Recommend
火龙果软件

互联网+用户画像大数据分析

统计学基础

- **集中量：**求和、平均数、中位数、众数、

- **差异量：**全距、标准差、方差、最小值、最大值、标准误

- **分布：**正态分布、峰度系数、偏度系数、正偏态、负偏态、高狭峰、低阔峰、离散变量、连续变量

内部资料，请勿外传

目 录

用户画像研究概述

用户画像研究流程

用户画像方法与案例演示

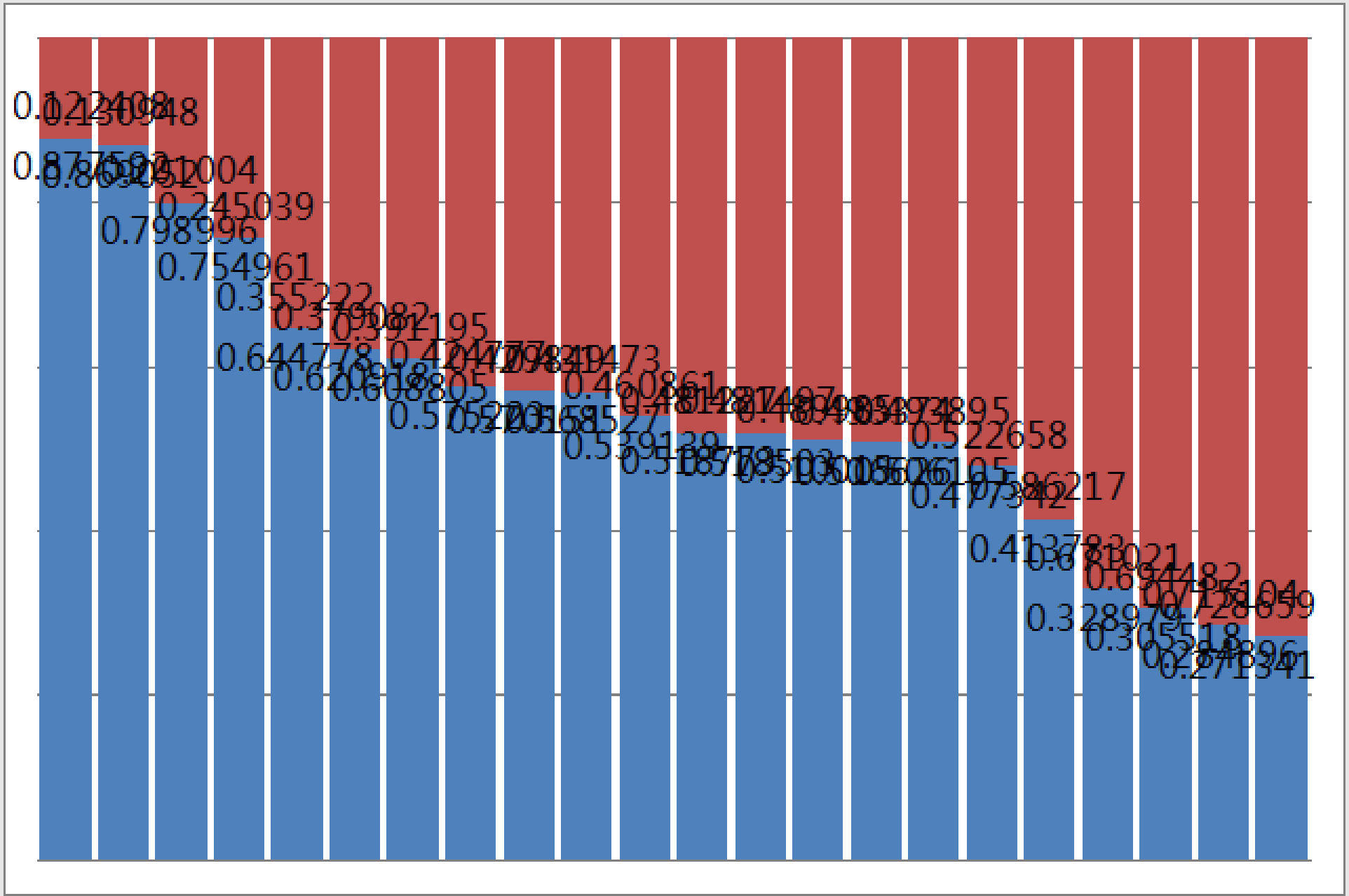
Recommend
火龙果软件

用户画像概述

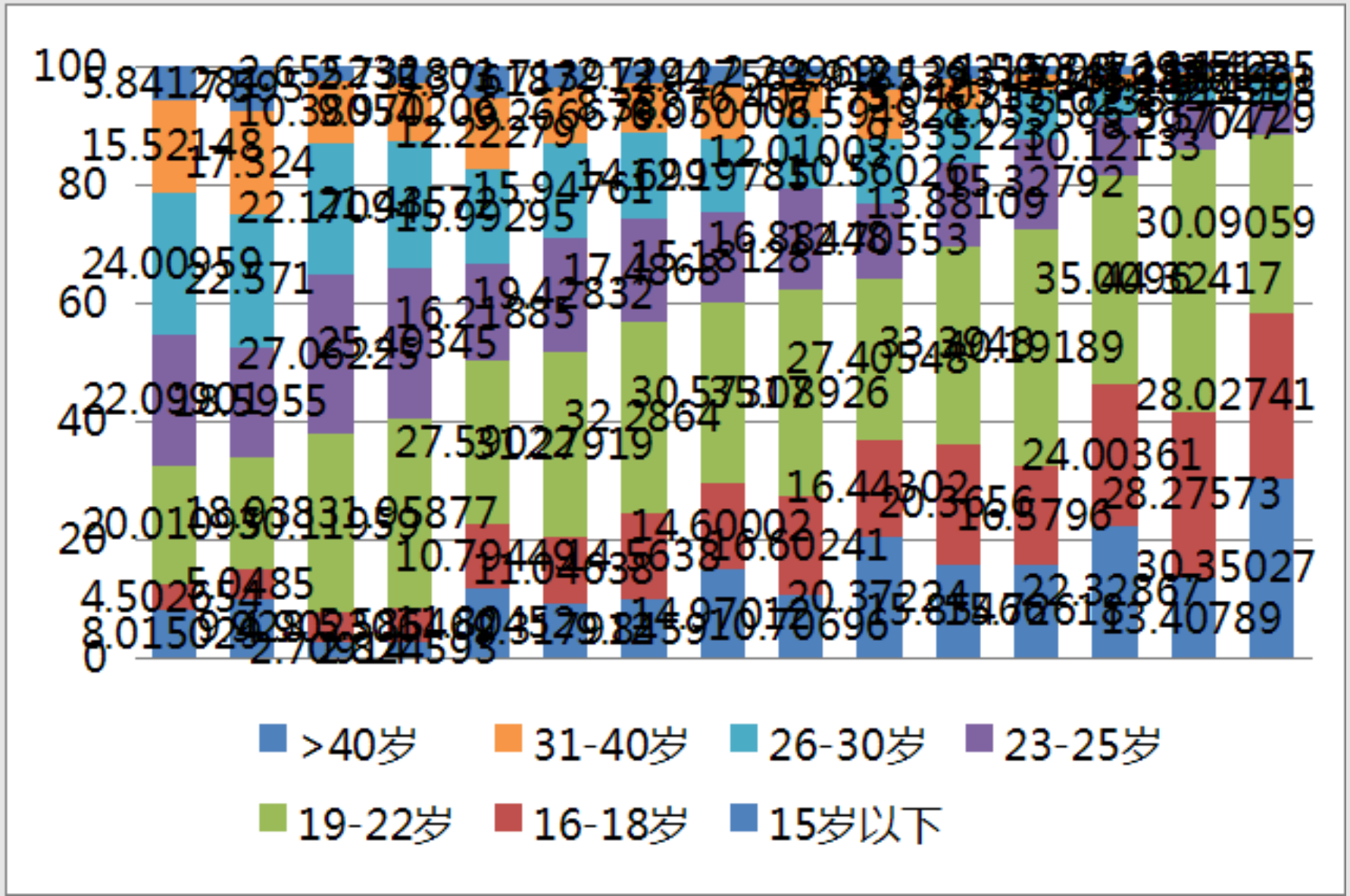
用户画像概述

一、群体用户定量描述统计

群体定量分类统计——各类用户性别构成



群体定量分类统计——各类XX用户年龄构成

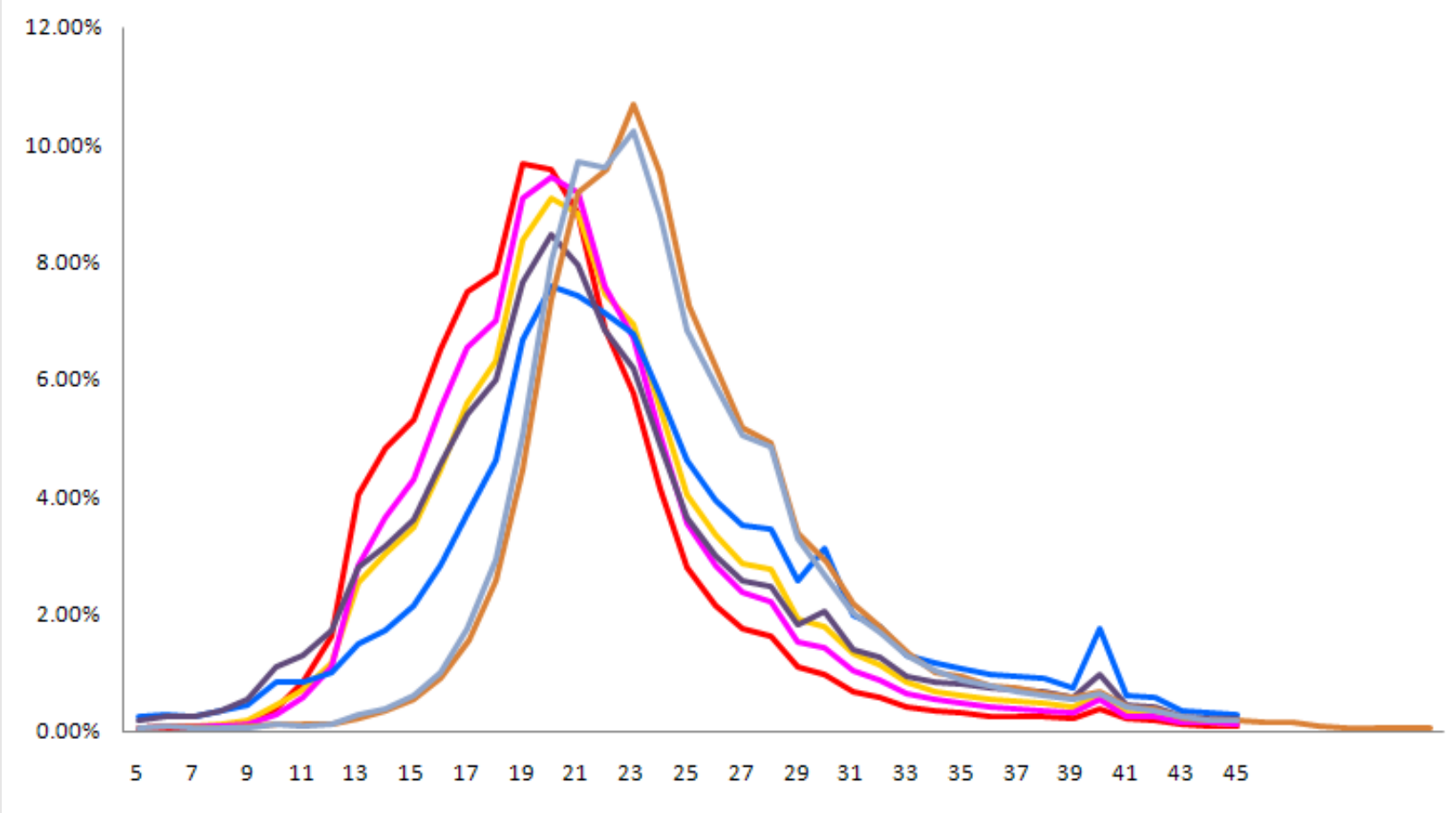


群体定量分类统计——各类XX用户年龄构成

年 龄	0-10岁	11-15岁	16-18岁	19-22岁	23-25岁	26-30岁	31-40岁	>40岁
CC频道	4%	4%	5%	20%	22%	24%	16%	6%
XX频道	6%	4%	5%	19%	19%	23%	17%	8%
AA成交	1%	1%	5%	30%	27%	22%	10%	3%
XX成交	1%	1%	6%	32%	25%	21%	10%	3%
XX活跃用户	5%	7%	11%	28%	16%	16%	12%	5%
XX登录用户	3%	7%	11%	31%	19%	16%	9%	4%
AA	2%	8%	15%	32%	17%	15%	8%	3%
XX	4%	11%	15%	31%	15%	12%	9%	3%
BB	2%	9%	17%	35%	17%	12%	6%	2%
XX商城用户	4%	16%	16%	27%	13%	11%	9%	4%
AA	2%	14%	20%	33%	14%	9%	5%	2%
AA	5%	11%	17%	40%	15%	8%	3%	2%
WOW活跃用户	6%	16%	24%	35%	10%	5%	2%	1%
大话西游活跃	2%	11%	28%	44%	8%	3%	1%	1%
XX烽火活跃用户	4%	26%	28%	30%	6%	3%	2%	1%

Recommend
火龙果软件

用户年龄分布图


12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45

二、个体用户定性描述

典型用户个体描述案例

王庆龙，

男，22岁，中专或以下学历，保安，2年工作经验，月收入在2000元以下，知识层次较低，理解表达能力较差。**性格内向，不爱说话，但是在网上却很活跃，是个很有代表性的用户。**使用XX4年，没有电脑，由于工作性质特殊**（用户的职业为保安），所以用户每天使用手机登陆XX及手机XX网**来打消无聊的时间，或下班去网吧上网。他的同事大都年龄较大，一般都是把打牌和喝酒作为娱乐活动，这是他所不喜欢的，与他同龄的同事大都喜欢把上网作为娱乐。他对电脑使用较为生疏。认为XX2009看上去不错，如果能把08的功能都加上再稳定些就更好了。与2009相比，更习惯使用2008。

