33(10): 887-895.
- [6] 彭 宇, 彭善元, 刘兆庆. 微粒群算法参数效能的统计分析[J]. 电子学报, 2004(2): 209-213.
- [7] Kennedy J, Eberhart R. Particle Swarm Optimization[C]//IEEE International Conference on Neural Networks (ICNN 95). Perth, Australia: [s. n.], 1995.
- [8] Parsopoulos K E, Vrahatis M N. Particle Swarm Optimizer in Noisy and Continuously Changing Environments[C]//Proceeding of the IASTED International Conference on Artificial Intelligence and Soft Computing. ICanun Mexico: IASTED/

ACTA Press, 2001: 289-294.

- [9] Parsopoulos K E, Vrahatis M N. Particle Swarm Optimization Method for Constrained Optimization Problems[C]//Proceedings of the Euro-international Symposium on Computational Intelligence. Konica, Slovakia: IOS Press, 2002.
- [10] 吴小红, 金炳尧. 前馈神经网络的一种优化 BP 算法[J]. 计算机科学, 2004, 31(10A): 240-241.
- [11] 潘 昊, 侯清兰. 基于粒子群优化算法的 BP 网络学习研究[J]. 计算机工程与应用, 2006(16): 41-43.
- [12] 曾万里, 危韧勇, 陈红玲. 基于改进 PSO 算法的 BP 神经网络的应用研究[J]. 计算机技术与发展, 2008, 18(4): 49-51.
- [13] 唐 俊. PSO 算法原理及应用[J]. 计算机技术与发展, 2010, 20(2): 213-216.