百度认证营销专家申请者论文

题目：内容大数据在汽车行业的应用及价值

姓名：王　郅

序号：04

2017 年 4 月 7 日

题　目

《内容大数据在汽车行业的应用及价值》

摘　要

搜索最本质的强大能力是抓取迎合用户的内容进行排序，提供有价值的流量。搜索结果页中，除去商业推广结果的分流外，另一部分核心流量化成了具有粘性的内容。通过结构化数据的采集，进行机器映射和人工甄别，对于内容语料进行细腻的关注点变化和情感分析，创新应用为汽车营销投放前、中、后期提供商业洞察服务。同时结合未来百度自然语言学习将其推及到数据工具和精准投放中，提高数据能效的同时辅助广告产品提升点击转化率。

关键词： 活大数据、内容大数据、语料分析、舆情监测，自然语言学习

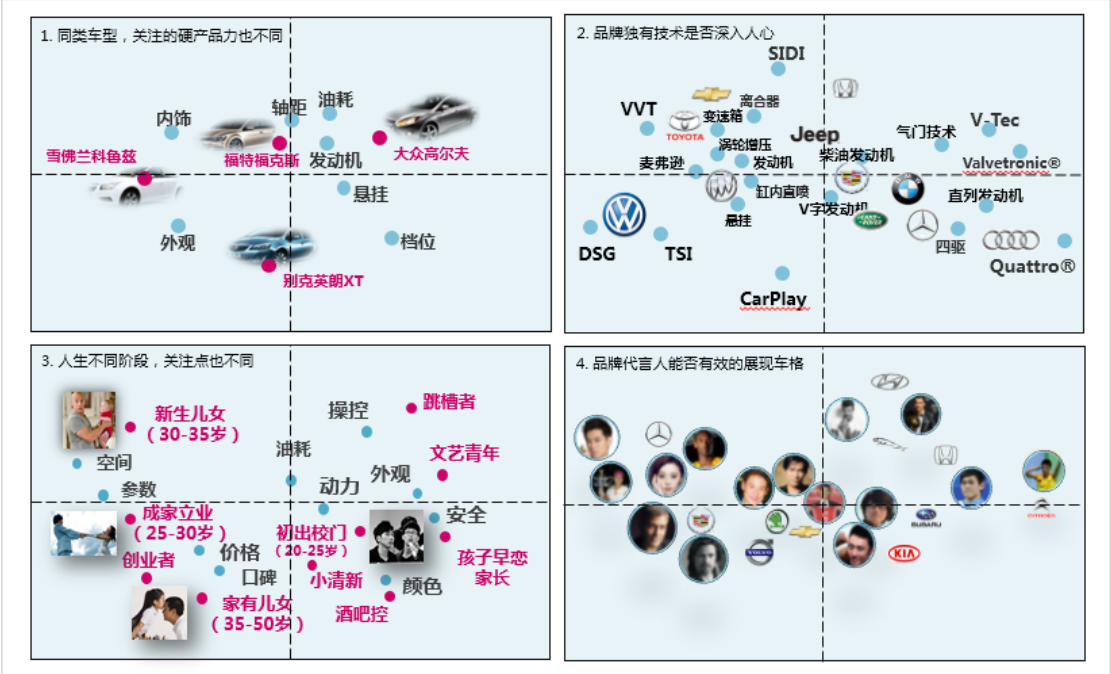
目 录

第一章：汽车行业传统搜索大数据

1.1 汽车行业传统搜索大数据应用的主要特点

先来看一组出自百度汽车行业峰会数据：“2014 年汽车行业相关关键词检索量为 3.4 亿，其中 1.6 亿为无线端检索量。”[1] 这是百度汽车行业第一次对外公布无线端用户的搜索量已占半壁江山，到 2017 年无线搜索已占 70%；“子品类方面 2014 年 SUV 同比增长 69%，新能源车品类同比增长比 358%。”事实也证明了，此后各大厂商均在 SUV 车型上做了密集的产品细分，无论是在市场关注度、搜索量或者销量上都呈现井喷，而新能源车品类更是在这两年成为了政府扶持和厂商发展的主旋律。这类大数据主要起的作用是**市场预测（Trend to Forecast）**，呈现的是**趋势性。**

2014 年汽车行业峰会还呈现了一组更形象的大数据（图 1）：“当选取四款热门紧凑型车，观察风格相似的车型在用户心中的关注点差异化时，可以看到用户对于核心技术的了解与品牌关联性。如法炮制还可以推及到人生不同阶段的关注点偏好，用户心中的品牌代言人等个性化洞察。”[2] 这组大数据是基于用户（cookie）历史行为的拓展分析，作用是**用户挖掘 （Relevance for Mining）**，呈现的是**关联性。**



（图 1）

到了 2015 年百度汽车行业峰会，公布了三个更为新颖的搜索大数据：

第一，“在北京，百度地图 LBS 日请求数为 150 万，其中有 100 万出现在日出行高峰时段。”

第二，“在上海，拍牌相关的需求检索量连续 16 个月走高。”

第三，“某家大型车企的官方整体降价公布后，蝴蝶效应震荡了全国汽车行业，“车型价格”相关的需求词搜索量在当月上涨了 40%。”[3]

这组大数据的变化集中地反映了当下用户生活的聚焦点。大数据与实事民生相关的**即时呈现（Instant from Native）**也是一个重要的**特征，**呈现的是**原生性。**

上述的应用也就呈现出了汽车行业传统搜索大数据的三个主要特征：**趋势性、相关性、原生性。**

1.2 汽车行业致力研究基于搜索的“活大数据”

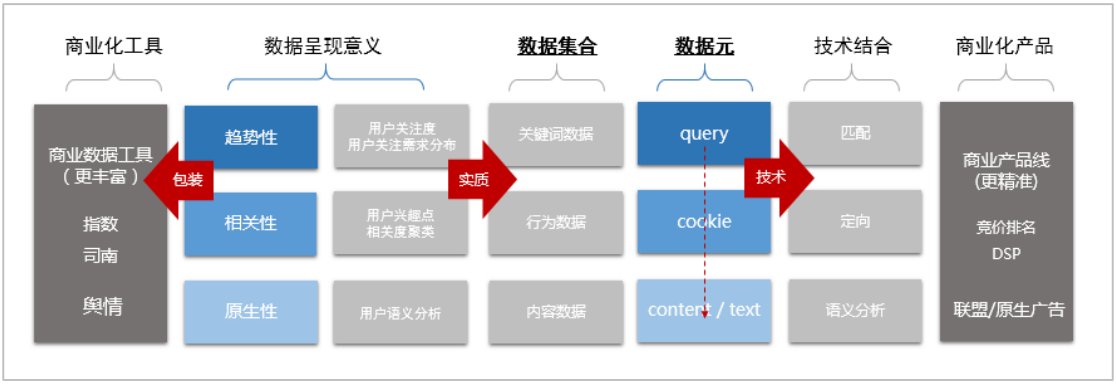
那么，具备了第一章 1.1 中所具备的三个特征的大数据是否就是“活的搜索大数据”呢？所谓**活的搜索大数据（Rolling Data）**，呈现的不能只是用户的片段式数据，不能只是用户的单维度数据[4]。

例如，汽车行业的广告主在搜索广告投放前、中、后期，都会关心百度指数这样**不断更新的量化数据**，从而直观了解其车型在互联网上的市场热度。而不是关心投放过程中的百度数据波动；再例如，汽车行业的广告主在了解 PC 端受众的 cookie 搜索路径行为后，也会同时想要了解 Mobile 端 ID 数据打通受众的搜索路径。这样才能构成所谓的**用户多维度同源数据**。“阿里巴巴也认为通过投放营销、购物消费、PC/移动设备兴趣轨迹、生活方式、视频娱乐、地理轨迹、社交评论等。一个汽车用户应该拥有 2.8 个 Cookie，1.7 个移动设备。通过同源数据与汽车用户沟通，才能提升效率。”；还有，从同源数据拆分的一种的维度，**内容化数据，它具有不断再生的属性**。社会化数据的更新十分快速，因为数据的更新方式用户自产，数据的呈现也十分透明，所以就不需要数据等待和运算呈现。

现在看来，能够不断更新的量化数据，能够不断丰富维度的同源数据，不断积累的 UCG 内容数据，具备这些属性的数据可以被叫做“活大数据”。本文就是针对这第三种数据进行创新探索，之后的第二章会就其灵活性和可塑性，进行一些高级分析和创新应用，对象就是基于搜索的内容大数据。

1.3“活大数据”对汽车行业搜索营销的服务作用

基于传统搜索大数据，<u>一方面，数据元经过整合分析能够呈现相应的营销意义，加以包装则可成为商业化的大数据工具，服务于汽车广告主。另一方面，数据元配上百度的技术可以开发成百度的商业推广产品或者某些产品技术功能点。所以，优秀的大数据是可以为数据工具的研发和商业产品的功能做贡献的。</u>



（图 2）

刚刚提到的某车型关键词或某车型关键词词包在百度的检索量、PV 数据经果算法加权后，已经被包装成像百度指数/专业版指数这样家喻户晓的数据工具，被广泛的使用来预估某个对象在百度的流量波动或流量量级服务关键词投放。

刚刚提到的基于圈定用户 cookie 相关的兴趣点和搜索路径则被包装成了百度司南工具，成为了准备 campaign 前的必须的人群画像部分（图 3）。汽车广告主也会参考用户的实际画像来选择精准广告投放中的兴趣偏好，提升效率。

而目前，只有内容大数据还没有被很好的整合或者使用。一直以来，百度似乎不是以内容见长的平台，但其实作为搜索引擎本身有这一个很大的技术优势，便是在于对于内容的抓取，梳理和排序。也就是说，庞大的数据只是“寄存”在百度的搜索平台中，要看到内容大数据就必须“破壳获取”。

【小结】汽车行业传统搜索大数据具有趋势性、相关性、原生性。如今百度已经成为一个多元化的服务平台，随着百度各种平台的技术整合，能够产生各种特色大数据，丰富同源数据的维度。而大数据的本身也是一条产业链，不难看出此前从数据元到数据分析的用户行为，到被包装成商业产品或者工具。**研究会用基于搜索的社交化内容证明：汽车行业活大数据是灵活度高，可塑性强的。**

第二章：内容大数据的应用分析法

2.1 内容大数据应用的理论依据：“S2C 思想”

“从搜索到内容“的思想，即从百度自身搜索平台和汽车行业垂直网站、论坛内抓取内容数据进行分析。那么为什么会有这个 Search to Content（以下简称 S2C）的思路的呢？

第一，基于搜索引擎的本质属性

搜索引擎最本质的强大能力是抓取迎合用户的内容进行排序，提供有价值的流量。在汽车用户的搜索结果页中，除去商业推广结果的分流外，另一部分核心流量化成了具有粘性的社会化原生内容。