访谈发现：1.对于广大低端用户来说，**易理解、简单、方便、快捷**是他们最需要的，也是他们不用MSN的原因之一；2.用户对XX依赖性很大，这样的用户希望XX的功能更强大，真正实现一站式在线生活。

来源：一次**XX2009** 用户访谈用户画像

典型用户个体描述

女，19岁，高中，学生，理解表达能力较好，性格对陌生人内敛对朋友外向活泼，不喜欢动脑，什么都喜欢方便的，最好只按一下就全部搞定的。**使用XX7年**，现实的社交圈基本局限在同班同学，但是网上却有很多不认识的好友，喜欢认识不同类型的人。虽然网龄较高但是与很多女孩子一样依然是个**电脑白痴女**，她喜欢操作越简单越方便越好。访谈过程中她说的最多的一句话就是**“我个人比较懒！”，最怕麻烦，**MSN就是太麻烦才不用的。对于电脑游戏喜欢互动性好，但是操作简单的，比如劲舞团、大话西游等。

访谈发现：1，很多女用户虽然使用电脑多年但是依然是个电脑白痴，对她们来说**所有操作如果超过两步就会晕**！对于设置性操作她们基本没有使用过，**她们只使用初始化设置**，希望在修改设置方面更简单！2，该用户的另一个特点就是“懒”，稍微有点麻烦或困难，她们就会懒得做，懒得想，如果有**傻瓜式，全自动式操作就很适合她们**。

来源：一次XX2009 用户访谈用户画像

三、用户画像数据挖掘

数据挖掘——付款用户【对应分析】

Recommend
火龙果软件

XX用户【聚类分析】

Recommend
火龙果软件

XX用户【聚类分析】特征得到的启发

高认同用户

低介入用户

新进用户

- 令这些用户改变(即令其从不付费使用到付费使用)的可能性较低：
  - TT龄长，使用XX历史较长，但仍不付费使用
  - 亦较少使用其它XXVIP业务
  - 换装频率不固定

- XX的核心用户的画像：
  - 19-25岁男女
  - 丰富使用XX各类业务/VIP业务的TT龄较长的活跃用户
  - 定期换装

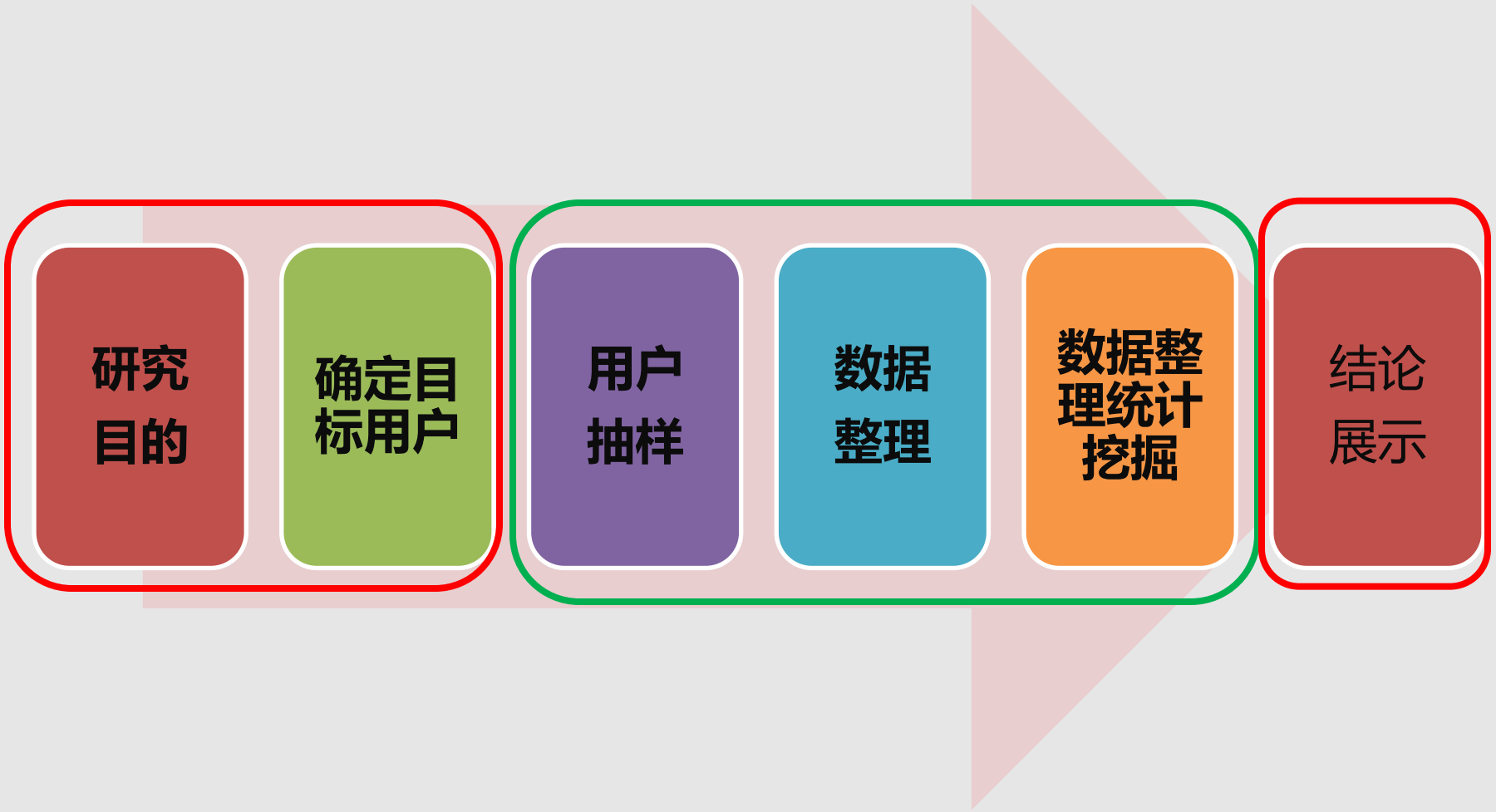
- 可拉动、改变的用户
  - 15-22岁男女
  - TT龄(2-5年)和使用XX历史较短
  - 定期换装

- 通过提升品牌满意度，增强其付费使用的信心

用户画像流程

Recommend
火龙果软件

用户画像流程



提取用户

Recommend
火龙果软件

抽样的几个概念

总体
Population

- 是所要研究的对象的全体。例如，考察XX农场用户体验，目标总体就是所有的XX农场的用户。抽样总体是用于从中抽取样本的总体。

抽样
Sample

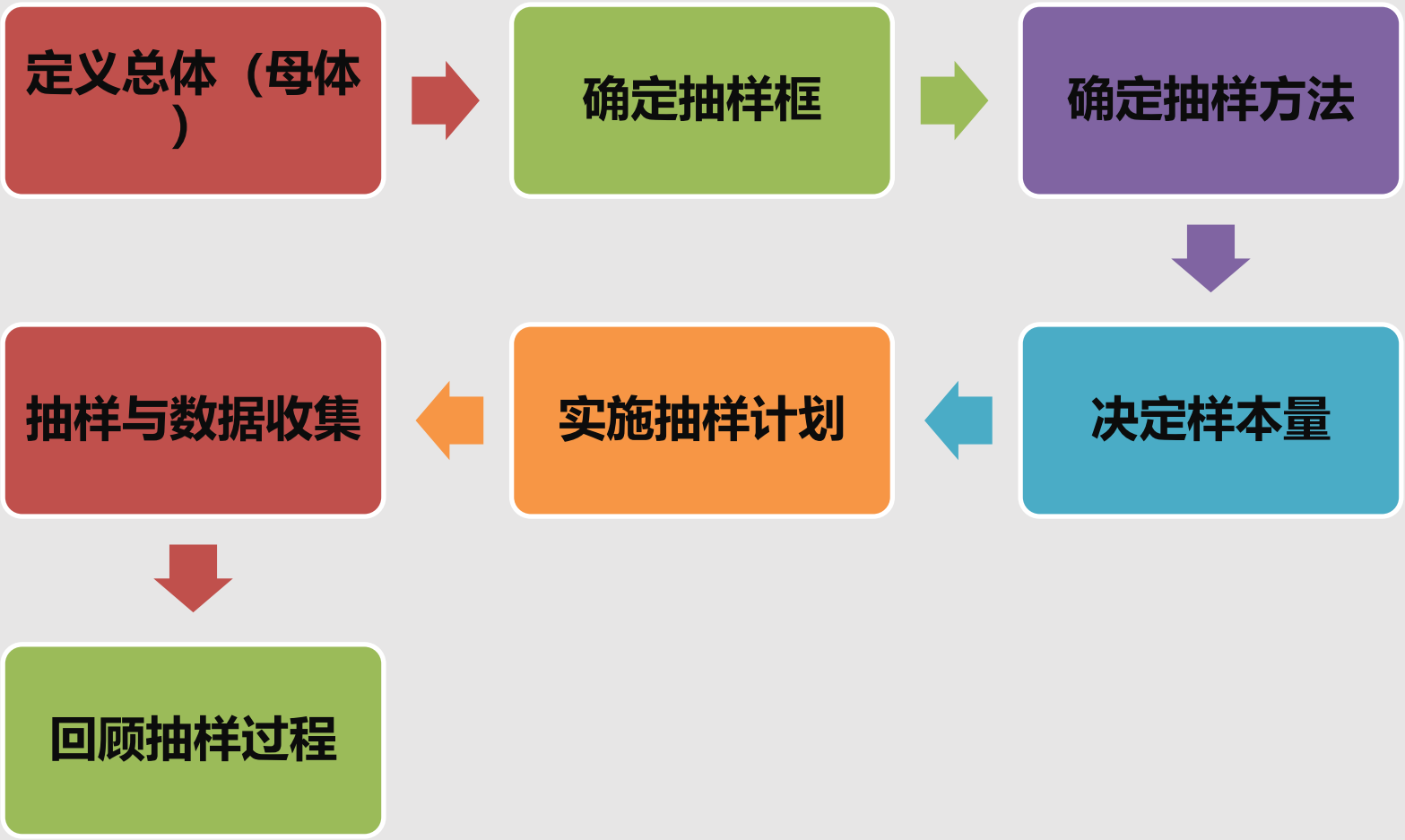
- 从目标总体（Population，或称为母体）中抽取一部分个体作为样本（Sample），通过观察样本的某一或某些属性，依据所获得的数据对总体的数量特征得出具有一定可靠性的估计判断，从而达到对总体的认识。

