大数据视野

袁岳博士

2015,9,11

大数据时代的数据，够大

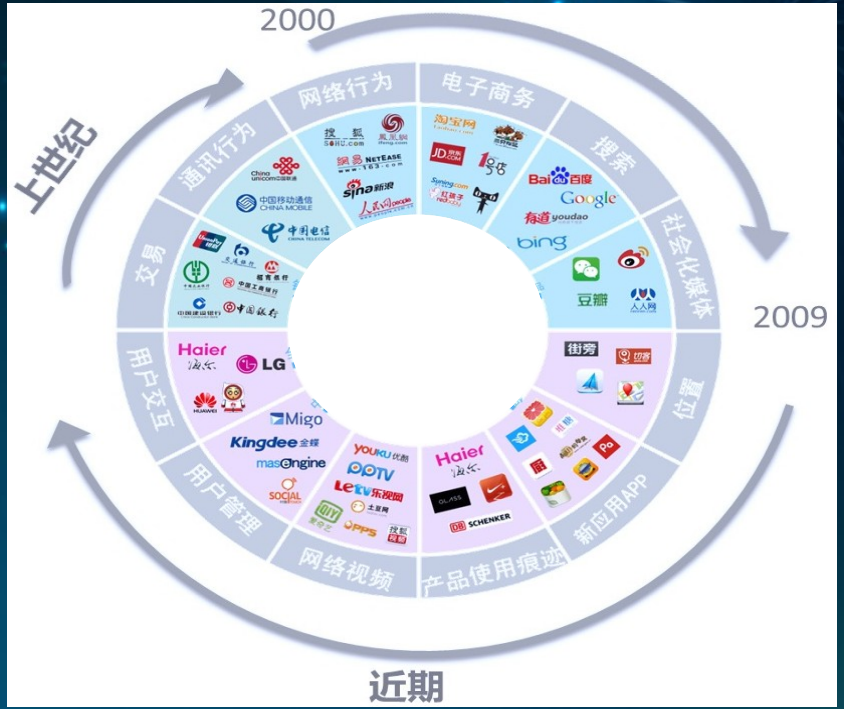

What Happens in an Internet Minute?
苹果的应用下载4.7万次
Pandora新增6万多首歌曲
发出2亿封邮件
Amazon销售8.3万美元
Flickr新贴3000张照片
Flickr有2千万次访问照片
Twitter上载10万条新微博
Facebook发生6百万次访问，新增28万用户
Twitter新增320个帐户
Google发生2百万次搜索查询
YouTube上载30小时的视频，发生130万次观看
639,800 GB of global IP data transferred
20
New victims of identity theft
47,000
App downloads
204 million
Emails sent
$83,000
61,141
Hours of music
3,000
Photo uploads
100,000
New tweets
135
Botnet infections
1,300
New mobile users
100+
New Linkedin accounts
277,000
Logins
6 million
Facebook views
6
New Wikipedia articles published
2+ million
Search queries
30
Hours of video uploaded
1.3 million
Video views
全球IP网一分钟传送639TB
And Future Growth is Staggering
Credit: Intel Corporation
Today, the number of networked devices
=
the global population
By 2015, the number of networked devices
=
2x
the global population
In 2015, it would take you 5 years
IP
to view all video crossing IP networks each second
在2015年要用5年时间才能看完在互联网上一秒内所传的视频