第二，基于流量分布的二分法

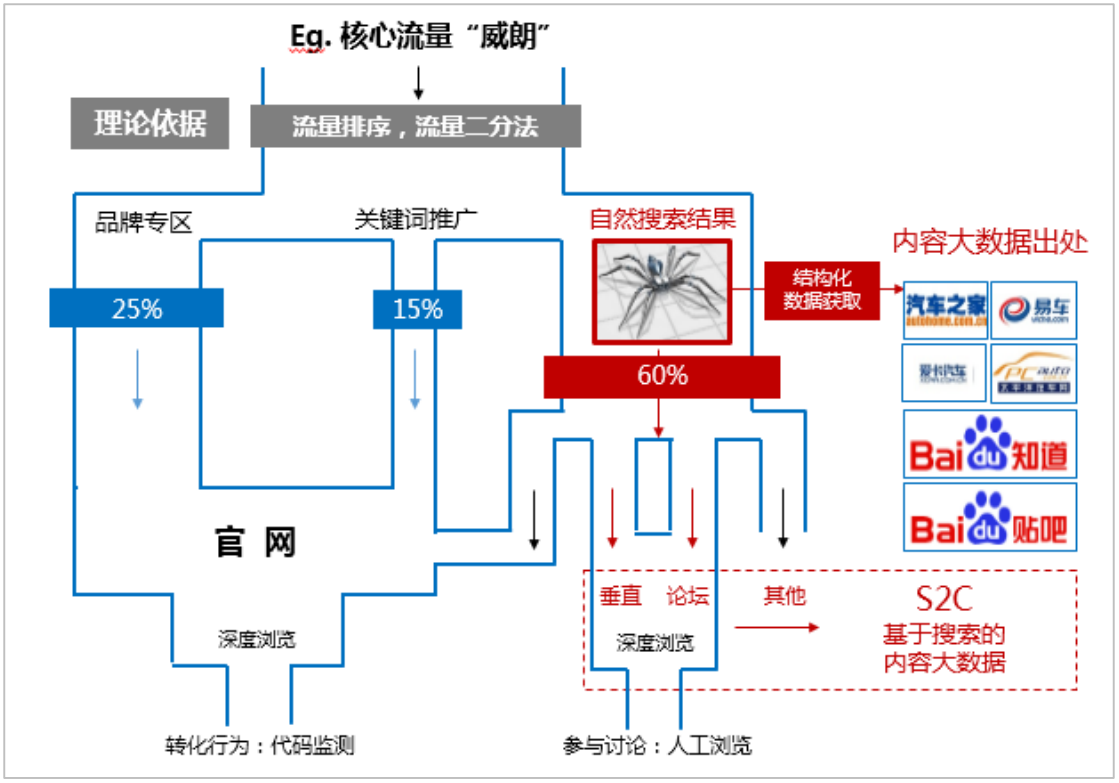
当以直观的搜索结果来看，一个用户搜索“别克威朗”这样的核心词时，在百度搜索结果页的呈现依次会是（图 3）：

第一屏，别克威朗的品牌专区、别克威朗的 SEM 广告；第二屏汽车之家威朗车型详情页、别克威朗的官方车型详情页、爱卡汽车威朗车型详情页；第三屏，威朗百度贴吧、威朗百科、威朗图片；翻页，汽车之家威朗车型详情页、爱卡汽车威朗车型详情页、别克威朗新闻聚合模块等。

（图 3）

当以流量的角度来分析时，不难发现（图 4）：搜索结果页中，一部分通过百度产品线带去的核心跳转流量 A 和流量 B，可以用监测代码来追踪其深度浏览数据和行为。但是我们关注的，是另一部分去到自然搜索结果的核心流量 C，**现在可以通过 S2S 的思想来采集流量 C 下的指定内容进行深度监测，这部分流量通常聚集较大的汽车垂直网站，百度贴吧，百度知道等，即自然结果名列前茅的内**

容平台。利用搜索抓取的思想，借特定社交环境下的内容就是搜索社会化抓取的对象。



（图 4）

在聚拢了如此庞大的内容数据以后，对于语料的解析就是一个“痛并快乐”的过程，其中会挖掘一些属于**内容大数据的语料特质**，本文第二章之后的篇幅就会对其进行拓展分析。

2. 2 内容大数据应用的分析法：“OTA 三步法”

面对汽车行业的社会化语料的分析，研究致力于从三步来切入。

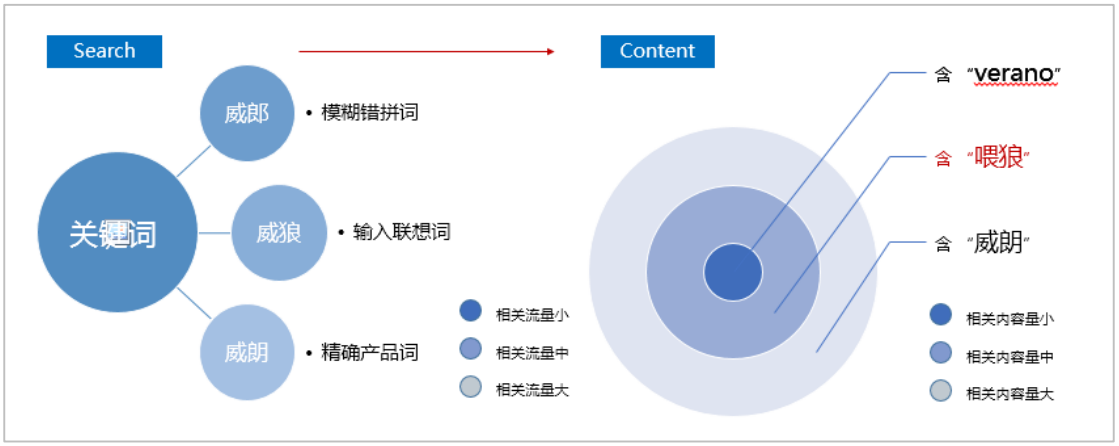
第一，如何决定汽车内容语料的**研究收集对象**？

2. 21 语料的对象（Object）

这一点就足以体现社会化大数据与传统大数据的不同之处。在关键词营销中，为了覆盖更多核心流量，会拓展一些**英文名、错拼词**和**输入法快速联想词**。如，威朗会拓展到威郎、威狼或者 verano。

而在社会化数据中，光光整入错拼词和英语名是不够的，还需要补充一些该**车型昵称**。例如，喂狼（别克威朗）；辣馒头（凌渡）。而这一类叫法通常在专业

社交平台中会引起共鸣，例如游戏论坛、体育论坛等。往往这样看似口水的叫法在内容量占有相当的比重。（图 5）



（图 5）

这里关于语料对象的词汇选择，我们要插叙一个社会化内容的特点，叫“行业黑话”。上过游戏或者体育论坛的用户都会很熟悉这类词汇，出于行业文化习俗与社交需要而创制的一些特色隐语，在游戏或者体育论坛中，比较多的存在于昵称。

例如，如果虎扑网上的头条标题是，**“辣棒鸡和萌库的对决”**，那么很容易就想到今晚是骑士队和勇士队的 NBA 焦点之战。如果搜狐视频上的头条视频是，**“大奉先上演帽子戏法”**，那么很容易就知道，今天晚上曼联靠伊布的三粒进球取胜了。再来看一个汽车行业的语料，例如：这么一比，**喂狼真的完爆辣馒头**了！

对于汽车行业来说，比较多的行业黑话就是对车型的昵称，这些昵称在垂直网站、贴吧等论坛上尤其容易被运用。例句中的意思是，在综合评测下，别克威朗的评分高于大众凌渡。如果不了解这种的黑话，在采集数据时很有可能会错过很大一段核心专业的评价内容。所以在这个基础上我们补充整理了汽车行业的现有各类车型的那些通俗叫法、谐音和昵称黑话。

第二，如何将收集到的内容语料进行**汽车关注点分析**？

2.22 语料的标签（Tag）

只要这个内容中有关注点，那么就有研究的意义，就可以自定义内容的标签维度。在这部分研究中，需要整理“汽车行业语料标签词典”。

词典是以表格的形式，主要分为“关注点”维度和“修辞”维度。如果判断

内容中含有“维度”词汇，且“修辞”词汇，则机器输出评分，由此该词典涵盖了 605 个不同“关注点+修辞”映射的结果（图 6）。

例如，该句含有“内饰”和“精美”这两个词，则程序默认这是一个好的评分，即会输出 2 分；如果该句还有“内饰”和“简陋”这两个词，则程序默认这是一个差的评分，即会输出 0 分；如果是一些中立的修辞或者其他一些机器无法辨识的情况，则输出分为 1 分，这些语料需要此后人工甄别再次判断评价的褒贬。

	维度	修辞	评价
131	动力	强	2
132	动力	好	2
133	动力	不错	2
134	动力	充沛	2
135	动力	足	2
136	动力	一般	0
137	动力	充足	2
138	动力	差	0
139	动力	完美	2
140	动力	弱	0
141	动力	出色	2
142	动力	不行	0
143	动力	小	0
144	动力	强劲	2
145	动力	不够	0
146	动力	不好	0
147	动力	要好	2
148	动力	还行	2
149	动力	不够强劲	0

（图 6）

第三，如何利用**机器和人工结合**的方式进行情感化分析？

2. 23 语料的映射（Algorithm）

内容聚集的自然语言中存在一些语境分歧。故有些目前内容只能靠人工来打标签甄别。如果程序无法输出中立的评价分，需要人工视情感语境甄别程度高低。以下两种情况在人工甄别的过程中较为特殊。

1） 流行语

例如：威朗的天窗简直是让人看得**蛋疼**！

“蛋疼，在语言表达应用此词，表示主观角度，外界其他事物对自己刺激较大，一时无法承受。” 这是百度百科词条对此的解释。像这样的流行语是层出不穷的，可能是因为某些热点事件，可能只是某些具有传播性的民生主题造成其一炮而红。

2） 反语

例如：威朗的价格低？我只能**呵呵**了！

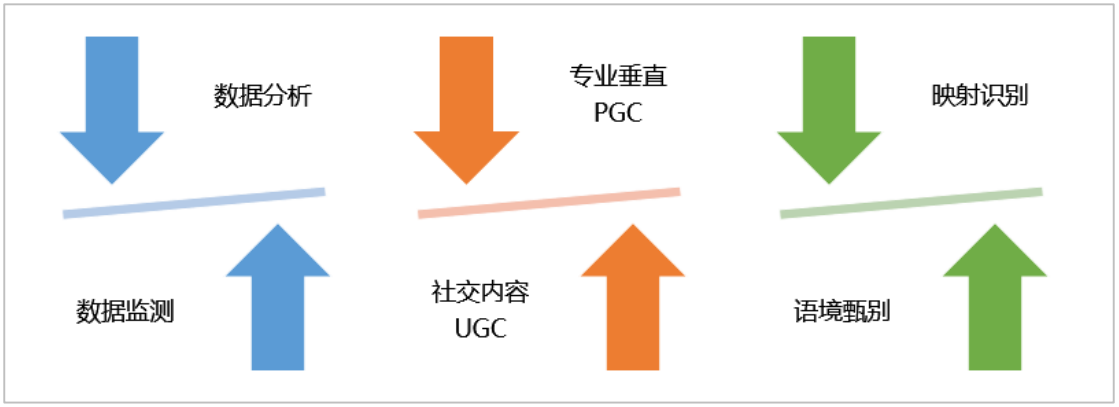
这部分的语义也是最难梳理的，倘若不结合上下文的语境很难判断这个词的褒贬性，程序更是无法准确直接判断其语义。

对于这些语词汇是随着时间而积累的，在积累的过程中无法穷尽，那未来这

就要依靠百度跨平台内容数据打通或者百度人工智能系统自然语言的完善和进步。既然百科词条中已有记载，相信这也已经成为百度的内容数字资产，可以被读取到相应的标签。

2.3 内容大数据应用的意义：“内容聆听”

像第二章中提到的研究方法，这样创新的内容大数据聆听，**1）即能起到数据分析的作用，又能体现数据监测；2）即聚拢了行业垂直网站的用户专业评论，又体现了社交化的原生内容；3）同时运用了机器语言的映射和自然语境的甄别**[5]（图 7）。事实证明了，这样的大数据可以贯穿于汽车营销的始终，并对广告主有着很多新颖的洞察。本文此后列举的两个案例就很好的体现了社会化内容数据的创新应用。



（图 7）

【小结】本质上，以搜索的流量排序和流量二分法为理论设想，确定汽车垂直类网站/社区及百度内平台为数据来源，制定出数据对应的标签规则，呈现创新的“内容聆听”用于分析。

第三章：内容大数据在汽车行业的应用案例

3.1 汽车投放中/后案例：上汽通用别克威朗“后视镜”项目

先来看一个内容大数据应用于车型投放售中及售后的实例。

案例一：上汽通用别克威朗新车上市

“后视镜”创新背景：在投放过程中紧跟监控数据量变背后的内容变化，高效做出投放策略调整，为未来的产品发展提供建议。

新车上市背景（如图 8）：

- 威朗为 2015 年别克新一代战略车型，于 6 月 27 日上市；
- 威朗定位高于英朗，低于君威，为别克紧凑级产品线中的旗舰车型；
- 威朗上市宣传主推包括：搭载 1.5T 涡轮发动机、7 速双离合变速器、风阻系数 0.27、Alcantara 运动座椅、高性能运动底盘调校、超大全景天窗、智慧互联系统等 15 个 FBI 维度……