抽样框

- 在抽样之前，总体应划分成抽样单位，抽样单位互不重叠且能合成总体，总体中的每个个体只属于一个单位。抽样框是一份包含所有抽样单元的名单。

Recommend
火龙果软件

抽样过程



Recommend
火龙果软件

抽样方法

简单随机抽样
simple random sampling

- 从总体N个单位中随机地抽取n个单位作为样本，使得每一个容量为样本都有相同的概率被抽中。特点是：每个样本单位被抽中的概率相等，样本的每个单位完全独立，彼此间无一定的关联性和排斥性。

系统抽样
systematic sampling

- 等距抽样。将总体中的所有单位按一定顺序排列，在规定的范围内随机地抽取一个单位作为初始单位，然后按事先规定好的规则确定其他样本单位。先从数字1到k之间随机抽取一个数字r作为初始单位，以后依次取r+k、r+2k……等单位。这种方法操作简便，可提高估计的精度。

分层抽样
stratified sampling

- 将抽样单位按某种特征或某种规则划分为不同的层，然后从不同的层中独立、随机地抽取样本。从而保证样本的结构与总体的结构比较相近，从而提高估计的精度。

整群抽样
cluster sampling

- 将总体中若干个单位合并为组，抽样时直接抽取群，然后对中选群中的所有单位全部实施调查。抽样时只需群的抽样框，可简化工作量，缺点是估计的精度较差

抽样效度与信度

Recommend
火龙果软件

分半信度

- 将同一批用户随机分成两组，计算关键指标的占比、相关系数。

复本信度

- 同一总体中随机抽取多个样本进行比较

内部一致系数

- 同质

一个教师的话

我不止一次给我的市场调研课学生说过，你们利用人人网之类做问卷调查基本是无效的，当个作业也就算了。

你们心里要明白，所谓物以类聚人以群分，利用社交网络的link做出来的问卷，没有随机性可言。包括你在msn、XX之类上传问卷给朋友请他们帮忙。缺乏随机性的样本，那就没有代表性。就好像小区的人，无法代表全体市民，你的朋友，无法代表全体学生。

摘自：魏武辉的BLOG http://weiwuhui.com/

Recommend
火龙果软件

数据整理

数据整理

Recommend
火龙果软件

数据检查

- 极端值处理；**心理学研究把超过2个标准差之外的值剔除。**

缺失值处理

- 没有观测到
- 有明显错误

数据分组

- 例如：年龄分段、选择处理等

数据检查——用户选择

描述统计量

	N	极小值	极大值	均值	标准差
成交年龄	182250	0	110	25.00	8.475
有效的 N（列表状态）	182250				

用户年龄取值范围：9岁—41岁

用户年龄占比分布

16岁—34岁 占比合计： 91.83%

年龄	占比	年龄	占比	年龄	占比	年龄	占比
6	0.07%	16	0.89%	26	6.20%	36	0.77%
7	0.06%	17	1.53%	27	5.18%	37	0.73%
8	0.05%	18	2.56%	28	4.90%	38	0.62%
9	0.06%	19	4.45%	29	3.36%	39	0.56%
10	0.12%	20	7.37%	30	2.89%	40	0.68%
11	0.10%	21	9.19%	31	2.16%	41	0.41%
12	0.10%	22	9.59%	32	1.78%	42	0.36%
13	0.21%	23	10.69%	33	1.33%	43	0.21%
14	0.34%	24	9.52%	34	0.99%	44	0.17%
15	0.55%	25	7.26%	35	0.93%	45	0.16%

Recommend
火龙果软件

用户画像数据挖掘实例演示

1. **对应分析**
2. **聚类分析**

用户画像方法

——“对应分析”实例演示

对应分析数据格式整理

特征	只看过	只关注XX	关注实物未购买	购买并付款
男性	48	36	28	10
女性	52	64	72	90
学生	67	65	63	22
非学生	33	35	37	78
工作	25	26	28	61
其他职业	8	9	9	17
14岁以下	15	14	18	-
15-18岁	45	41	38	2
19-22岁	27	26	23	23
23-25岁	8	11	11	23
26-30岁	3	5	6	13
31岁以上	2	4	4	5
有网购AA经验	27	33	35	100
无网购AA经验	73	67	65	-
无网购	73	67	65	-
网购50元以下	6	5	5	9
50-100元	10	13	14	24
100-200元	7	10	11	24
200元以上	4	5	5	5
喜欢XX会买实物	25	29	37	31
不会喜欢XX会买实物	36	31	21	15
说不清喜欢XX会买实物	39	40	41	18

Recommend
火龙果软件

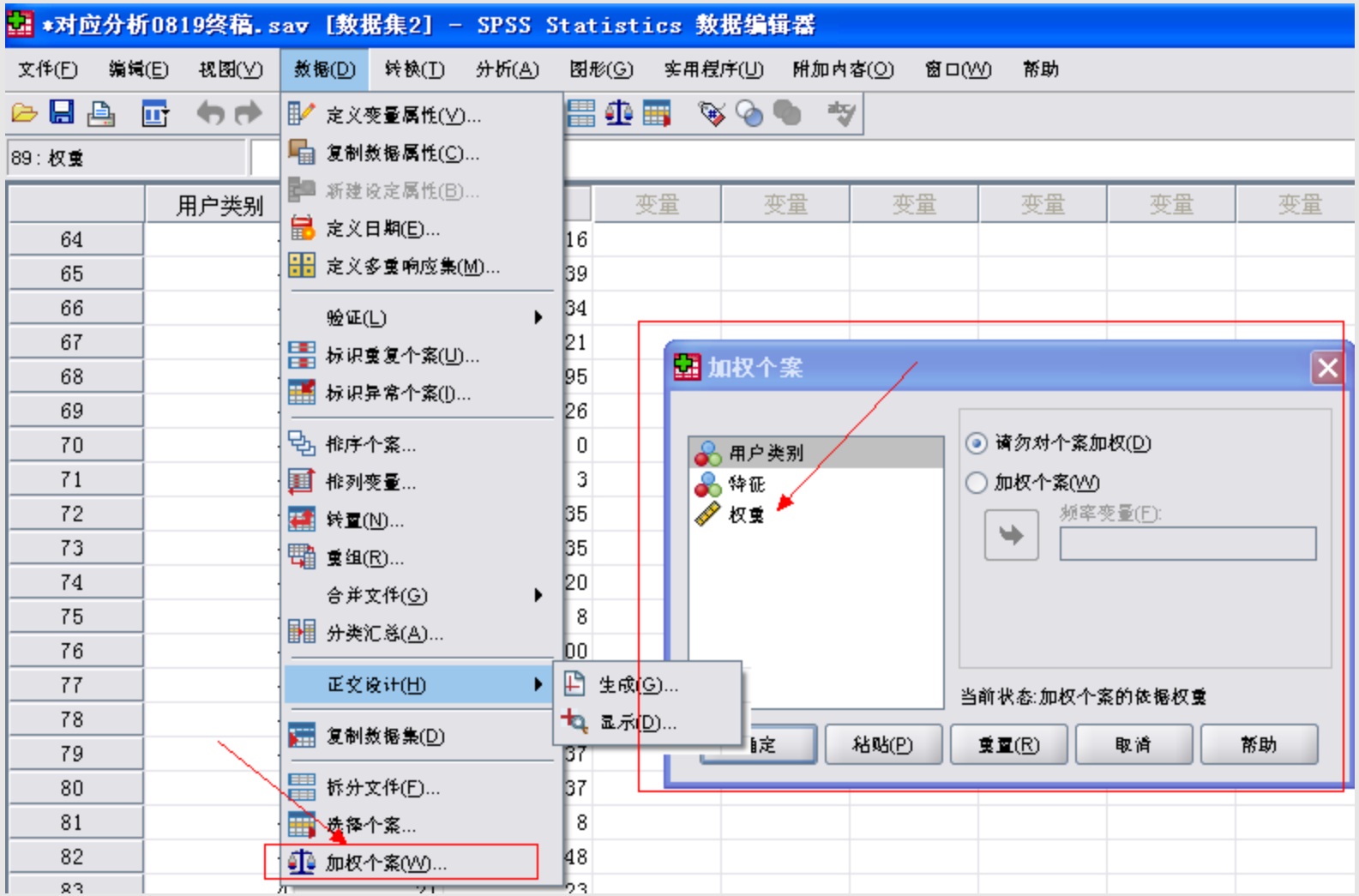
	用户类别	特征	权重
1	1	1	48
2	1	2	52
3	1	3	67
4	1	4	33
5	1	5	25
6	1	6	8
7	1	7	15
8	1	8	45
9	1	9	27
10	1	10	8
11	1	11	3
12	1	12	2
13	1	13	27
14	1	14	73
15	1	16	6
16	1	17	10
17	1	18	7
18	1	19	4
19	1	20	25
20	1	21	36
21	1	22	39

	用户类别	特征	权重
22	2	1	36
23	2	2	64
24	2	3	65
25	2	4	35
26	2	5	26
27	2	6	9
28	2	7	14
29	2	8	41
30	2	9	26
31	2	10	11
32	2	11	5
33	2	12	4
34	2	13	33
35	2	14	67
36	2	16	5
37	2	17	13
38	2	18	10
39	2	19	5
40	2	20	29
41	2	21	31
42	2	22	40