大数据时代数据，够广

新数据源

不断产生

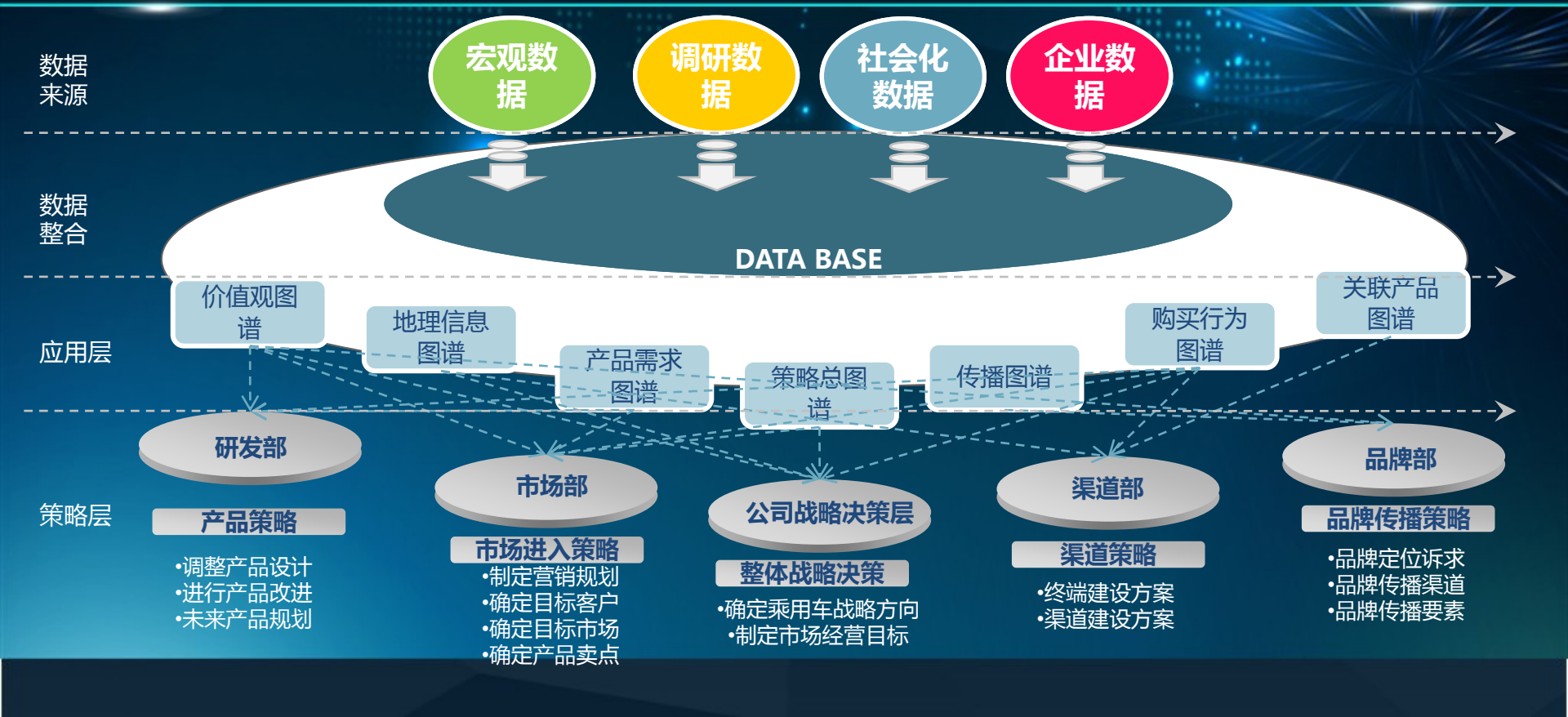
并继续产生



但现有的数据，不够联

没有跨界的大数据不是真正的大数据

打破数据孤岛，构建数据价值

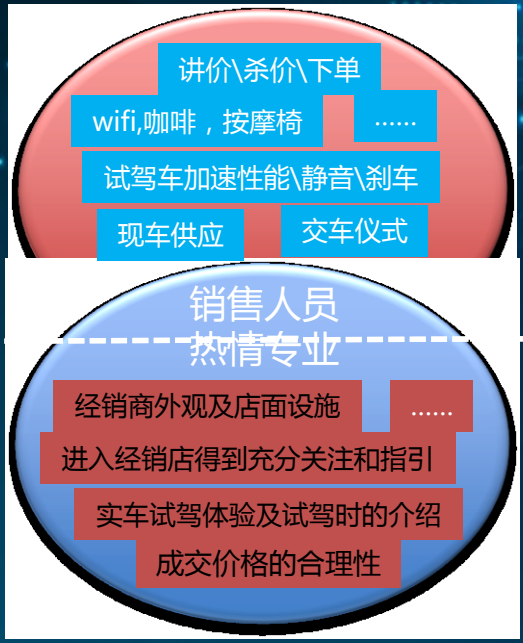


网络论坛&线下用户调研结合，信息互补形成新洞察

网络论坛

某品牌-2012网络论坛监测论坛声量分布

信息分布



网络评估信息：

反应用户感受、感性触点，与促进成交相关内容多

线下用户调研

某品牌-2012车主调研购车信息源分布

购买决策信息分布

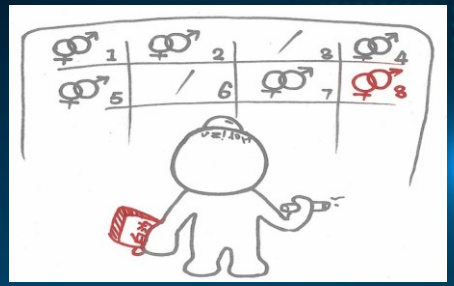