（图 8）

基于广告主给到的背景信息，

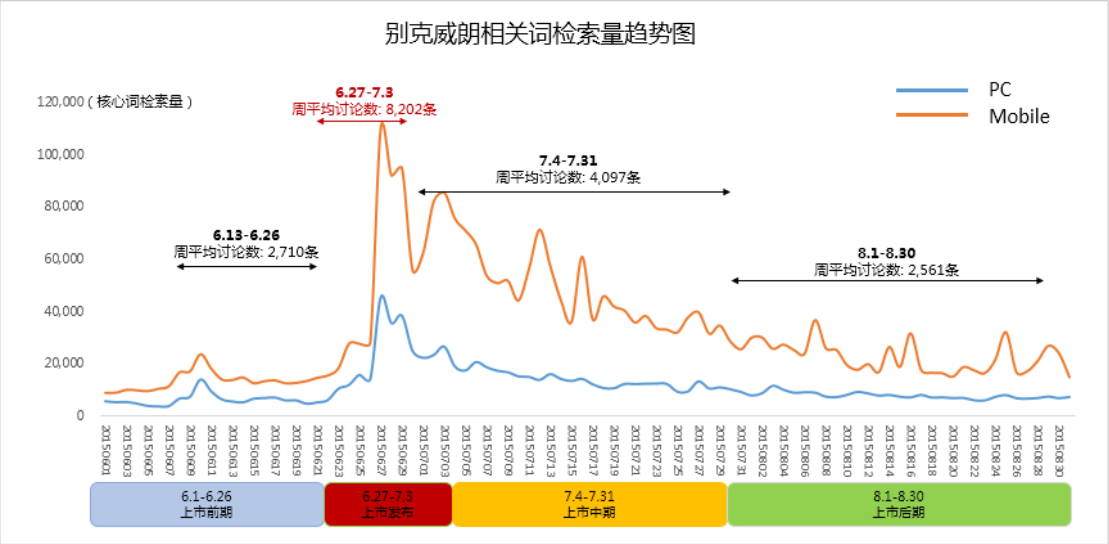
语料对象：确定的词根定为:“威朗”、“喂狼”、“verano”

语料来源：圈定了四大垂直网站（汽车之家、易车网、爱卡汽车、太平洋

汽车网）中威朗相关的论坛 BBS 的 URL，以及百度贴吧威朗吧和百度知道中与威朗相关的 Q&A 内容。这些内容数据便成为了研究的主体语料。

分析进度：大数据研究（图 9），从上市预热前期到上市 campaign 投放结束（6.1-8.28），by week 的监测分析了内容化大数据，在上市进程中给予广告主及时的反馈，在 campaign 结束后汇总各项数据，对车型的现状和未来的研发做出建议。如下为完整的数据梳理：

整个过程中，基于别克威朗搜索核心关键词包，周平均内容数的波动与核心词包检索量同步波动，即**搜索热度与社交评论热度呈现正相关**。（图 10）故这样的数据呈现有一定深化研究的意义。



（图 9）

然后，我们仔细来分析一下“破壳抓取”到的内容数据（剔除无效及明显灌水内容后，数据共计 9,638 条评论）。社会化数据标签的维度逻辑如下，整个研究分析围绕着这四块内容展开，即：

- 关注别克威朗的核心人群特征（什么样的人？）
- **关注别克威朗的核心人群需求（关心什么内容？）**
- **关心别克威朗的核心人群态度（以什么样的语气？）**
- 关注别克威朗核心的人群口碑（做出什么样的评论？）[6]

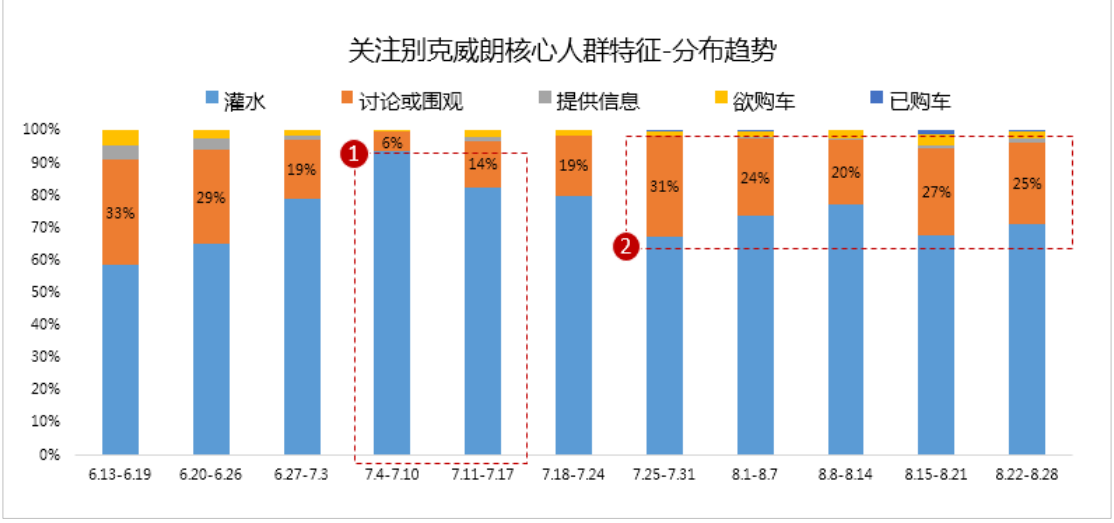
第一，**用户是什么样的人**？

根据内容评论内容的原则，将内容产生方分成这五大类人群（图 10）：

	用户特征	分类定义	评论举例
1	灌水跟帖	原帖不含有效信息	“支持发帖”或频繁出现的相同内容
2	提供信息	多为转发的新闻稿，并包括论坛内的疑问回答	“如果我的回答对您有帮助，请设为最佳答案，谢谢！”
3	*讨论围观*	*原帖包含有效信息，即围绕威朗车型各维度的讨论*	*“内饰不错”、“价格太高”等*
4	欲购车人	潜在车主，密切关注威朗或有看车或试驾行动	“就等这辆车……”、“降价两万会考虑”等
5	已购车人	多为提车反馈帖	“终于到提车了……”、“……终于入手威朗”等

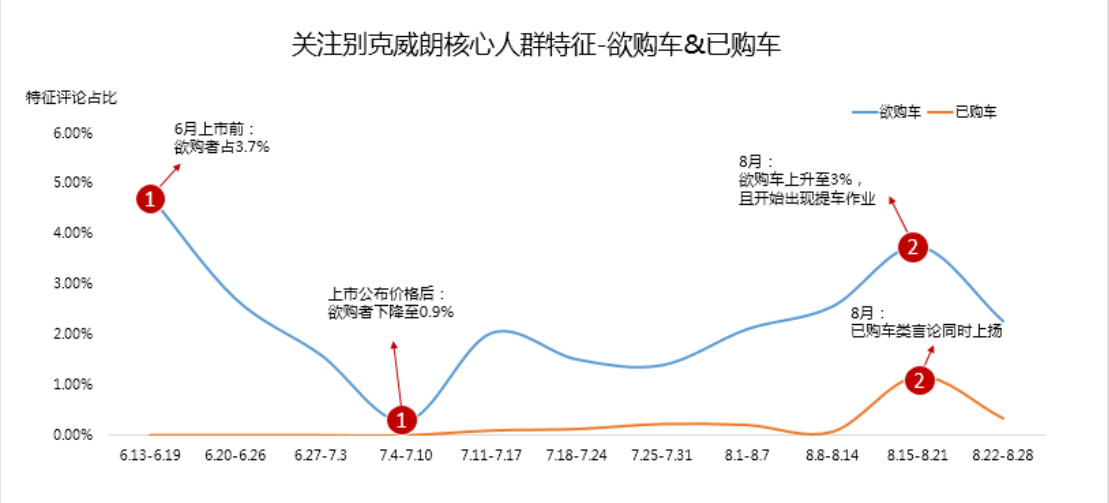
（图 10）

整理之后，从关注别克威朗核心人群特征的分布来看：**社交化体现在评论内容较为主观、直接。尤其在上市之初，人为宣传意图明显；**内容化体现：讨论围观类的评论占主要空间。在上市中期，讨论围观类言论基本占据 25%左右，中后期欲购车评论渐多，如（图 11）



（图 11）

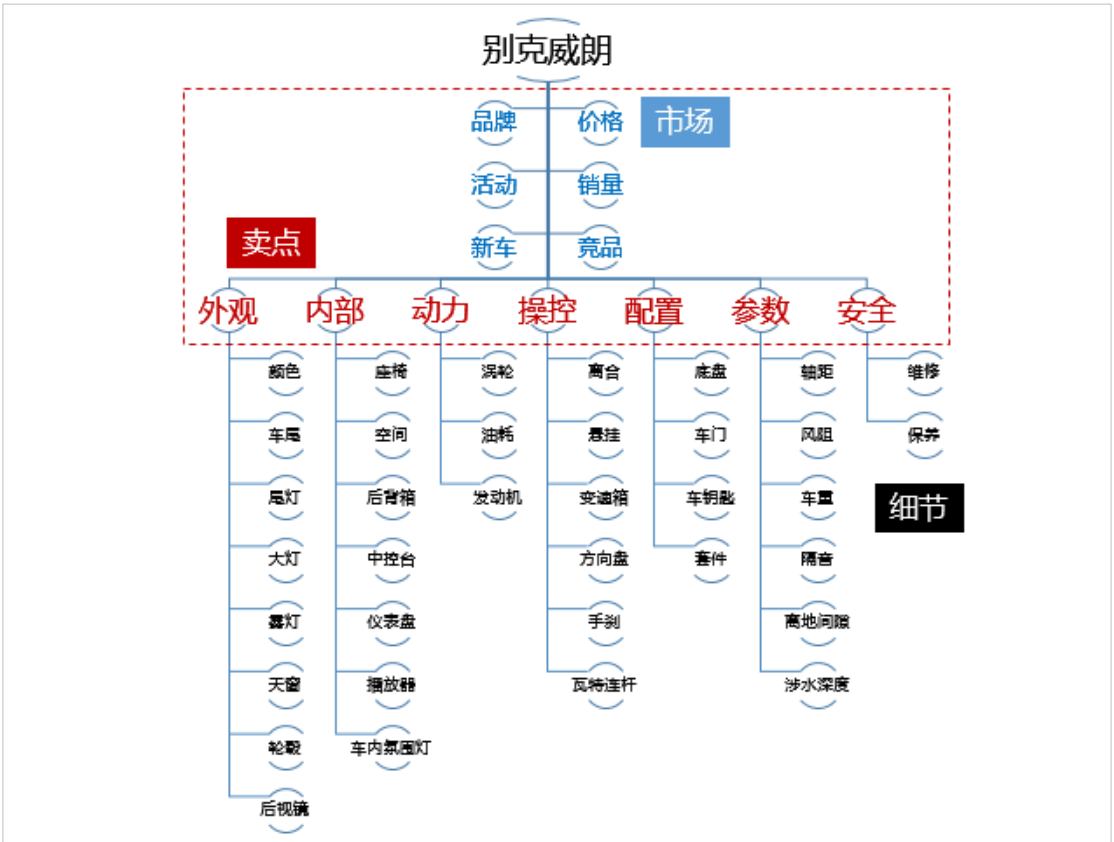
当我们将**欲购车&已购车**的评论深入研究后，发现（图 12）：上市前的预热期，类似欲购车的评论约占总评论数的 4%，当价格公布后，此类言论迅速下降，一直到上市中期才缓慢回升；**此外，在上市中期 8 月，出现了欲购者类评论和已购车类评论同时上扬的态势，并且有了若干提车作业类型的评论。**