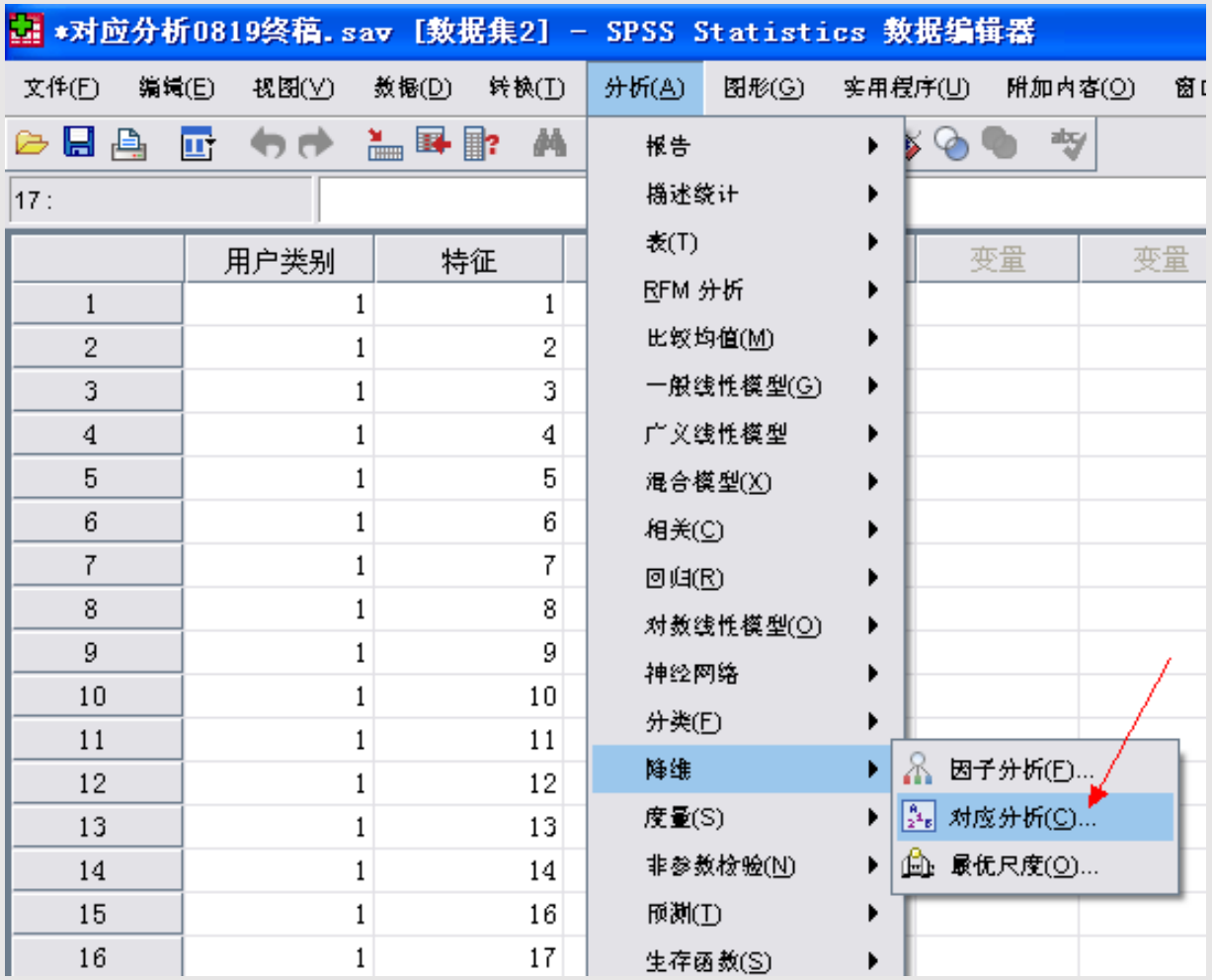
	用户类别	特征	权重
43	3	1	28
44	3	2	72
45	3	3	63
46	3	4	37
47	3	5	28
48	3	6	9
49	3	7	18
50	3	8	38
51	3	9	23
52	3	10	11
53	3	11	6
54	3	12	4
55	3	13	35
56	3	14	65
57	3	16	5
58	3	17	14
59	3	18	11
60	3	19	5
61	3	20	37
62	3	21	21
63	3	22	41

	用户类别	特征	权重
64	4	1	16
65	4	2	139
66	4	3	34
67	4	4	121
68	4	5	95
69	4	6	26
70	4	7	0
71	4	8	3
72	4	9	35
73	4	10	35
74	4	11	20
75	4	12	8
76	4	13	100
77	4	14	0
78	4	16	14
79	4	17	37
80	4	18	37
81	4	19	8
82	4	20	48
83	4	21	23
84	4	22	28

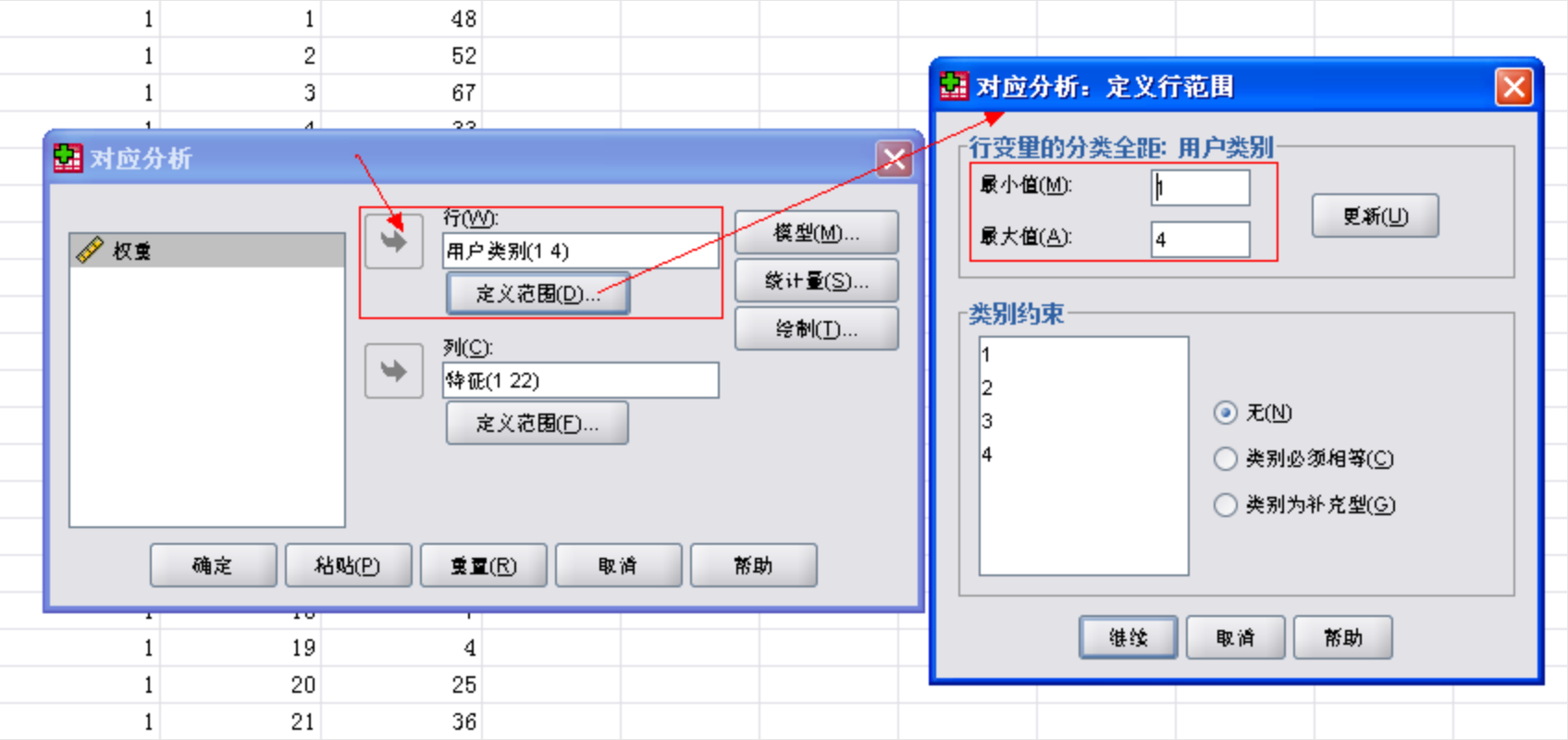
加权个案


*对应分析0819终稿.sav [数据集2] - SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 图形(G) 实用程序(U) 附加内容(O) 窗口(W) 帮助
89 : 权重
用户类别
定义变量属性(V)...
复制数据属性(C)...
新建设定属性(B)...
定义日期(E)...
定义多重响应集(M)...
验证(L)
标识重复个案(U)...
标识异常个案(I)...
排序个案...
排列变量...
转置(N)...
重组(R)...
合并文件(G)
分类汇总(A)...
正交设计(H)
生成(G)...
显示(D)...
复制数据集(D)
拆分文件(F)...
选择个案...
加权个案(W)...
加权个案
用户类别
特征
权重
请勿对个案加权(D)
加权个案(W)
频率变量(F):
当前状态:加权个案的依据权重
确定
粘贴(P)
重置(R)
取消
帮助

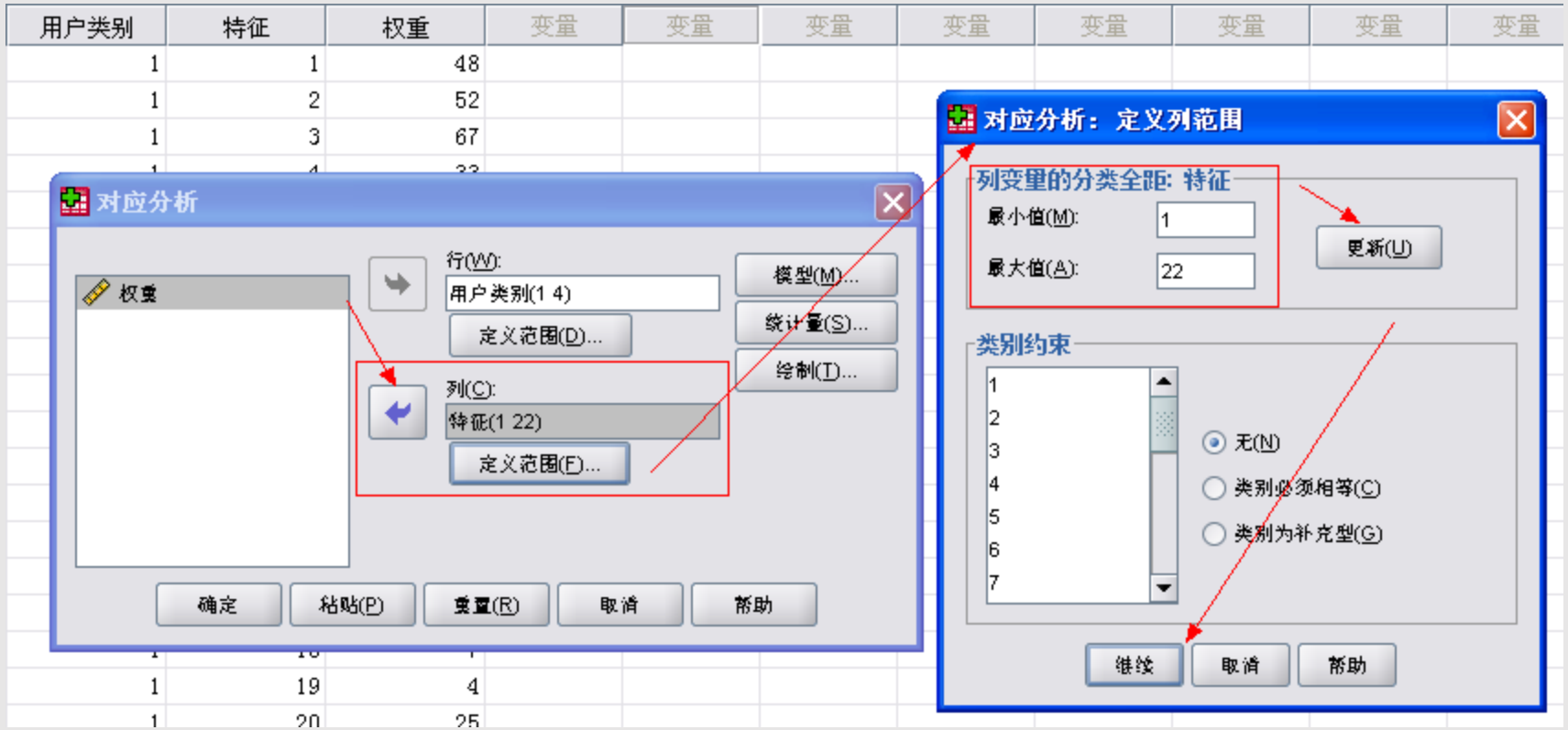
对应分析过程


*对应分析0819终稿.sav [数据集2] - SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 图形(G) 实用程序(U) 附加内容(O)
17 :
用户类别 特征
变量 变量
报告
描述统计
表(T)
RFM 分析
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
因子分析(F)...
对应分析(C)...
最优尺度(O)...