调研评估信息：

侧重企业规范、与管理相关内容多

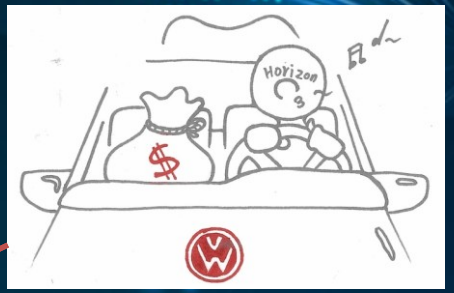
两种信息综合运用才能让4S店管理

“情理”兼备

数据贯通，洞察生活方式图谱



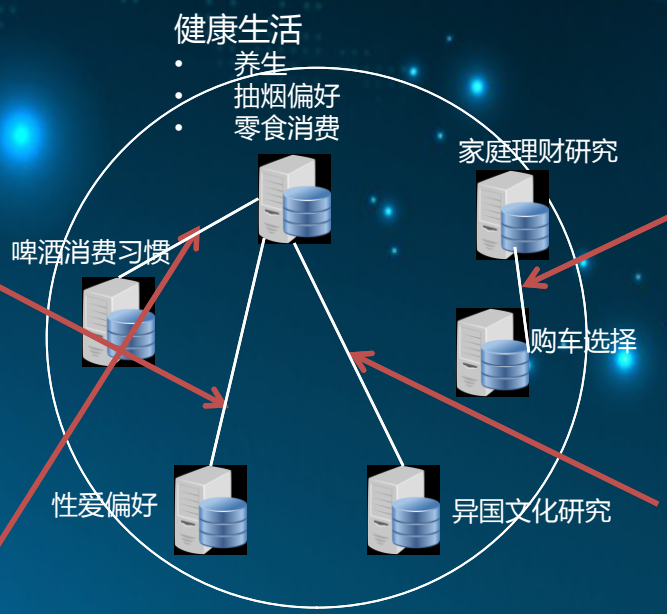
抽白沙烟的人，性爱更频繁



开德系车的，更爱藏私房钱

喜欢哈啤的，更喜欢零食

爱韩国，爱养生



某地产企业，整合多渠道客户信息

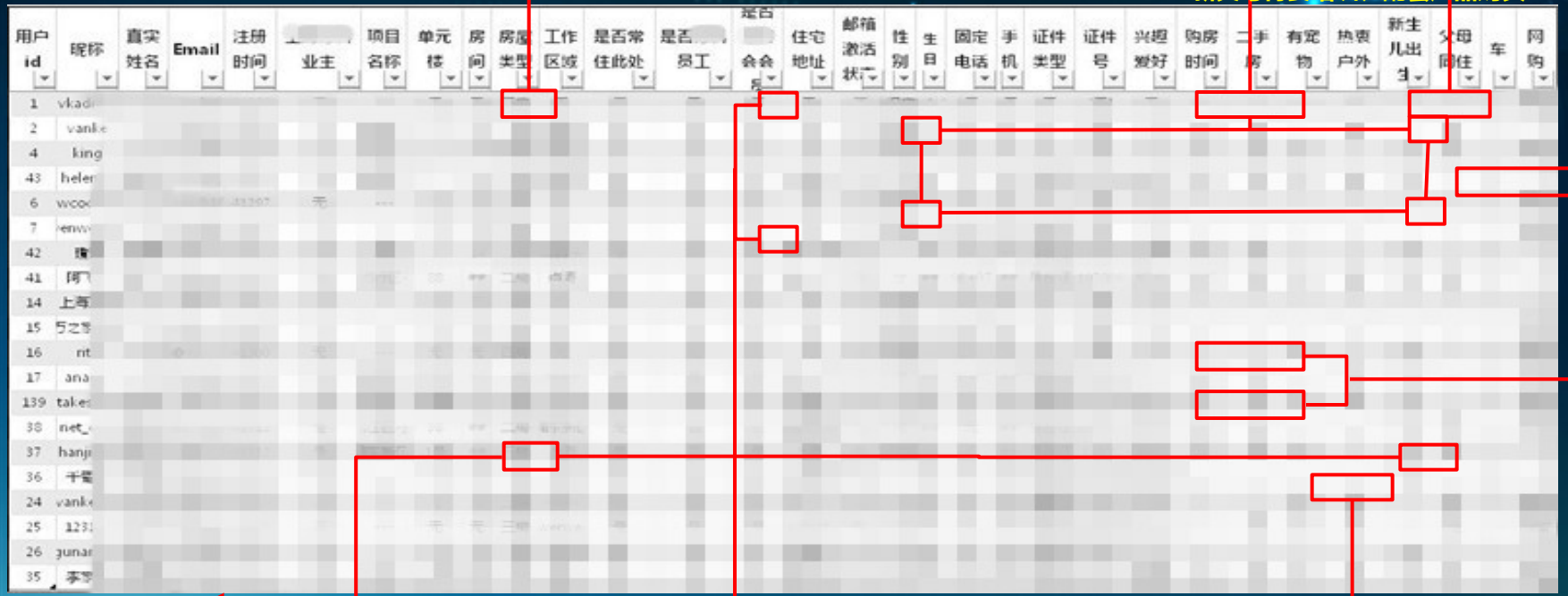
企业内部与社会化信息

客户信息数据

一体化信息平台

对数据库整合分析，为目标人群设计需求产品，并进行营销沟通