（图 12）

第二，**用户关心什么内容？**

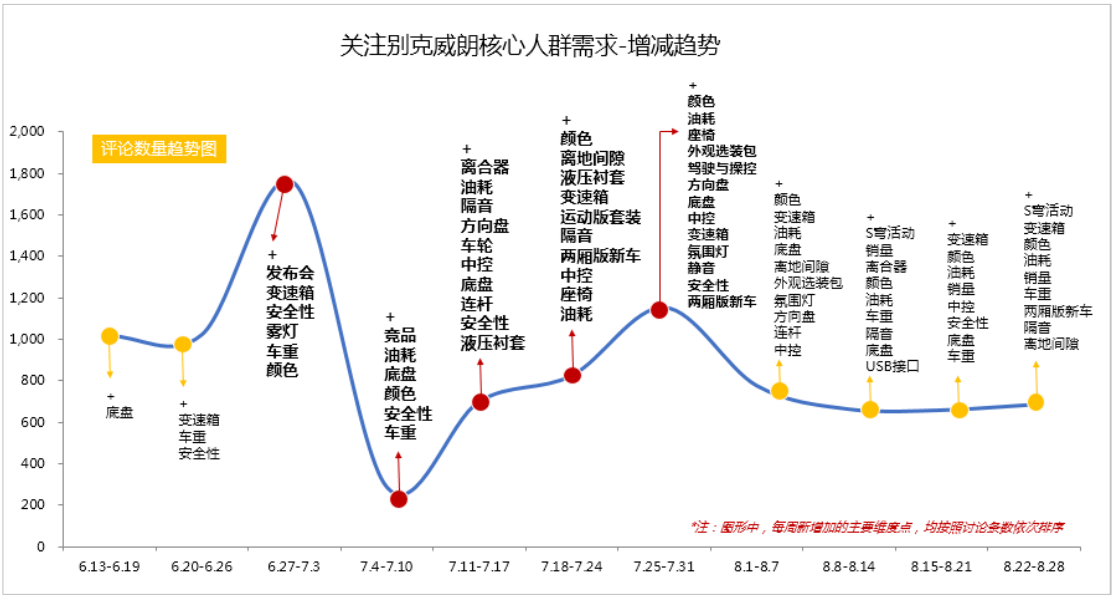
这个层级的梳理无疑是最复杂且琐碎的，在于需求层级的标签琳琅满目，那么如何在数据处理过程中先界定好标准呢？



（图 13）

如图 13，研究确定了**第一层级的市场需求（蓝字）**，大致包含但不限于**品牌、价格、活动、销量、新车、竞品相关的内容主题。**而车型的卖点构成了**第二层**

级的卖点需求（红字），大致包含但不限于**外观、内饰 、动力 、操控 、配置 、参数、安全等**；随着上市发布后，各个车型卖点在各传播渠道陆续露出，讨论的热点维度逐步增多，且讨论愈发专业且细致化，也就**细分出了第三层级的需求，即细节需求（黑字）**。

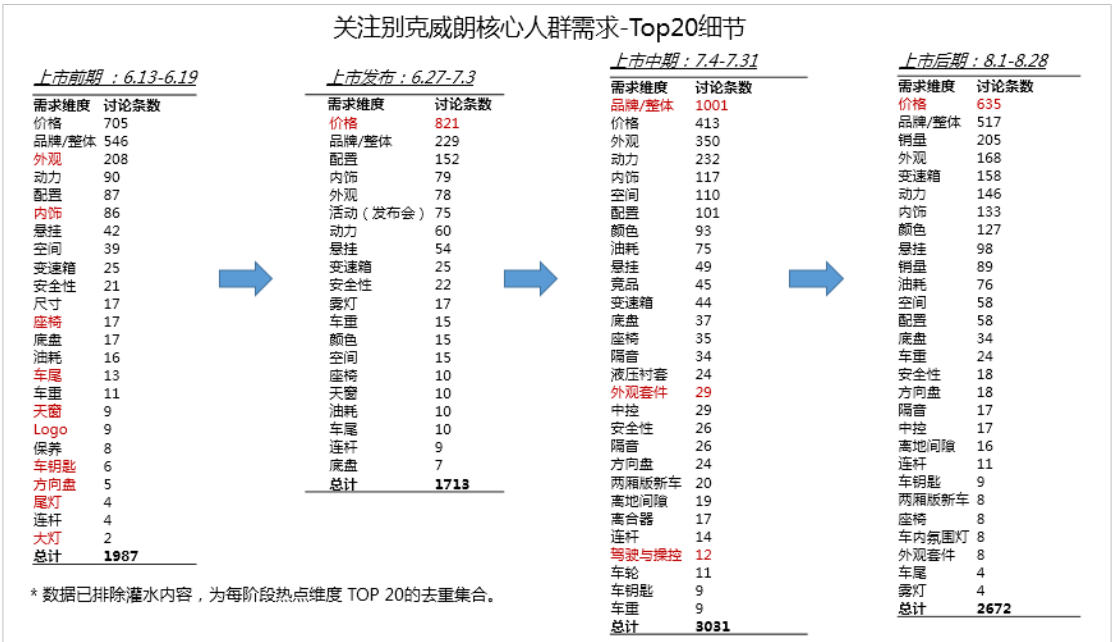


（图 14）

基于以上细致的标签，将关注点内容铺到时间轴上，发现（图 14）：

- 威朗上市预热期，讨论多为对**价格的猜测**，提供车型**基本信息与参数**，同时对**细致的外观与内饰**讨论较多；
- 威朗上市发布后的第一周，讨论的维度略有减少，内容**明显集中在价格**上；
- 威朗上市的中期，逐步从对价格的集中讨论转移到更多维度中，同时随着**看车与试驾的增多**，对车型的**综合评价、外观的改装、操纵感受**的讨论明显增多；
- 威朗上市后期，讨论又有集中于价格的趋势，多为**希望降价或优惠的呼声**，同时随着**试驾与提车的增多**，整体综合反馈也有所增加。

经过内容的数量汇总，图 15 更为直观的呈现每个不同阶段归出热点维度 Top20。



（图 15）

在这些维度中，可以特别强调一下**竞品维度**的洞察，这个数据呈现基本可以解决此前广告主提出的 3 个疑惑：

疑惑 1：别克威朗在品牌内的车型中定位是否准确？

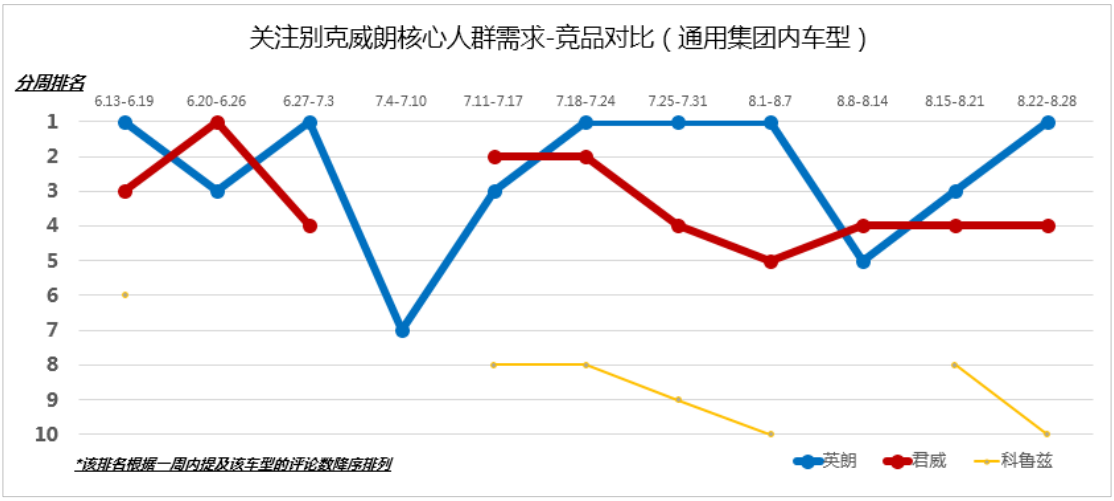
疑惑 2：别克威朗对比兄弟品牌雪佛兰的车型，是否有差异化？

疑惑 3：除了“南凌渡”和“北速腾”外，还有什么市场认知的竞品？

带着这三个问题，根据每周评论内容中提及竞品的评论数做了汇总，（图 16）按周的内容数排名，从中还区分了集团内、外竞品。

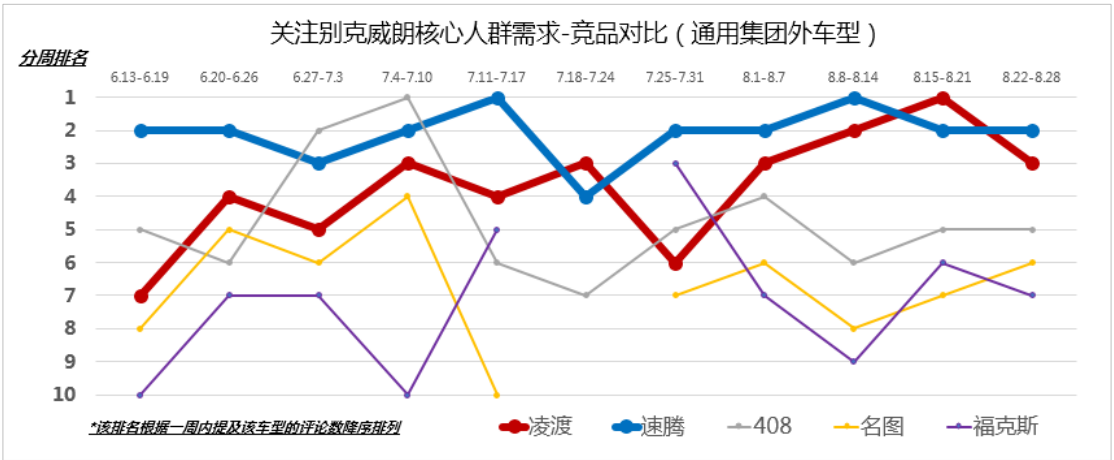
（图 16）

1. 威朗作为“大英朗小君威”的定位，别克品牌内的车型话题**显然集中于此二者**，且英朗是话题量最大的，君威其次；
2. 雪佛兰科鲁兹在研究时段中正值上市，也成为比较对象加入到集团内的竞品中，作为威朗的下游竞品被用户所讨论；
3. 还有一些集团内竞品也阶段性出现在话题中，如别克凯越，别克昂科威，雪佛兰迈锐宝。（图 17）



（图 17）

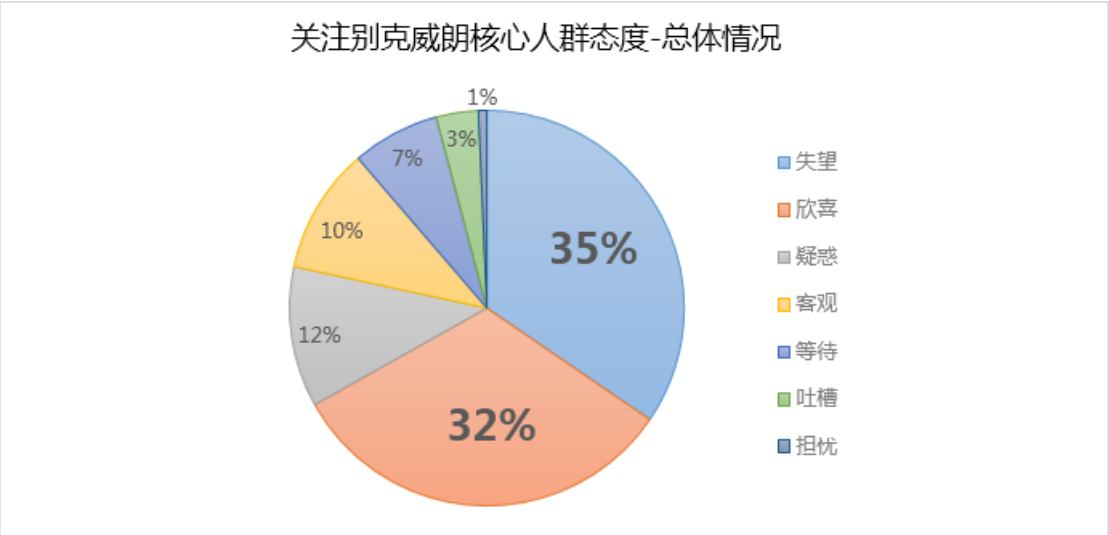
而在集团外竞品（图 18），可以看到**上汽大众的凌渡和一汽大众的速腾**，始终是核心的竞品话题，且在上市中期两者的评论量占据话题量前两位。随着威朗上市的进程，凌渡的话题量逐渐上升，于市场预期的那样成为威朗核心的对比竞品。此外标致 408、北现名图和福特福克斯也榜上有名，其中标致 408 在整个进程中的话题量比较稳定，仅次于两核心竞品。