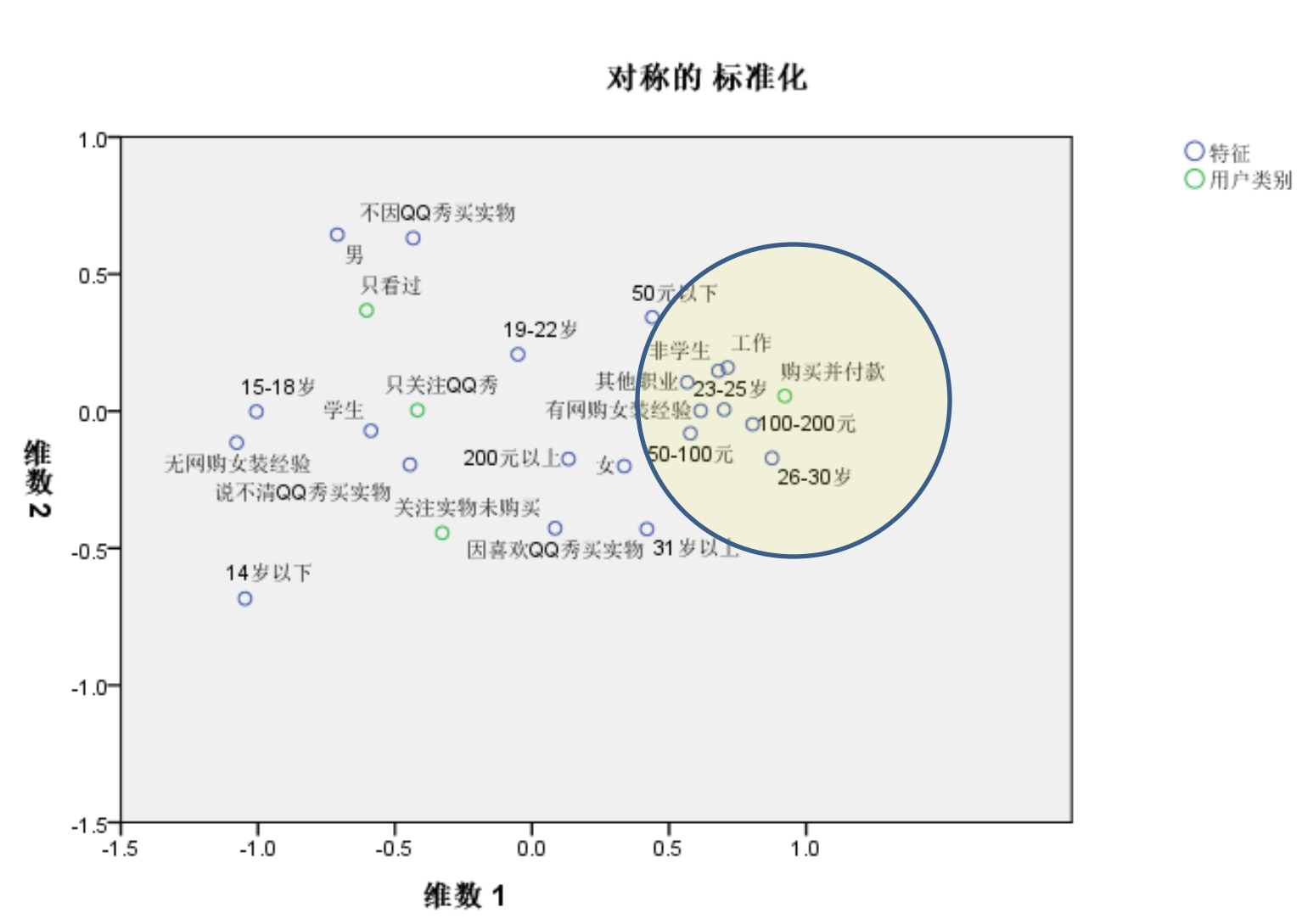
对应分析——定义行范围（用户类别）



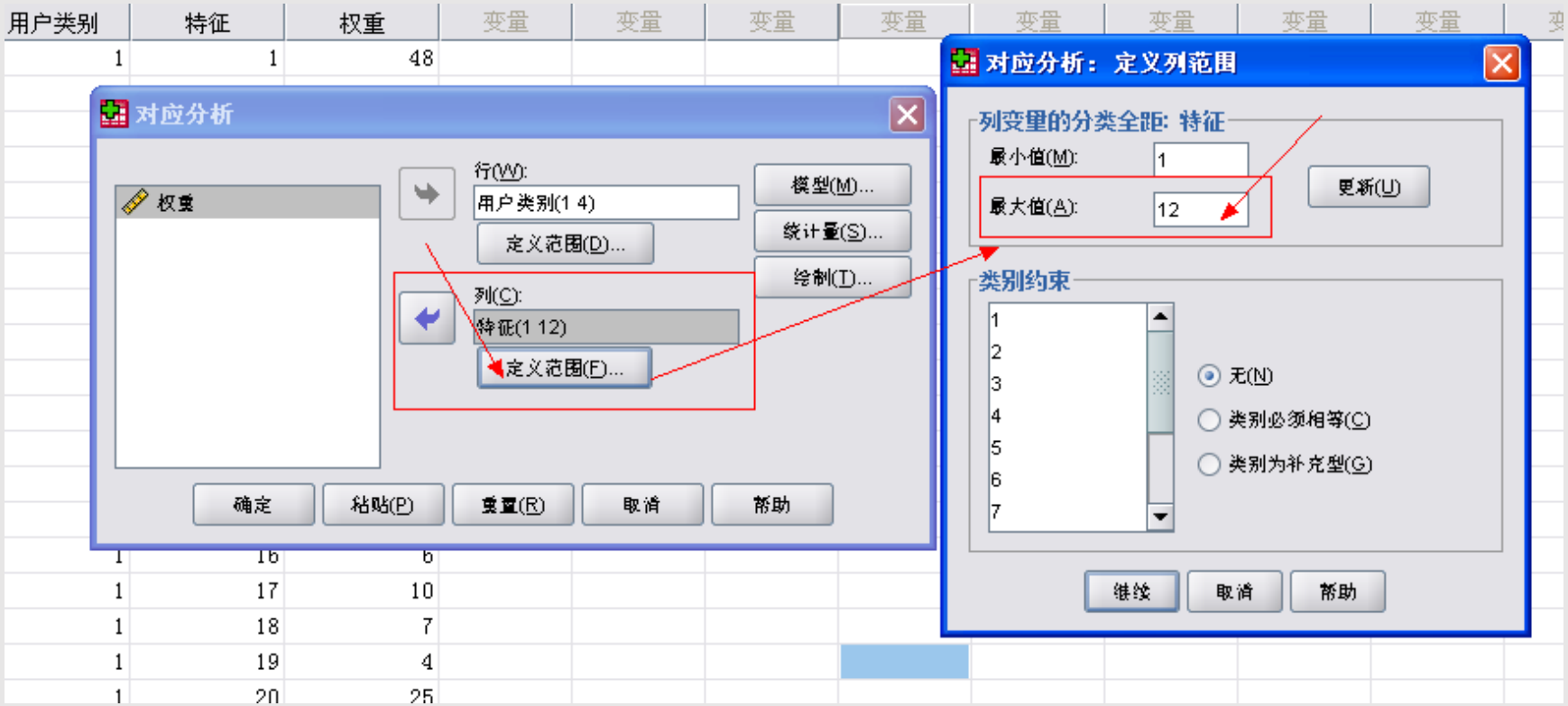
对应分析——定义列范围（用户特征）



对应分析结果图


对称的 标准化
特征
用户类别
不因QQ秀买实物
男
只看过
50元以下
19-22岁
非学生
工作
购买并付款
15-18岁
只关注QQ秀
其他职业
23-25岁
学生
有网购女装经验
100-200元
无网购女装经验
200元以上
女
50-100元
26-30岁
说不清QQ秀买实物
关注实物未购买
因喜欢QQ秀买实物
31岁以上
14岁以下
维数 2
维数 1
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.5
-1.0
-0.5
0.0
0.5
1.0