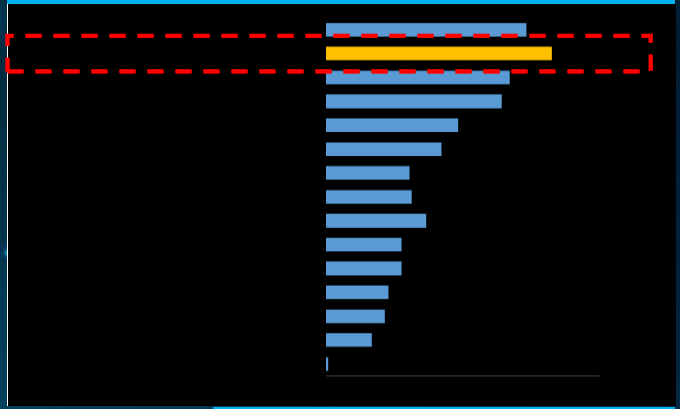
通过多渠道信息，找到企业新价值驱动



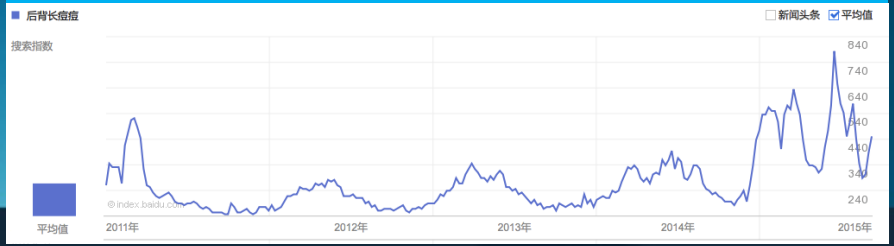
某企业通过多元数据，获得创新产品概念

通过线上社媒（微博／论坛）碎片化信息萃取，发现很多用户在洗澡时都有“搓澡”的习惯

《零点2014年城市居民洗浴习惯调查》：一个人洗澡时搓背很麻烦是近5成北方城市居民在洗浴过程中普遍遇到的问题



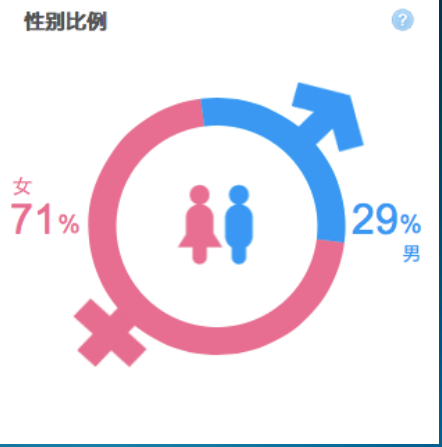
百度指数显示，较历年相比，2015年网民对“后背长痘痘”的关注程度明显提升，远高于平均值