（图 18）

第三，**用户以什么样的语气？**

这部分将所有语料的情感标签归纳为以下几种（图 19）：欣喜、客观、等待、疑惑、失望、吐槽、担忧。从核心人群的总体态度来看，喜忧参半，也有部分人表现出疑惑求解，客观中立，等待观望的态度。

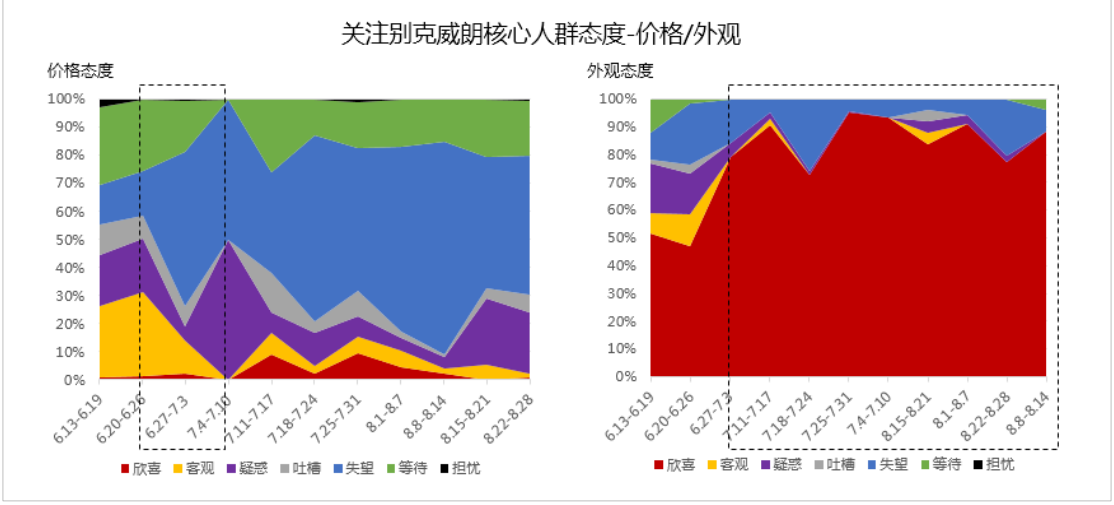


（图 19）

跟着时间轴看，在上市前期，“期待”与“疑惑”的态度居多，同时亦有不少客观提供信息者的存在；在上市发布当周，“失望”的态度激增，且“期待”锐减，同时“等待”的声音多为期待降价；在上市中期，虽然“失望”态度较多，但总体处于讨论的状态；在上市后期，随着试驾与提车反馈的增多，“欣喜”的态度有所回升，同时对威朗更为细致的提问与“疑惑”增多。这里还结合了关注点的维度细看，不同时期用户的情感倾向是什么样的。不难发现，**价格和外观是用户表达情绪比较集中的两个特征**。（图 20）

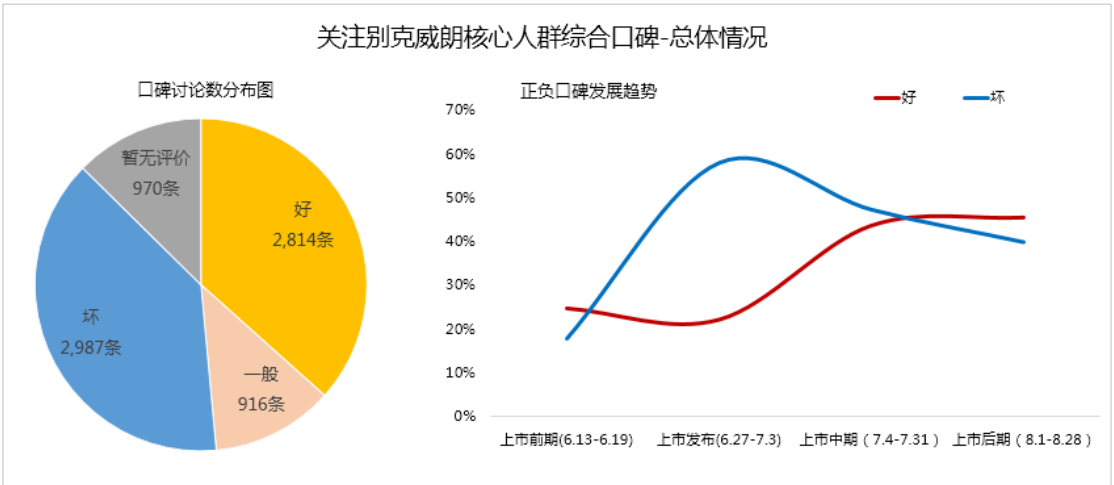
价格维度：核心人群在上市前，多持有等待/疑惑/客观的姿态关注，但**当价格发布后普遍觉得价格略高表示失望**，此后也有部分人在试乘试驾后，表示该车**信价比是在接受范围内**的；

外观维度：核心人群在威朗对外问世后，**对于外观的评价都非常积极**，可以说外观及其细节是这辆车在上市后最为显著的公众卖点。（图 21 为百分比面积图，其中数据为（每周讨论）条数占比，且已排除灌水内容）



（图 20）

而总体来看（图 21），正负口碑各执一词，评论的绝对值数量相当；而给予中评和不予评价的数量也近乎相当；从进程来看，**当价格公布后，市场反馈的确较为激烈，这样的评论延续了将近两周**。

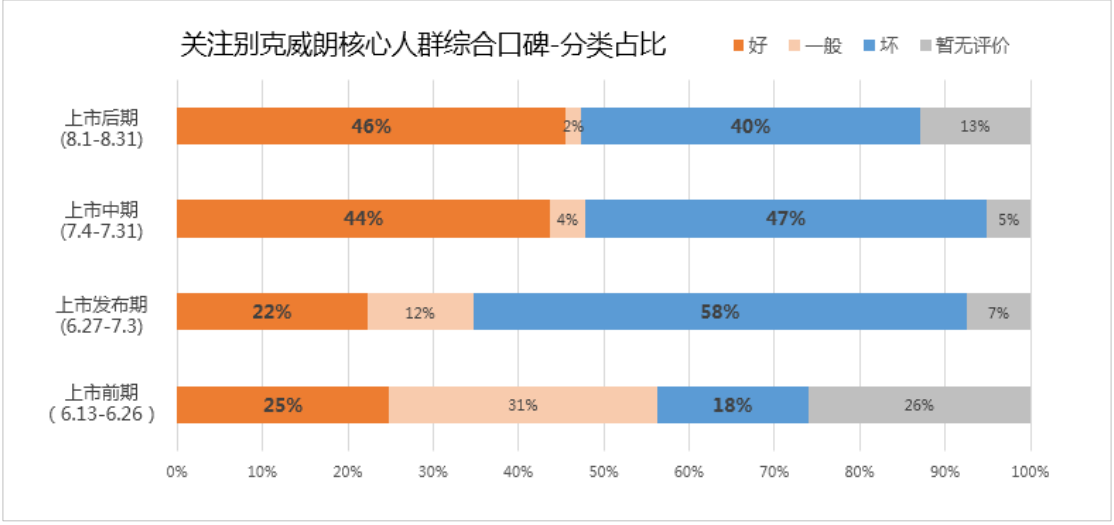


（图 21）

第四，**用户做出什么样的评论？**

这一部分标签可以被叫做综合口碑，简单来说就是没有过多的情感色彩的直观态度。研究中分为好、一般、坏、暂无评价这四种综合口碑。

如图 22，威朗上市前期 ，猜测与期待的好评和一般评论较多，观望不评价的言论也占据一部分；威朗上市后，特别是在官方价格发布后，多数网友认为定价略高，且有宣泄情绪的恶意灌水行为；随着威朗试驾与官方活动的开展，舆论在后期出现回暖，特别是在 8 月，口碑逐渐回升。



（图 22）

威朗核心人群综合口碑的媒体倾向（图 23 以下仅以此案例的数据为参考）

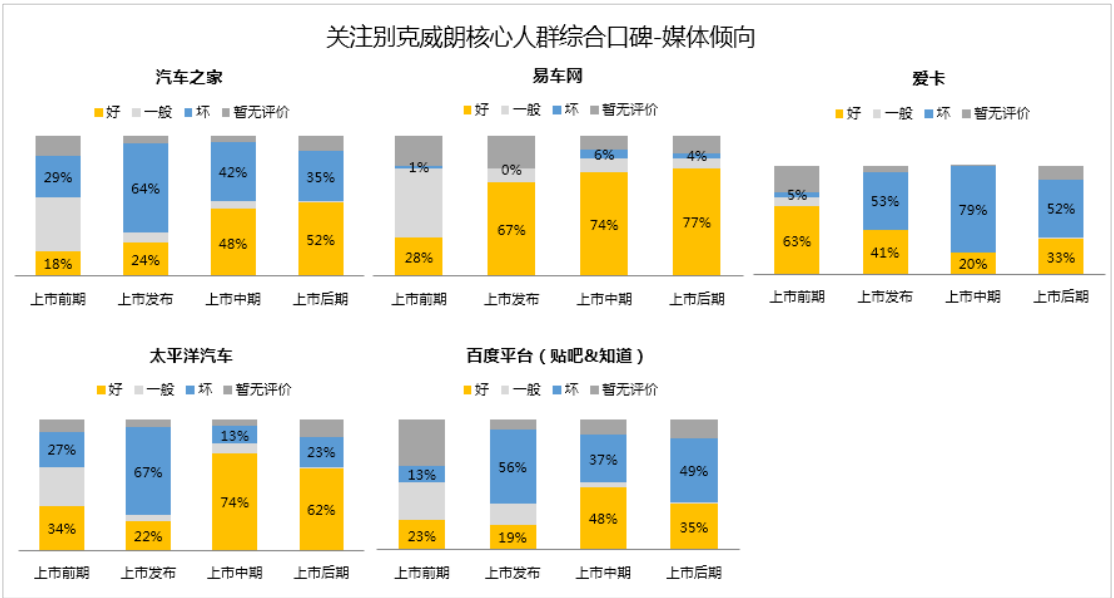
汽车之家：评论较为深度且细化的中立媒体论坛；

易车网：多为疑问解答（“暂无评价”中包含大量中性回答），且好评率较高；

爱卡论坛：在 7 月有明显涌入“车黑”现象，大量言论并无具体内容指向，如“威朗垃圾”、“反正我不买”等；

太平洋论坛：上市后，好评率大幅增加；

百度知道/贴吧：较为中立的社交平台，同时知道平台中也含较多中立的回答。



（图 23）

综上刚刚展开的四大点，研究做了如下归纳：

人群特征：上市过程讨论/围观类言论基本占据 25%；上市中后期欲购车类评论渐多；评论中能甄别出较多的灌水内容。

人群需求：热点需求为价格，品牌/整体、外观、悬挂等。随着发布后讨论的 FBI 越发细化；前期对于价格猜测，发布期对于配置的各执一词直到中后期渐缓；集团内竞品为英朗/君威/科鲁兹，集团外竞品为凌渡/速腾/408 等。

人群态度：比重最高的两群人可谓喜忧参半，也有部分人表现出疑惑求解，客观中立，等待观望的态度；在价格维度始终的激烈反馈的众矢之地，而在外观一边倒呈现好评，配置方便主要对于独立悬挂的问题略表遗憾。

人群口碑：总体来看，正负口碑各执一词，评论的绝对值数量相当；而给予中评和不予评价的数量也近乎相当；舆论倾向方面，汽车之家的确较为专业，深度且中立，百度社交平台在言论方面也较为客观。

基于上述的内容化数据提炼，在投放全过程中对关键词投放和精准展示类投放做出了实时的策略优化调整（图 24），包括但不限于竞价调整，拓量增加维度，创意更新等。并且，在内容大数据分析中，很好的体现出了广告主要推广的卖点的确被用户所广为讨论。于是在投放维持阶段，广告主顺水推舟与百度地图做了更炫的创新合作：“天生爱跑”踩点任务。将内容大数据中广为讨论的卖点对应到地图上的 15 种不同的场所地点，与用户进行深度产品力沟通。