减少用户特征

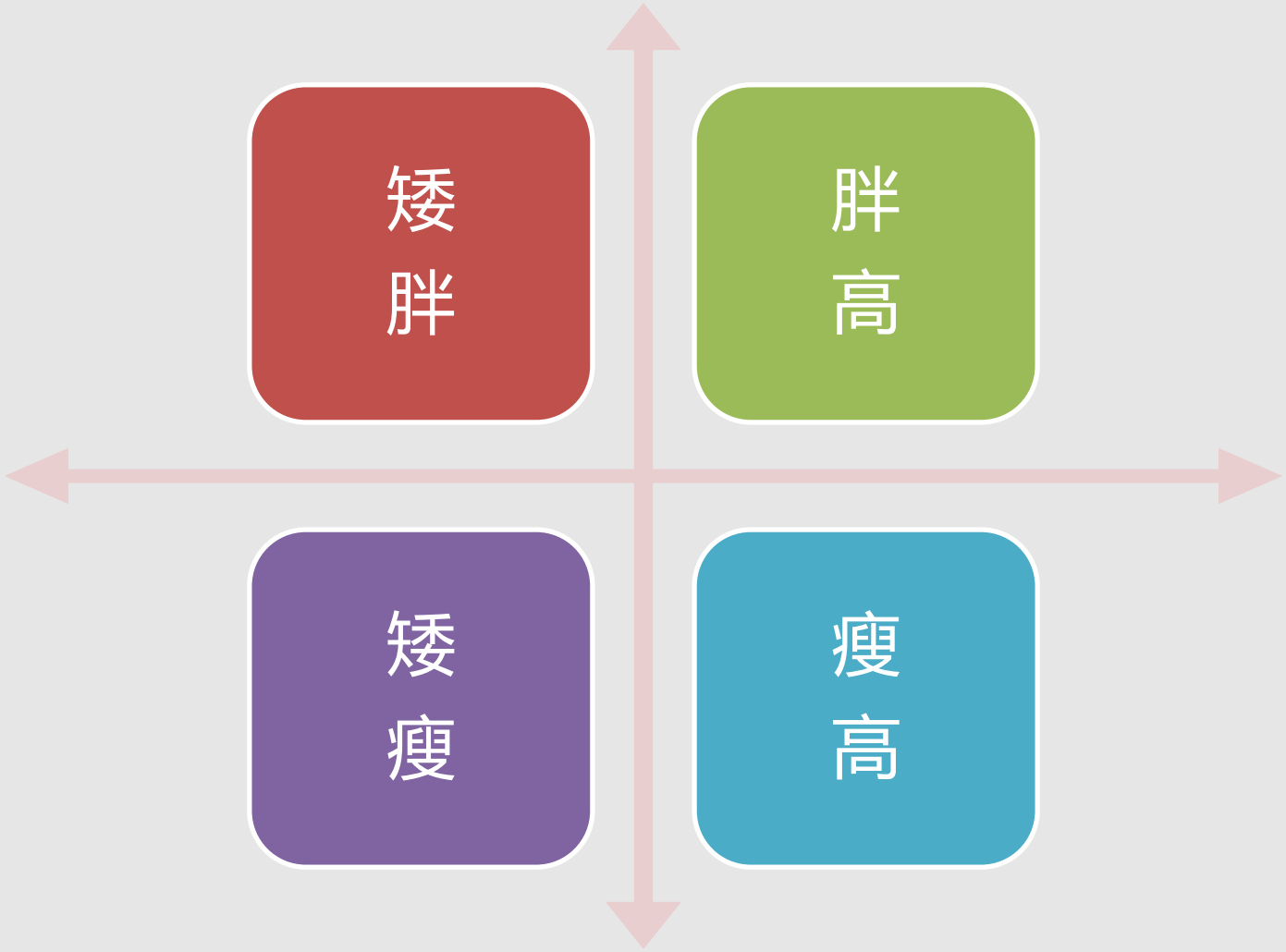

用户类别
特征
权重
对应分析
权重
行(W):
用户类别(1 4)
定义范围(D)...
列(C):
特征(1 12)
定义范围(F)...
模型(M)...
统计量(S)...
绘制(T)...
确定
粘贴(P)
重置(R)
取消
帮助
对应分析：定义列范围
列变量的分类全距: 特征
最小值(M):
1
最大值(A):
12
更新(U)
类别约束
无(N)
类别必须相等(C)
类别为补充型(G)
继续
取消
帮助

Recommend
火龙果软件

用户画像方法

——“聚类分析”实例演示

人群划分

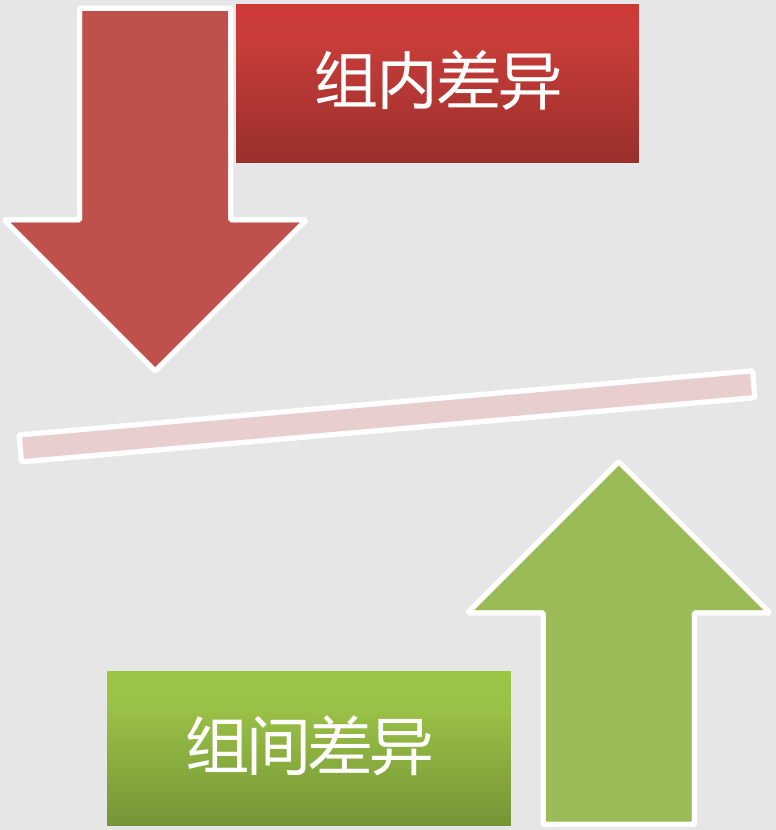


Recommend
火龙果软件

聚类分析

概念

- 根据数据本身结构特征对数据进行分类的方法——聚类分析，通过聚类分析，可以把数据分成若干个类别，使得类别内部的差异尽可能的小，类别外部差异尽可能大。



聚类分析的种类

层次聚类法

非层次聚类法

首选方法

二阶聚类法

K-均值聚类法

连续变量

样本量>1000

数据检查

描述统计量					
	N	极小值	极大值	均值	标准差
年龄	79132	0	110	23.93	6.543
成交笔数	79132	1	759	3.62	6.160
成交金额	79132	1.01	32,866.43	192.76	354.07
有效的 N （列表	79132				

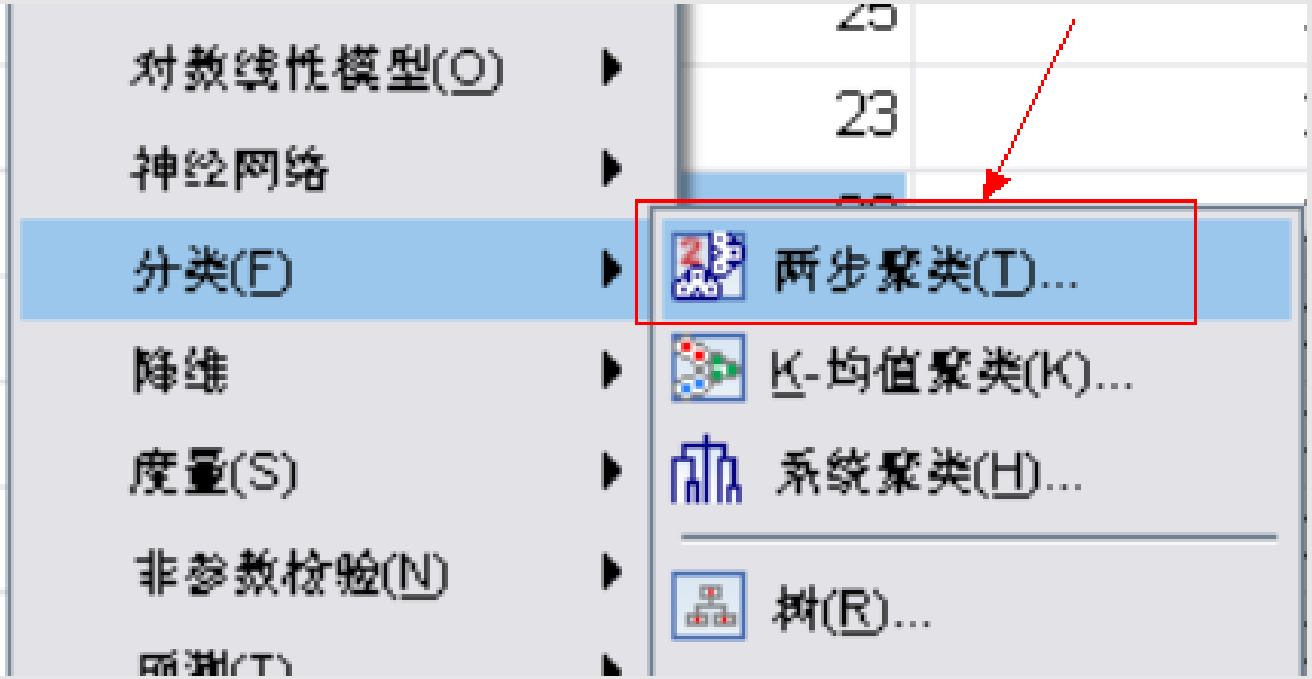
样本选择：

年龄：17——31 （24+13， 24-13)

笔数：1——16

金额：1——900

SPSS 两步聚类分析



Recommend
火龙果软件

AA聚类分析结果

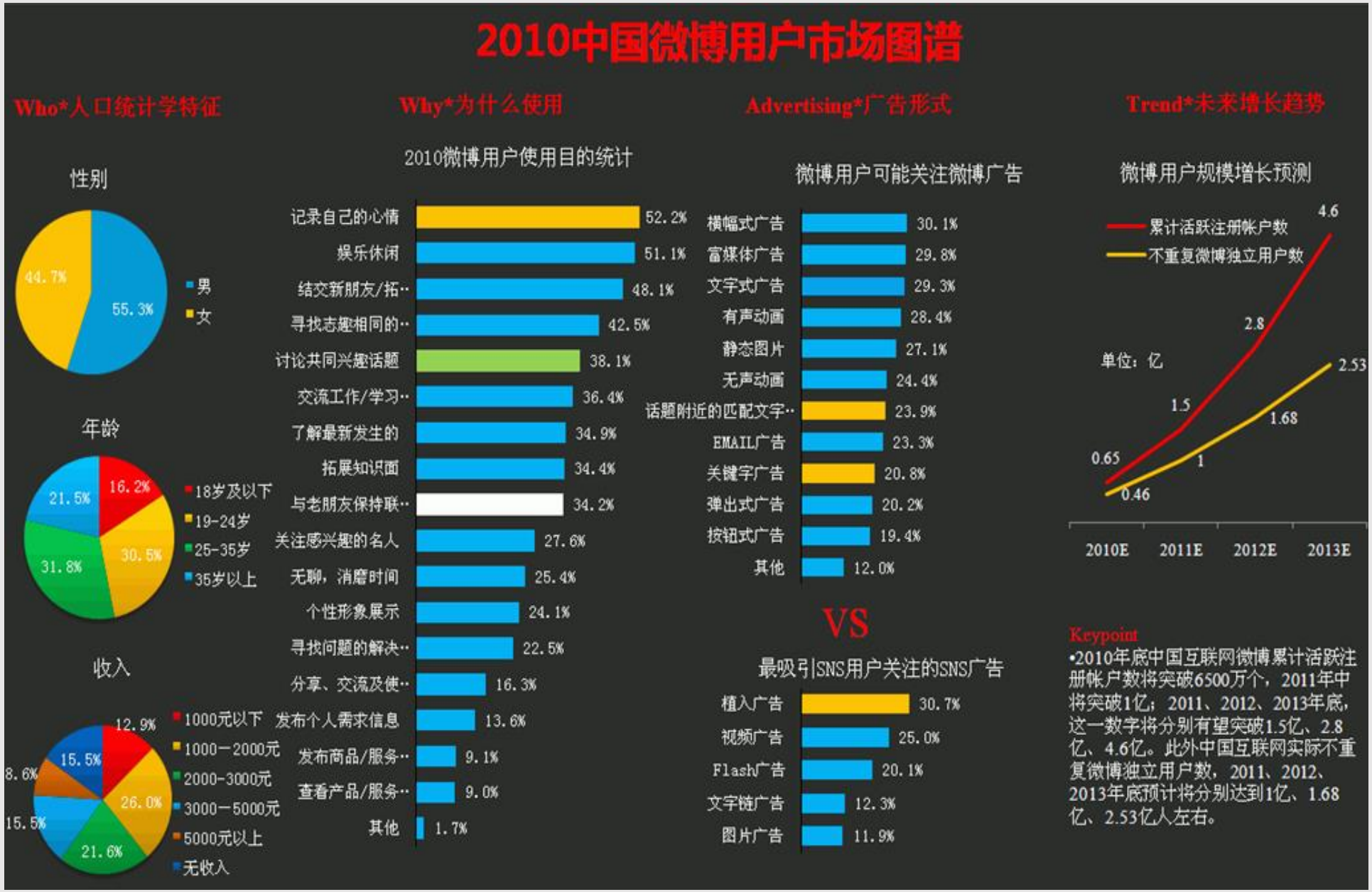
		年龄		成交笔数		成交金额	
		均值	标准差	均值	标准差	均值	标准差
聚类	1	23.42	4.773	2.12	1.312	104.69	76.87
	2	26.47	12.232	6.84	2.340	412.18	308.48
	3	23.52	4.774	1.73	1.047	97.12	73.01
	组合	23.91	6.527	2.67	2.254	147.06	176.03