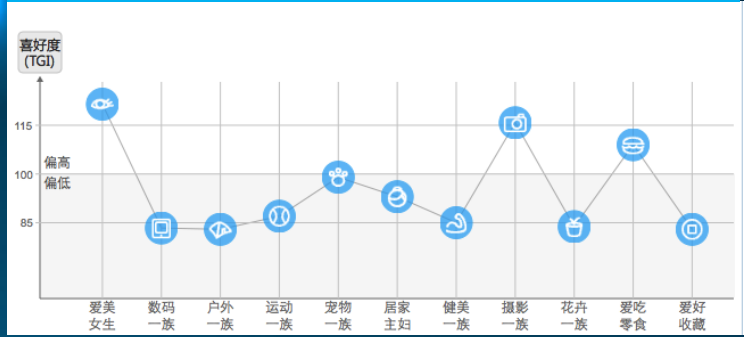
一款新品洗脸刷在淘宝／天猫热卖，2015年七月爆发式增长

某企业通过多元数据，获得创新产品概念

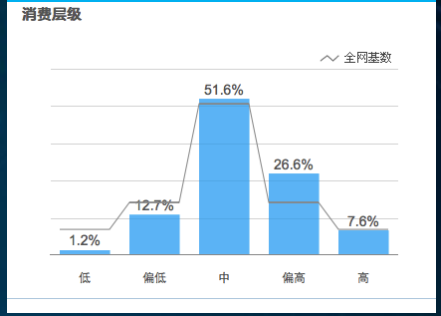
女性为主



爱美、爱拍照



中等偏上收入



某企业通过多元数据，获得创新产品概念

皮肤清洁和花洒合二为一

可拆卸刷头：
拆掉刷头就是花洒，安上刷头就是洗背神器。可以不同家人用不同的刷头

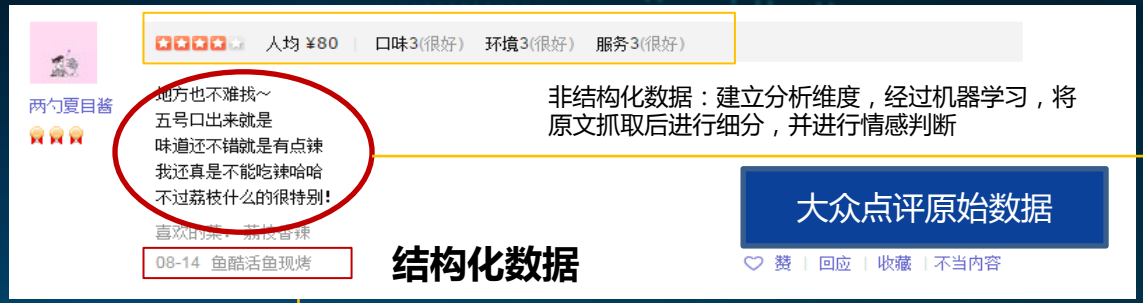
采用**超声波摇摆振荡技术**：
以温和摇摆方式与肌肤天然弹性相结合，在深入清除油脂和污垢的同时，将摩擦降至最低，保护肌肤胶原蛋白。

强度挡选择——清洁强度分温和，中等，加强，方便使用者自选；

提醒功能——根据医学原理测算皮肤耐受该产品清洁的最大时间；到达阈值刷子会自动声音提醒，一分钟自动关闭，也可以保护产品过热损坏。

餐饮企业，通过网络信息抓取，解决企业经营困境

智能信息抽取--将杂乱无章的文本变成可交叉分析的多维度变量



非结构化数据：

- 地理位置评价信息：地方不难找/五号口出来就是
- 味道评价信息
  - ✓ 口味种类评价：味道不错/有点辣
  - ✓ 口味合适度评价：不能吃辣
  - ✓ 口味独特性评价：荔枝什么的很特别

结构化数据：对结构化数据（星级评价、人均、日期、推荐菜、品牌、店铺名称）进行抓取并归类和统计。

结构化数据归类整理

社媒数据数据库

原文抓取存放

分维度进行分析

信息抽取引擎技术优势

- 采用Semi-Markov Model，可做结构化预测
- 通用引擎，可配置额外内容抽取
- 词典与上下文融合，与Wiki融合
- 综合词语、语法、语义特征
- 分布式API，单机可达1000篇/秒