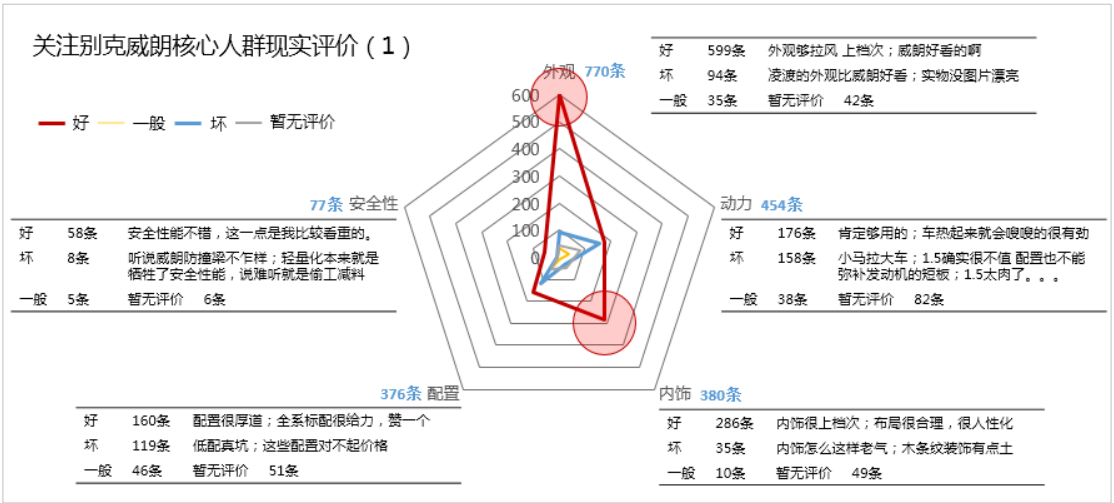
		预热期	上市发布期	上市中期	上市后期
百度产品线投放	关键词触发类	1. 价格类词：↑、✪ 2. 外观类词：↑、✪ 3. 母品牌词：↑、↔	1. 15 FBI类词：↑、↔、✪ 2. 价格类词：✪ 3. 市场竞品词：↑、↔	1. 参数配置类词：↔、✪ 2. 外观内饰类词：↑、✪ 3. 讨论竞品词：↓、↔、✪ 4. 试驾购买类词：↑	1. 产品+垂类词：✪ 2. 搜索类竞品词：↔、✪ 3. 竞品比较类词：✪ 4. 试驾购买类词：↑、↔
	精准展示类	1. 关键词定向：↑、↔、✪ 2. 兴趣点定向：↔	1. 关键词定向：↑、↔ 2. 竞品词定向：↑、↔ 3. 兴趣点定向：↓	1. 配置类词：↔、✪ 2. 外观内饰类词：↑、✪ 3. 讨论竞品词：↓、↔、✪ 4. 试驾类词：↑、↔、✪	1. 产品+垂类词：✪ 2. 竞品比较类词：✪ 3. 试驾类词：↑、↔、✪
	原生信息类	1. 关键词定向：↑、↔、✪ 2. 兴趣点定向：↔、✪	1. 关键词定向：↑、↔、✪ 2. 竞品词定向：↑、↔、✪ 3. 兴趣点定向：↓、↔、✪	1. 配置类词：↔、✪ 2. 外观内饰类词：↑、✪ 3. 竞品词：↑、↔、✪ 4. 试驾购买类词：↑、✪	1. 价格类词：↑、✪ 2. 竞品比较类词：↑、✪ 3. 试驾购买类词：↑、✪

* 标识意义：↑提价，↓降价，↔关键词拓量/增加维度，✪更新创意

（图 24）

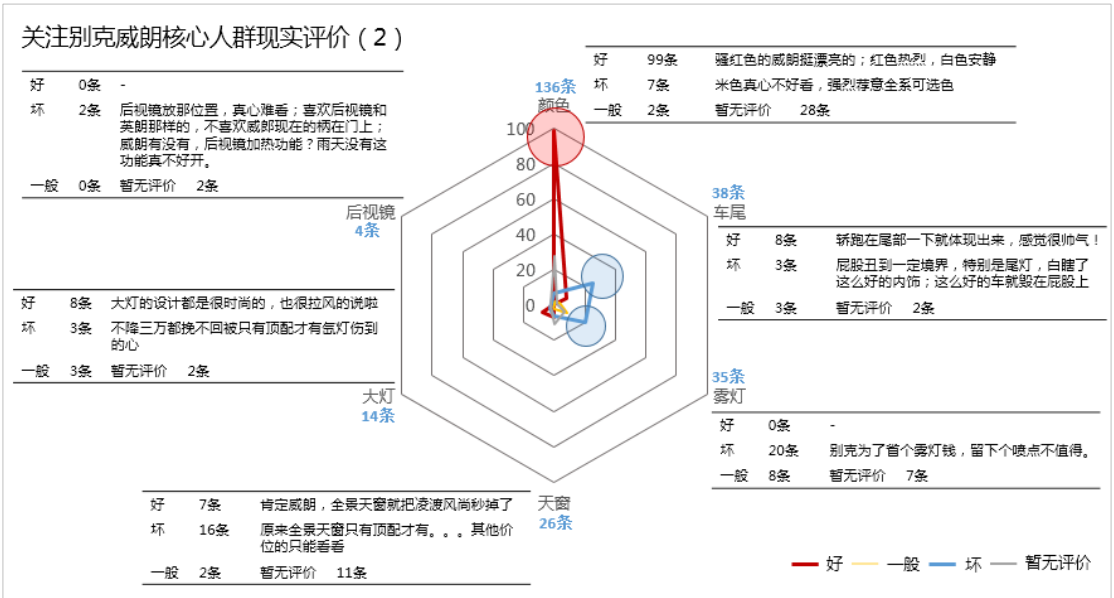
上面这部分是思路是偏数据分析和策略承接的，而开篇提到了的大数据需要有预测性、相关性和多维性。那么这些数据在投放结束后可以给未来该车型的市场策略，新品研发什么启示呢？先看看关注别克威朗核心人群现实评价，选取几个维度做出雷达图可以直观的看下内容评分。

维度 1：外观、动力、内饰、配置、安全五大属性（图 25），不难发现外观和内饰是用户心中威朗的最大看点，作为别克旗下的动能车型，威朗的配置和动力在未来仍有提升的空间。



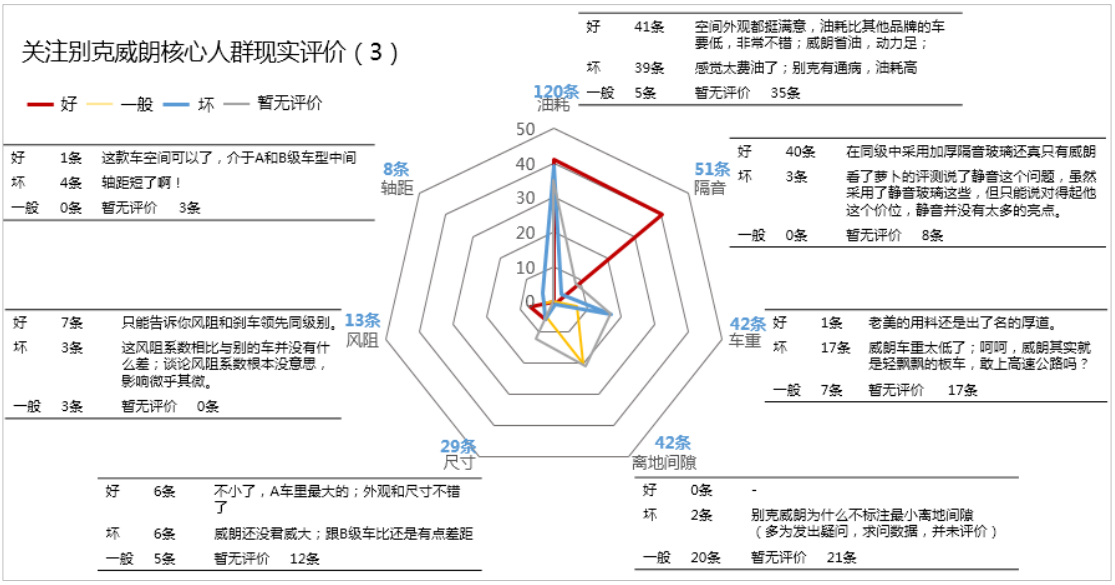
（图 25）

维度 2：基于外观的细节列举，包括后视镜、大灯、雾灯、天窗、车尾、颜色的比较（图 26）。



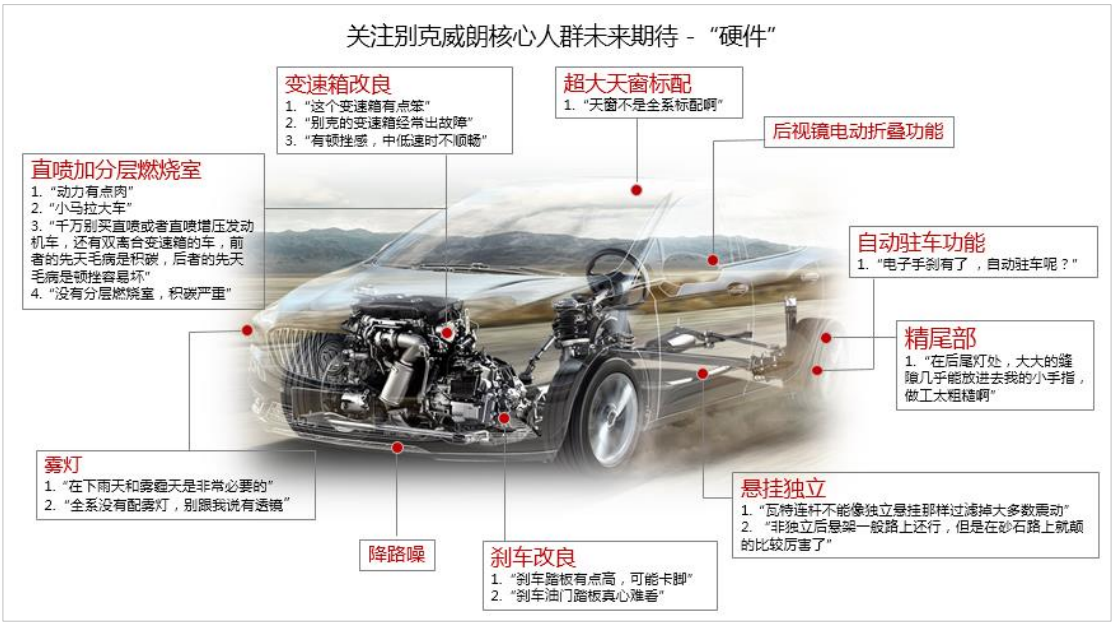
（图 26）

维度 3：基于参数的细节列举，包括油耗、隔音、车重、离地间隙、尺寸、风阻、轴距的比较（图 27）



（图 27）

当然实际研究还做了很多维度的对比分析，报告中先不赘述。那么针对现实的评价，也就可以得出用户眼中理想的下一代威朗，可以在哪些方面改进。到底是配置、外观、内饰？还是价格呢？由此构制除了核心人群对于威朗未来在“硬件”或是“软件”上的期待。（图 28、29）



（图 28）

（图 29）

【小结】整个案例为广告主提供的服务不止于量化的数据分析，而是进一步深挖真正的核心人群背后的内容和情感偏向。对于广告主来说，在投放过程中高效地得到很多细腻的用户关注点需求，且实时监控社会化情绪信息，为广告主提供更为及时的投放策略变化建议。当投放结束后汇总信息和反馈，为未来广告主的市场策略，产品研发提供前瞻性的用户反馈信息。