用户画像在工作中的实际应用

Recommend
火龙果软件

- 抽样依据
- 产品优化
- 用户拉新
- 产品定位
- 用户挽留

Recommend
火龙果软件

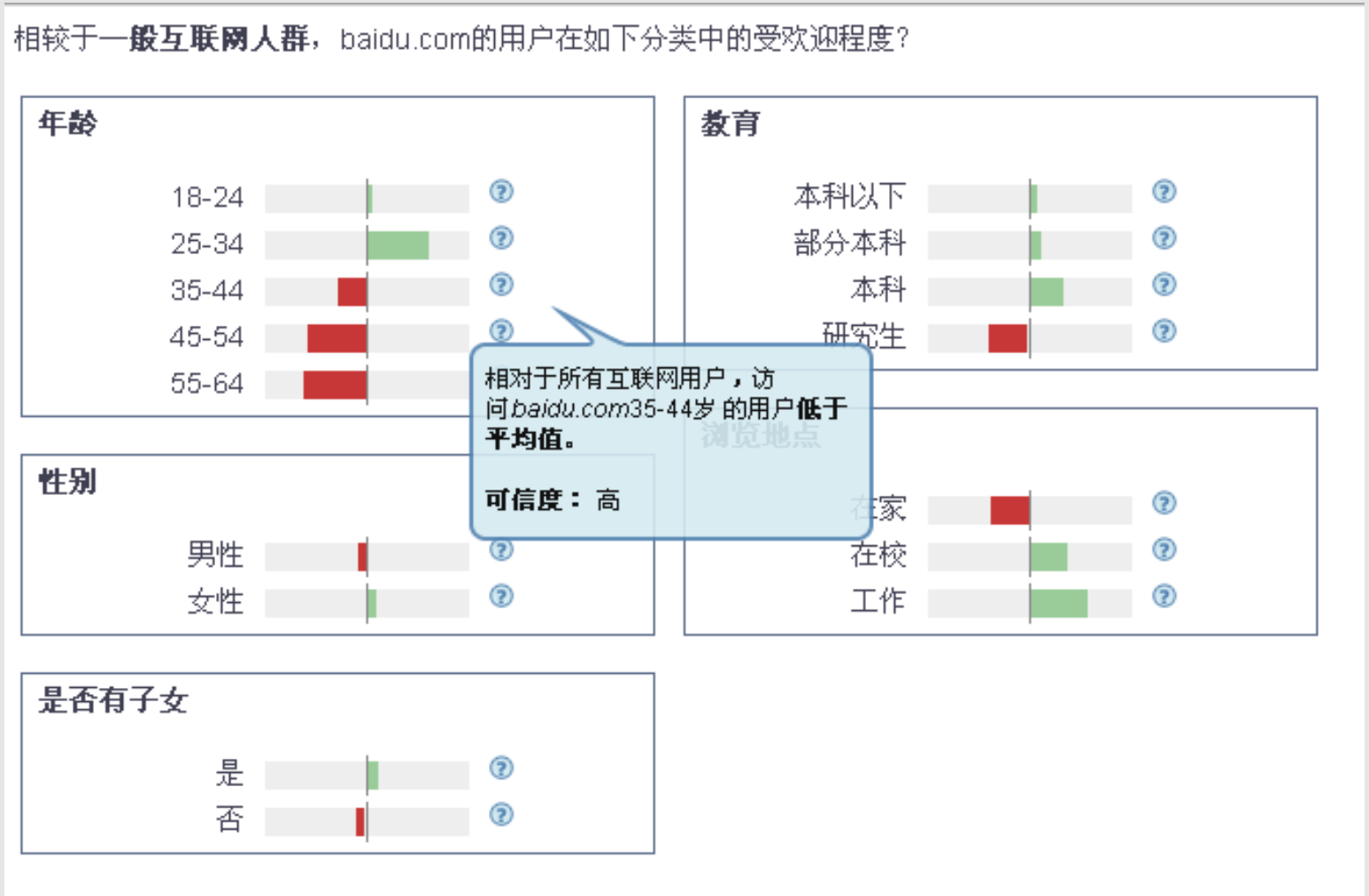


来源; KM平台，**中国微博用户市场图谱：2013年微博用户将突破2亿** http://km.oa.com/group/693/article_view/56793

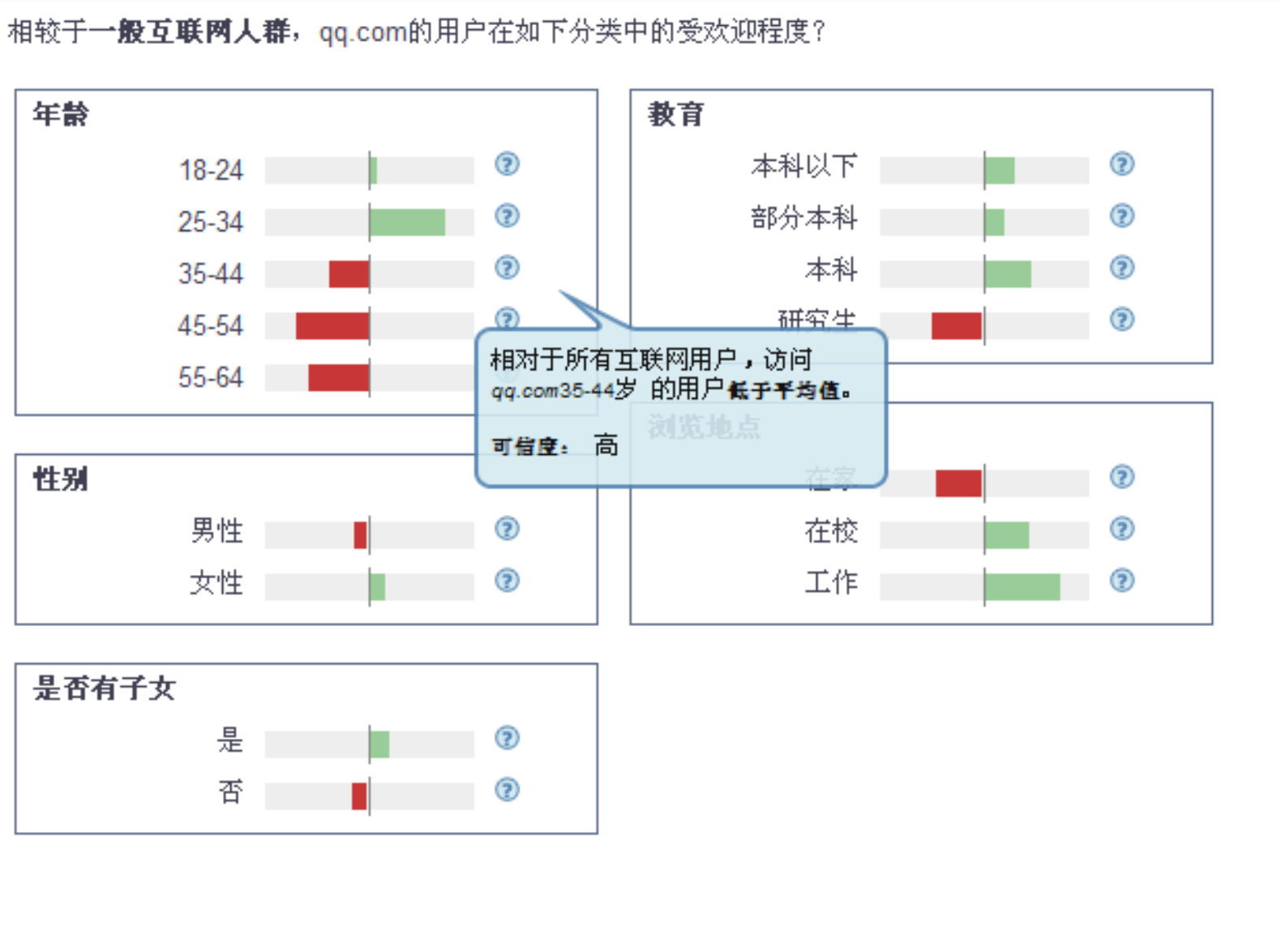
可以看看的几个网站

- **统计专业网站：http://www.statistics.com**
- **UCD大社区：** http://ucdchina.com/
- **Alexa中文官方网站：** http://cn.alexa.com/
- **淘宝用户体验设计：** http://ued.taobao.com
- **Excel Home:** http://www.excelhome.net
- **图研所：** http://www.tuyansuo.com/