关键词方法仅能获得已知的实体存在。只有**拥有语义分析能力**的实体抽取引擎才能**自动的根据文章所表达的内容**对实体信息进行抽取，从而**发现未知信息**。

餐饮企业，通过网络信息抓取，解决企业经营困境

品牌优劣势分析 - 具体标签分析

可以从聚类标签来看，烤鱼入味、味道浓是顾客认为口味好的主要内容；而导致他们负面评价的主要原因是，口味太咸、太酸、太辣是主要不满点

某品牌口味维度正面评价内容

聚类标签	提及次数
口味较浓	889
烤鱼很入味	706
味道特别浓	68
鱼肉很入味了	32
浓香入味	20
而且很入味	17
老酸菜味道很足	14
还有汤汁也很够味道	7
汤汁很入味	6
整体感觉川味很浓重	6
浓郁的重口味	4
烧的倒满入味的	2
口味正宗	103
味道很纯正	32
味道正宗	22
正宗的味道	18
味道正宗	23
但是味道很纯	4
很纯的味道	4

聚类标签	提及次数
口味淡	189
味道比较淡	125
口味比较淡	36
蘸汁太淡	8
汤味道很淡	7
味道清淡比较适合我们的口味	7
味道以清淡为主	4
吃惯了清淡口味的	2
酸辣程度	593
酸酸甜甜的很爽口	197
酸酸甜甜	263
自制酸奶很酸	124
酸的够劲爆	3
沙棘汁酸酸的解腻	2
小甜小酸	2
香香苦苦甜甜的	2

某品牌口味维度负面评价内容

聚类标签	提及次数
偏咸	444
就是有点咸	319
就是偏咸	102
唯一的缺点就是咸了点	7
可是数量咸没别的了	5
封缸肉豆腐太咸了点	3
那天的汤咸到不忍下咽	2
奶茶太咸	2
其中有个榆林豆腐汤感觉有点咸	2
蒜泥西兰花稍微有点咸	2
偏酸	851
酸奶很酸	699
酸奶太酸了	121
酸死我了	23
而酸味不够	3
单吃酸奶酸涩感略重	3
只是最后上的酸奶就太酸了	2
偏重	49
功夫鱼软烂入味	34
口味略重	9
味道都算比较重的	4
菜色口味有点重啊感觉	2

聚类标签	提及次数
偏辣	48
稍微有点辣	27
大盘鸡太辣	4
味道是小点儿小辣	4
不够辣	2
只辣无其他味道	3
包心菜有点辣	2
凉菜几乎都是辣的	2
每个菜总是有那么点辣的	2
该辣的不辣	2
偏淡	132
味道还淡	101
就是蘸酱太清淡了	17
有点粤菜的清淡	2
寡淡无味啊	2
羊肉串也不太入味儿	8
汤味道不足	2

餐饮企业，通过网络信息抓取，解决企业经营困境

品牌优劣势分析 - 留存流失*原因分析

通过他们提及的下次再来、多次来店分析得出，导致他们留存的主要原因是口味好，一如既往地好吃而且选择较多

2014-2015年各品牌留存/流失比*（%）

	2014-2015整体*	某品牌	B品牌	鱼酷	C品牌	D品牌	E品牌	F品牌
多次来店	16.1	22.0	20.7	18.2	13.3	11.5	13.0	9.4
下次再来	11.7	12.1	11.3	11.8	13.0	11.6	12.1	12.2
留存*（多次来店+下次再来）	27.8	**33.4**	31.3	30.0	26.3	23.1	25.1	21.6
流失*（下次不来）	4.3	**4.3**	4.7	6.8	4.1	3.7	3.6	3.2
留存流失差	23.6	**29.1**	27.6	23.2	22.2	19.4	21.5	18.3

聚类标签	提及次数
味道好吃	667
一如既往地好吃	549
口味选择多	127

某品牌的留存原因（n = 6,323条，%）

注：【2014-2015年各品牌留存/流失比】指抓取时间为2014年1月份-2015年7月份的各个品牌留存/流失比例。其中，某品牌留存比=某品牌评论中提到“多次来店” “下次来店”的评论数量/某品牌总人气；某品牌流失率=某品牌评论中提到“下次不来”的评论数量/某品牌的总人气；某品牌留存流失差=某品牌留存比-某品牌流失比

全面互联 始终有数