3.2 汽车投放前案例：上汽大众斯柯达新晶锐“快乐营销”项目

再来看一个搜索社会化大数据应用于车型投放售前的实例。

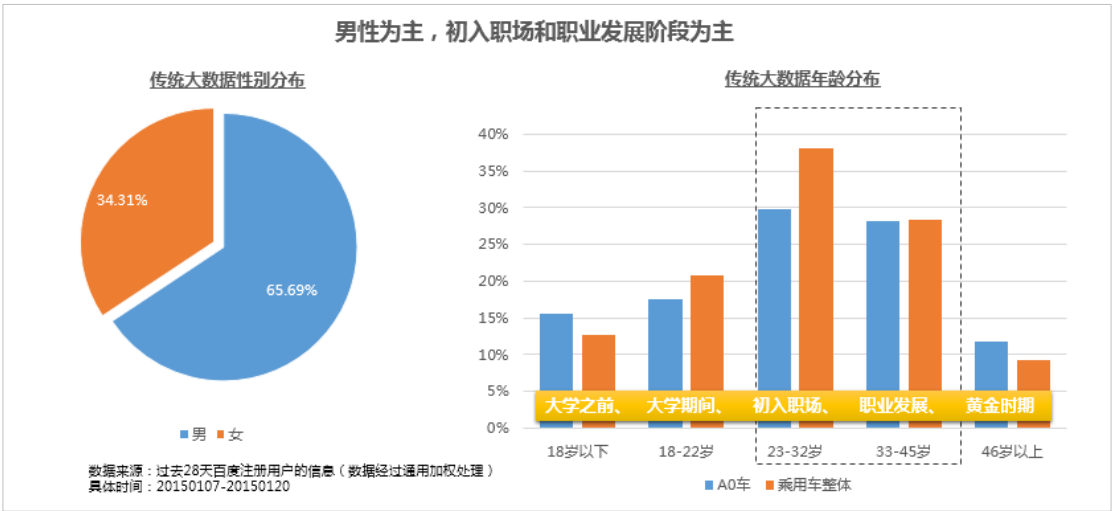
案例二：上汽大众斯柯达新晶锐新车上市

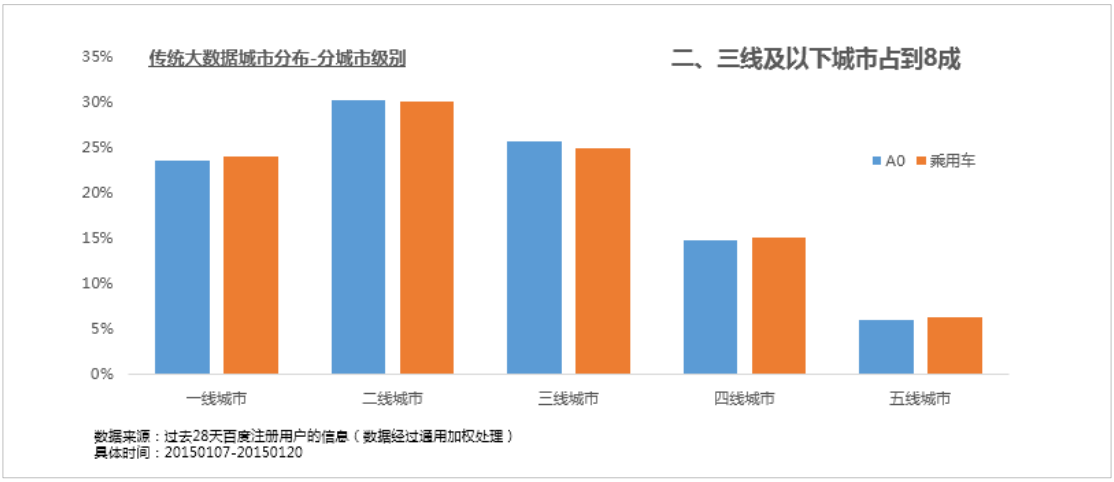
斯柯达新晶锐相关背景信息：

- 品牌传播的挑战：绕开配置和性价比，**以态度和价值观打动消费者**；理解当下年轻消费者，以朋友的身份平等沟通。
- 品牌的大数据方案：从消费者的日常言论中来**提炼情感共鸣点**；从大数据中全方位洞察生活情境。

传统的大数据无论是在自然属性或是市场分布（图 30 和 31），都无法给出 A0 级市场上明显的人群差异化特征，仅仅只是得出一群男性、初入职场和处于职场发展，以二、三线城市为主的 TA，像这样的人群画像几乎入门级的车型和一些国产

车都比较雷同，所以这样的人群画像**尚缺灵魂**。



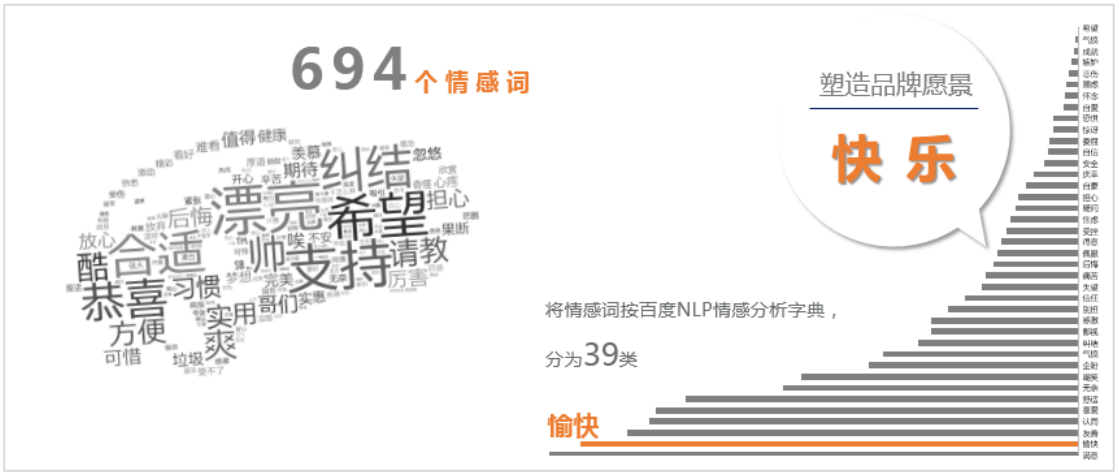


（图 30、31）

那么基于**内容聆听**出的大数据研究有何不同呢？

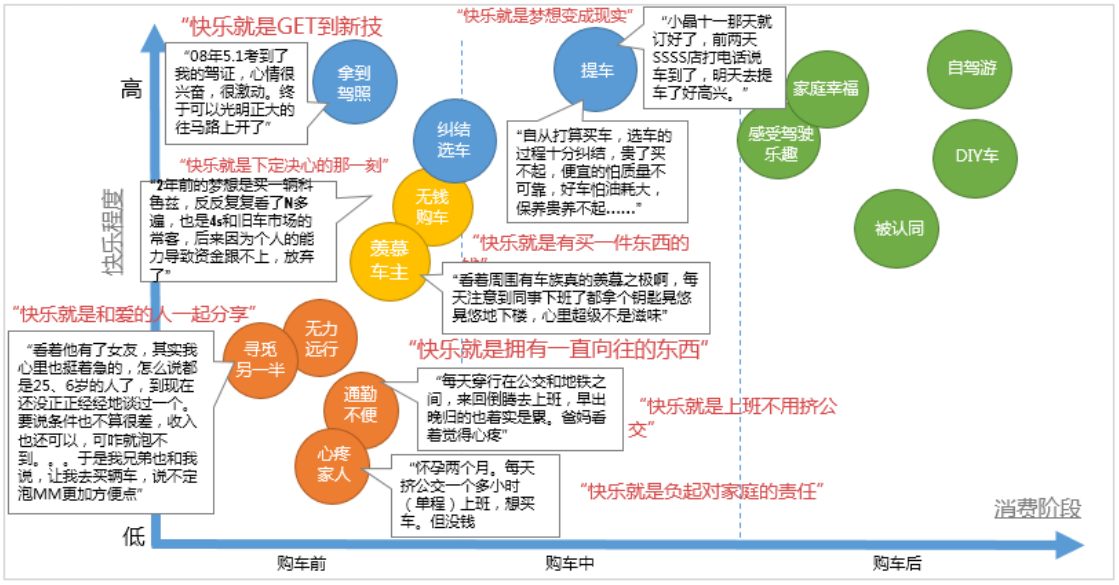
第一，从百度内容大数据到“快乐”，寻找情感共鸣点，塑造品牌愿景

基于传统的大数据分析结果并没有带来更多差异化的信息，所以想到了从垂直网站的近亿个页面中提取和 A0 级车用户和潜在用户公开讨论的近 200 万条的内容，利用百度的 NLP（Natural Language Process）自然语言情感分析，将含有情感关键词的讨论内容提取并进行聚类分析，如下图 32。



（图 32）

这当中比较具有抽象意义的情感词聚类有 694 个，继续进行排序和二次聚类便会发现“快乐”是具有典型特征的用户语境。所以基于这样的大数据提炼，我们与广告主共同确认了“快乐”这个品牌愿景。这些各种各样的快乐相关的语料总结成一句话就是：“有了车，就一下子有了实现各种快乐的可能性”（图 33）



（图 33）

第二，内容大数据为广告主找到了人群的情感特征，便有了承接性的推广方案

1. 在关键词 SEM 广告中，加入的多套创意轮替中，**均体现出了“一切为了快乐”**的 TA 主题，购买词为 **小型车行业词词包。（图 34）**

   SEM 文案标题 1：“小型车首选，全新超大空间斯柯达晶锐 一切为了快乐”

   SEM 文案标题 2：“小型车 2015 排行，全新斯柯达晶锐 一切为了快乐”

（图 34）

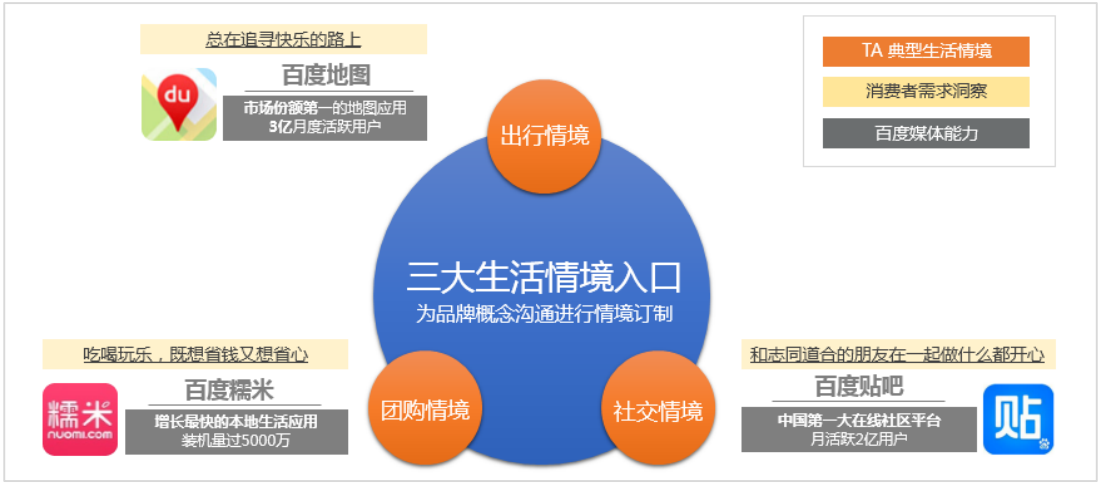
2. 在精准展示类广告物料中使用多种色彩的物料画面，提及快乐上市的信息；Eg. DSP 广告物料上加入各种与快乐相关文字元素的 slogan 拼合成的空气动力效果，配以不同颜色，突出“只为快乐”的主题。**（图 34）**
3. 在知道平台，针对各种 A0 级行业的问题页，嵌入了知道专题页，互联网快乐青年报告，从数据到人群，再到实实在在的车型卖点信息，以专题的形式原生的呈现在问题与问题中，如下**图 35**：

（图 35）

4. 在创新平台的合作中，选取生活中的出行场景、团购场景和社交场景分别配以百度的地图、糯米和贴吧，均突出 TA 典型的生活形态，三个平台分别与快乐接轨。（图 36）