BAIDU.com 人群——cn.alexa.com



XX.com人群——cn.alexa.com



GOOGLE人群——cn.alexa.com

Recommend
火龙果软件

相较于**一般互联网人群**，google.com的用户在如下分类中的受欢迎程度？

年龄

18-24
25-34
35-44
45-54
55-64
65+

教育

本科以下
部分本科
本科
研究生

性别

男性
女性

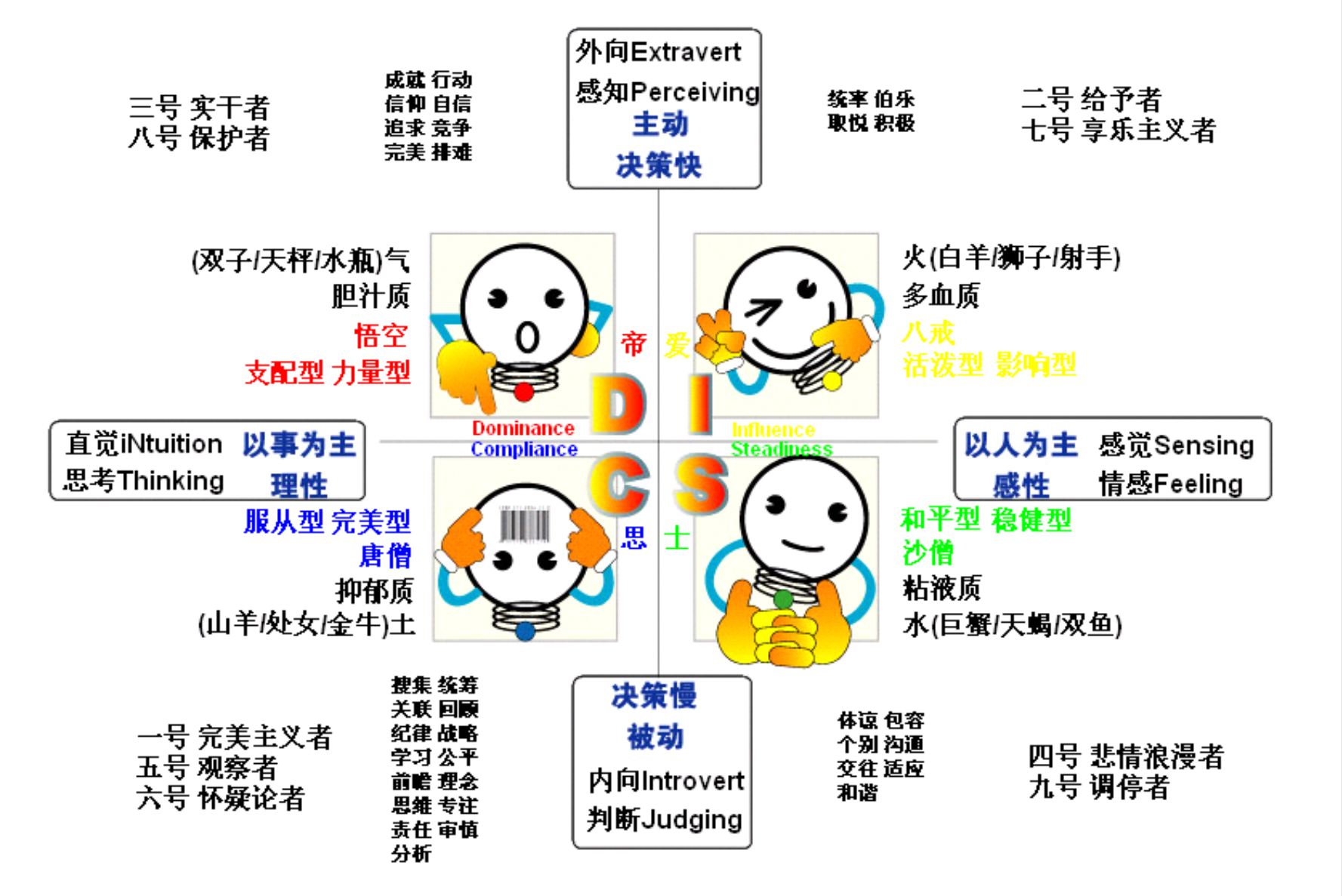
浏览地点

在家
在校
工作

是否有子女

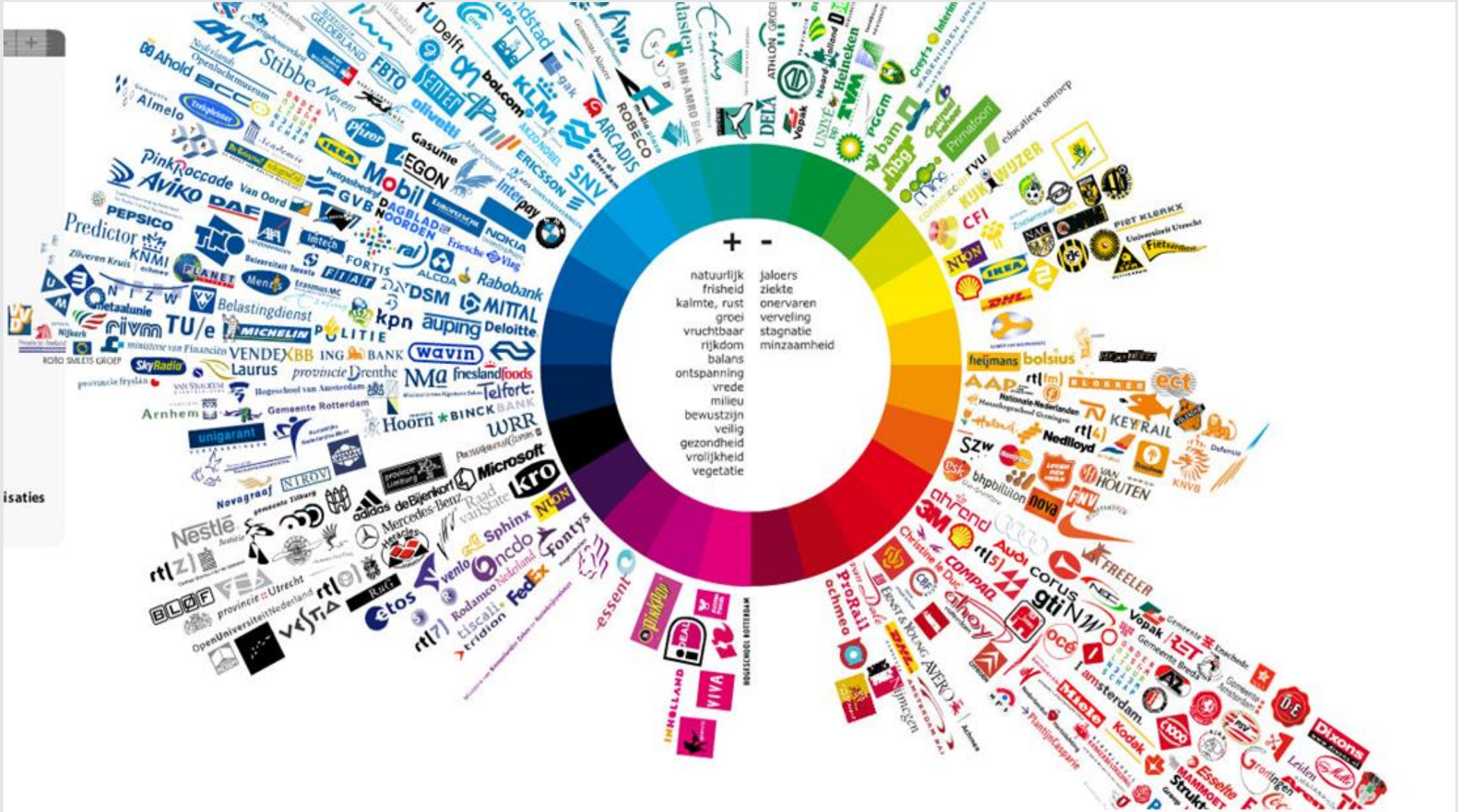
是
否

星座与气质用户画像

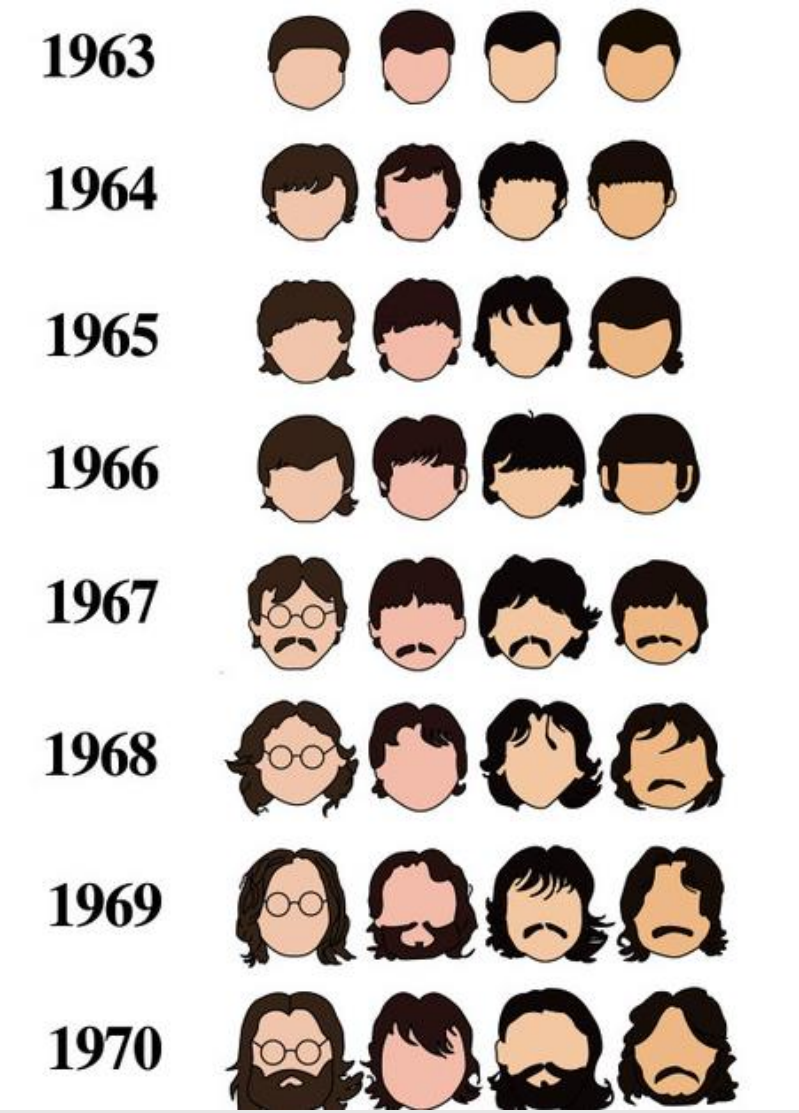

外向Extravert
感知Perceiving
主动
决策快
三号 实干者
八号 保护者
成就 行动
信仰 自信
追求 竞争
完美 排难
统率 伯乐
取悦 积极
二号 给予者
七号 享乐主义者
(双子/天秤/水瓶)气
胆汁质
悟空
支配型 力量型
帝
爱
火(白羊/狮子/射手)
多血质
八戒
活泼型 影响型
D
I
Dominance
Compliance
Influence
Steadiness
直觉iNtuition
思考Thinking
以事为主
理性
以人为主
感性
感觉Sensing
情感Feeling
C
S
思
士
服从型 完美型
唐僧
抑郁质
(山羊/处女/金牛)土
和平型 稳健型
沙僧
粘液质
水(巨蟹/天蝎/双鱼)
决策慢
被动
内向Introvert
判断Judging
一号 完美主义者
五号 观察者
六号 怀疑论者
搜集 统筹
关联 回顾
纪律 战略
学习 公平
前瞻 理念
思维 专注
责任 审慎
分析
体谅 包容
个别 沟通
交往 适应
和谐
四号 悲情浪漫者
九号 调停者

Recommend
火龙果软件

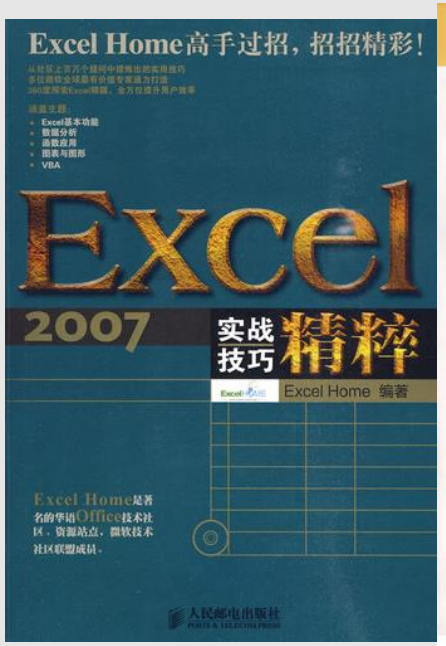
品牌画像——颜色维度



甲壳虫乐队发型变化图表



可以读读的几本书



￥45.00 ￥88.00 ￥30.96 ￥28.00

收集数据的十个原则

一、思考研究问题之初，也要开始思考回答问题需要收集的数据类型。

二、在思考数据类型的时候，也要思考从何处获得数据。

三、保证你用于收集数据的表格简单容易。

四、始终记得对数据文件进行备份，并存储在不同地方。已经丢失&即将丢失。

五、不要依赖他人收集或转换数据。

六、计划详细的何时何收集数据的日程表

七、只要有可能，就为你的项目培育可能的数据来源。

八、尽力追踪遗漏了的测试或者访谈对象。

九、永远不要销毁你的原始数据。

Recommend
火龙果软件

Recommend
火龙果软件

汇报完毕 感谢聆听