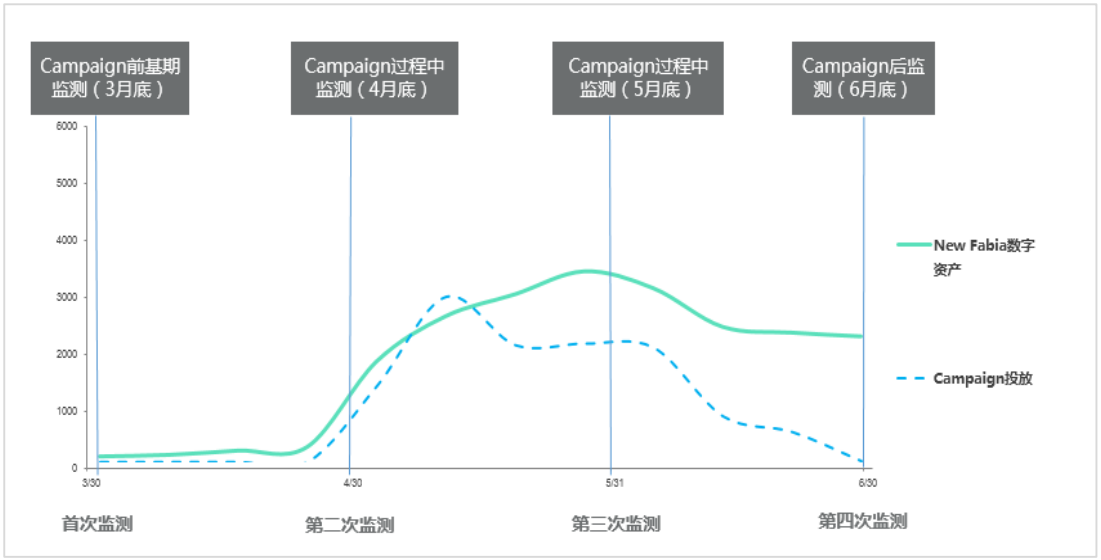
   ➢ 百度地图-总在追寻快乐的路上；

- 百度糯米-吃喝玩乐既想省钱又想省心；
- 百度贴吧-和志同道合的朋友在一起做什么都开心



（图 36）

当这波活动投放结束后，研究发现 New Fabia 在百度的数字资产的趋势稳定且有明显提高，夯实了该车型的经典底蕴，如下图 37：



（图 37）

【小结】整个 New Fabia 案例一气呵成，既有内容大数据分析论证的新颖人群画像，又有产品承接解决方案，还有创新合作作为亮点。无论是常规产品线投放，还是软性内容专题，抑或是地图、O2O、社交工具平台的契合都无一例外的表达了快乐的人群主题，而这个主题也正是研究使用大数据为客户找到的最初的 idea。

第四章：内容大数据的应用拓展

在第三章中，已经可以窥见社交化数据的特征：灵活性和可塑性.以一个新颖的角度为广告主做大数据分析，所呈现的变化和洞察是非常丰富的。那么过程中，发现两个问题：一，数据采集过程需要程序技术支持，这样的技术百度现在是可以做到的，被叫做“网页结构化数据的提取”。二，在数据识别过程中，机器识别量和人工识别量的比重直接确定了该项目的效率。针对这两个问题，研究在第四章拓展社交化大数据的未来。

4.1. 内容大数据的工具化：汽车舆情内容的数据监测

一．数据采集

1.. 语料对象定义：

（A）汽车车型名称词包

（B）汽车车型昵称词包

2. 语料来源 URL 定义:

（A）五大垂直网站的车型相关论坛 URL

（B）百度贴吧相关吧百度知道相关内容

3. 时间区间定义。无论从数据量或时效性来说，按周获取为宜。

二．数据识别

1. 标签映射规则定义，如之前第二章中提到的，汽车基础行业 605 个不同“关注点+修辞”映射的结果

2. 汽车行业社会化语料词典，包括：

（A）汽车行业车型名称及对应的昵称，（将用于数据采集中的对象词包定义）

（B）汽车行业专业修辞及“行业黑话”（将扩充数据识别中“关注点+修辞”）

三．数据结果

1. 标签映射内容监测

对于内容数据中的关注点提取，现在已经可以做到比较好，程序会自动输出关于不同内容渠道提取后的关注点维度分布，如图。

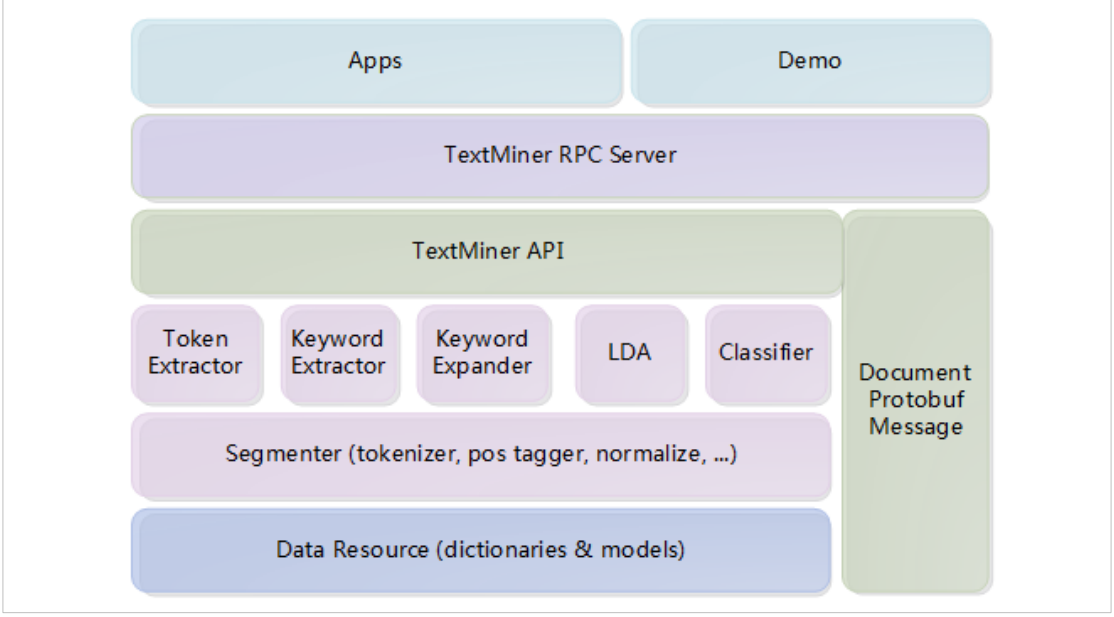
2. 情感语境内容研究

这部分研究对于内容分析的要求较高，如斯柯达新晶锐案例中所讲到的。目前机器能做到的是提取修辞词汇进行聚类，看哪些情感修辞语所占的比重较高，来确定对象内容的主基调。未来这部分的社会化聆听的发展有赖于自然语言学习的提升或人工智能的发展，像度秘这样偏机器学习、人工智能的产品的进步，相信其实对于情感语汇的学习已经可以落地做到，只不过这样的技术需要被应用的产品范围需要扩大，需要被推及。

4. 2. 内容大数据的产品化：汽车广告品效的产品优化

第一，**语料规则积累**。内容大数据的产品化，对于自然语言的机器学习有一定的促进作用，并提供应用环境。现在简单的程序编写可以大致映射内容的观点标签、情感标签，但是这也只是随着语料库的人工扩充而变得更丰富，更智能。这个还依赖于未来人工智能的技术，汇聚成海量大数据挖掘系统。[6]

真正需要高效、精准判断数据的情感，“腾讯的一种常见的方法是文本分类，由于对标注语料库的依赖，类别规模一般不会太大，粒度较粗。还有一种方法就是文本聚类，挖掘语义主题标签，更细粒度的理解文本意思，隐含语义分析技术逐渐发展成为常用的解决方案。这些不同维度的文本分析模块，包括词袋、关键词提取、关键词扩展、文本分类和 Peacock 模型等。”（图 38）



（图 38）

既然百度大脑现在自然语言、语音技术、图像技术、用户画像、机器学习和 AI 上都已经有所发展，相信对于数据的情感语义判断完全可以被提高。

第二，**内容环境的优化**。当大数据的内容语义积累的越来越多，越来越深之时时，自然而然就形成了这个行业语料库，之后的事就是将智能化应用到广告产品的功能中去，无论是对于**广告环境的判断、原生内容主题的判断、还是推广内容的推荐**都是可以拓展应用到商业产品功能中去的。

第三，**广告效能提升。**用户获取信息另一个重要的渠道就是 Feeds，正因为 Feeds 的资讯是个性化，语义理解就能发挥其中的作用了。如果能通过机器学习给内容或文档打上各种各样的标签，例如本文第三章案例中提到的主题标签、情感标签和口碑标签。**这些标签可以从不同角度描述一个对象内容，以满足不同应用需求，并于不同的搜索 query 相关联，可以形成了关注点标签图谱。**[7]这种关注点标签能更好地描述用户与对象内容之间的关系，因为它能同时对用户和内容进行推理和计算。如此若可以应用在广告产品上，可以提升广告相关性，为广告找到相似的用户，这样也同时可以提高广告的点击率。

例：腾讯丰富的产品线拥有中国互联网公司最多的用户，有着海量、丰富的用户关系和行为数据，如 QQ 好友关系，QQ 群关系，电商浏览、交易，新闻浏览，查询 Query，UGC 内容（如微博、说说等），移动 App 安装，微信公众号文章阅读和广告点击行为等。通过用户行为数据的挖掘可以帮助我们更好的了解用户，以推送精准的广告。

而这些数据都可以形式化为**用户-物品矩阵**，如用户-用户、QQ-QQ 群，用户-应用（Apps），用户-搜索词（或搜索 Session），用户-URLs 等。**腾讯利用 Peacock 系统对上述用户-物品做矩阵分解，从不同数据来源，多视角理解用户兴趣，进而挖掘相似用户，提供给广告主丰富的定向策略，如用户商业兴趣定向、关键词定向和 Look-Alike 定向等。**同时，获取到的用户特征，也可以作为广告 CTR、CVR 预估系统的重要特征。[8]

对于百度商业产品来说，精准展示类 DSP 广告、百度的原生信息流广告、凤巢的关键词相关性匹配都可以使用基于用户行为数据的理解，从中归纳用户的兴趣和需求，提供广告精准定向技术，为广告主做到更高效的推广。

第五章：结论

汽车行业传统搜索大数据的三个主要特征：趋势性、相关性、原生性。而搜索引擎最本质的强大能力是抓取迎合用户的内容进行排序，提供有价值的流量。针对搜索核心流量分布的拆分，延展到对于内容大数据的结构化抓取，将这部分“活大数据”进行语料解析和深挖，找到用户的情感和诉求，这对于汽车搜索营销乃至汽车全媒体营销都具有崭新的意义。

诚然，内容大数据的创新应用，可以服务于汽车营销投放的全程。将内容大数据投入实践，就可为营销者迅速收集市场信息，做出新颖的用户洞察，丰富的舆情监控，高效的策略优化，开放的发展引导。

不仅如此，当海量内容数据加以百度的机器学习技术，未来可巧妙地融入行业舆情数据工具。而在各种广告商业产品上，内容数据的计算识别更可以为汽车广告与用户的相关性提升、广告点击率提升和用户体验做出巨大贡献。

参考文献与注释

[1] 《百度 2014 年汽车行业峰会：大数据驱动汽车营销》大数据部分, 2014.6

[2] 《百度 2014 年汽车行业峰会：大数据驱动汽车营销》司南相关性数据部分, 2014.6

[3] 《百度 2015 年汽车行业峰会：无线畅想 双擎互动》大数据部分, 2015.6

[4]《计算广告 互联网商业变现的市场与技术》刘鹏、王超著，人民邮电出版社; 第 1 版 (2015 年 9 月 1 日)；

[5]《社交的本质》 兰迪•扎克伯格 著 中信出版社; 第 1 版 (2016 年 5 月 1 日)；

[6]《用户网络行为画像:大数据中的用户网络行为画像分析与内容推荐应用》 牛温佳, 刘吉强, 石川等著, 电子工业出版社; 第 1 版 (2016 年 3 月 1 日)；

[7]《大数据智能:互联网时代的机器学习和自然语言处理技术》 刘知远著， 电子工业出版社; 第 1 版 (2016 年 1 月 1 日)；

[8] 《让机器搞懂 100 万种隐含语义，腾讯 Peacock 大规模主题模型首次全揭秘》2016.2