DTCC
2017第八届中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2017

商业语义原生广告核心技术

李东军
Deepleaper

个人简介

- 北京跃盟科技CTO
- 曾任百度搜索时效性排序团队负责人
- 8年前沿技术经验，在搜索、自然语言处理、机器学习、大规模数据处理方面有深入研究

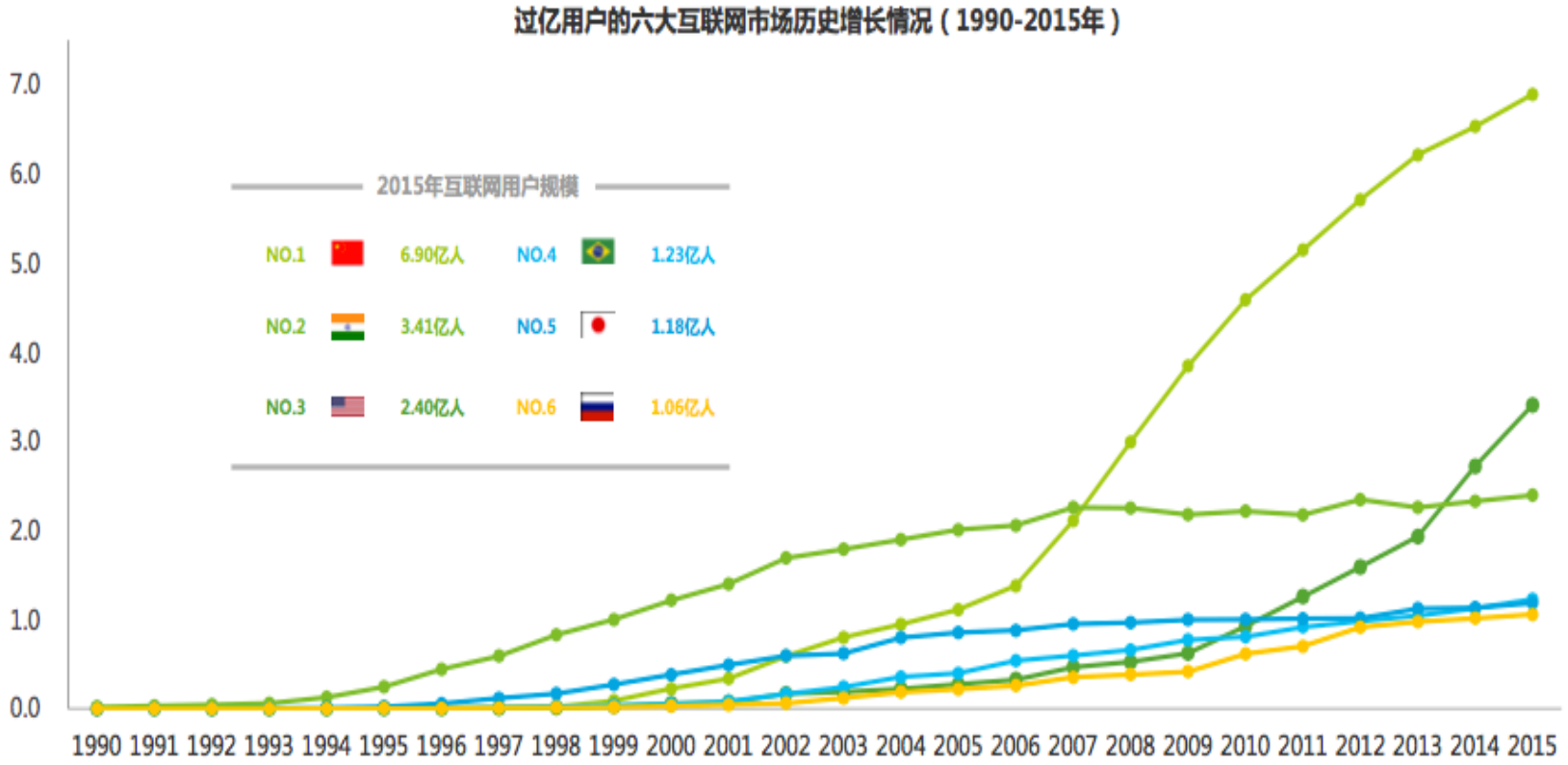
商业语义原生广告介绍

商业语义原生广告核心技术

商业语义原生广告介绍

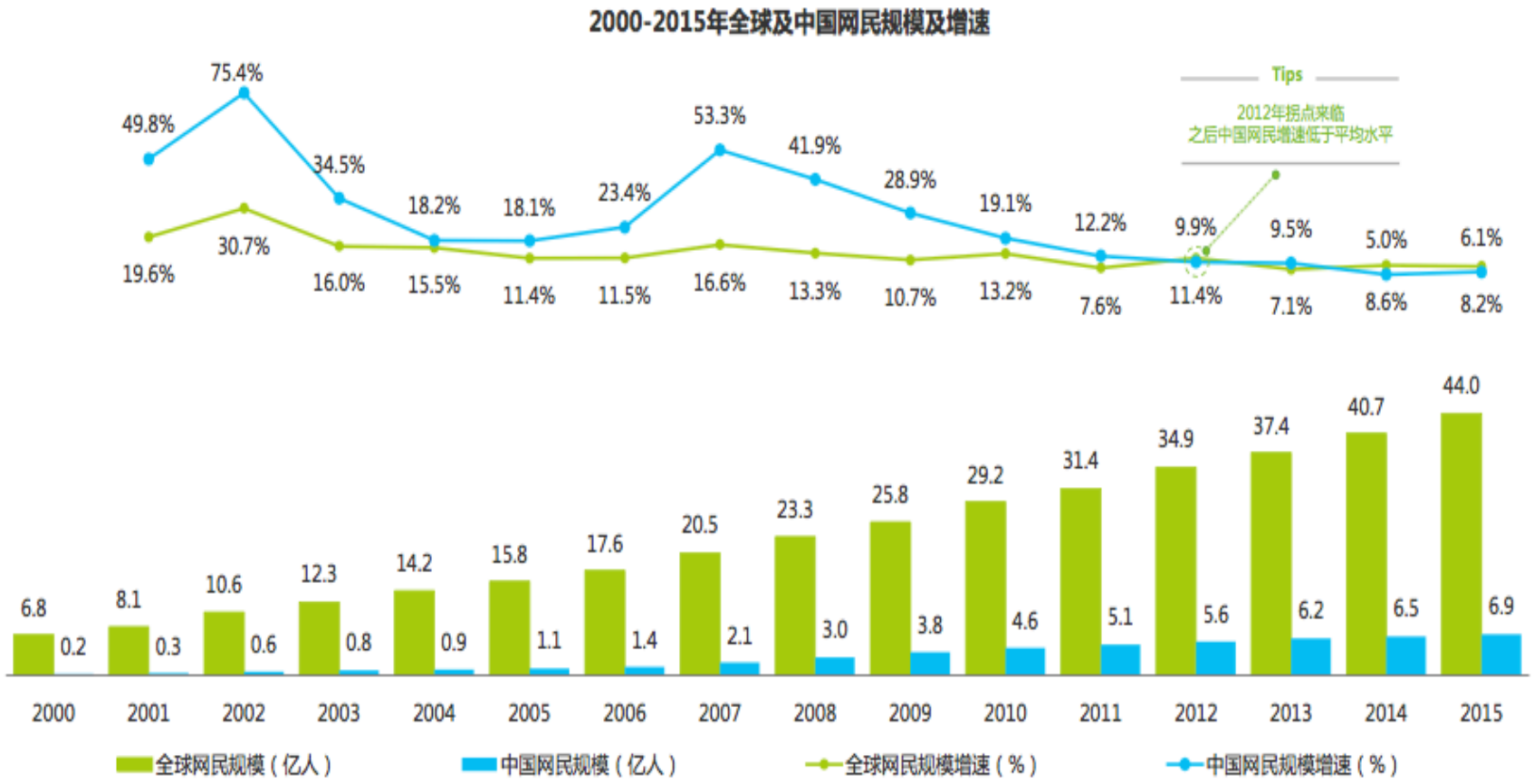
商业语义原生广告核心技术

中国是全球最大的互联网用户市场



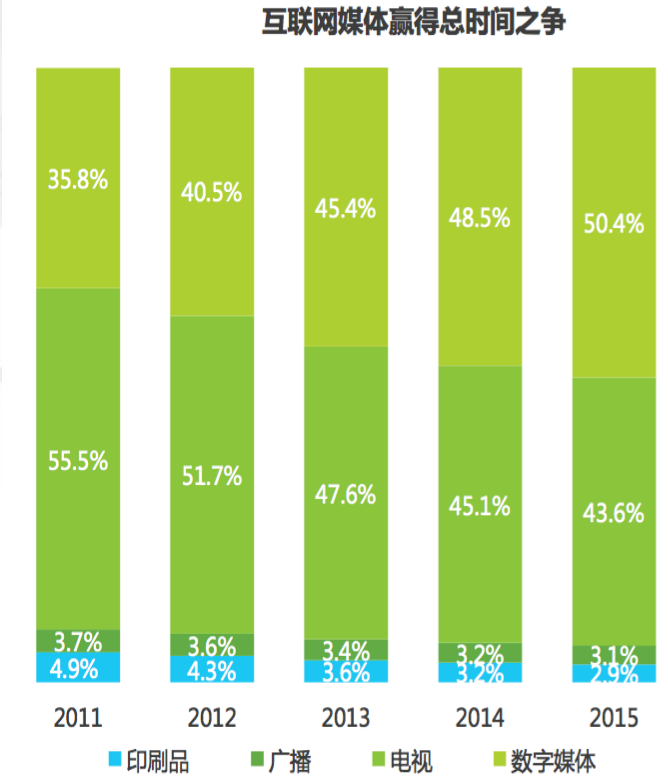
艾瑞

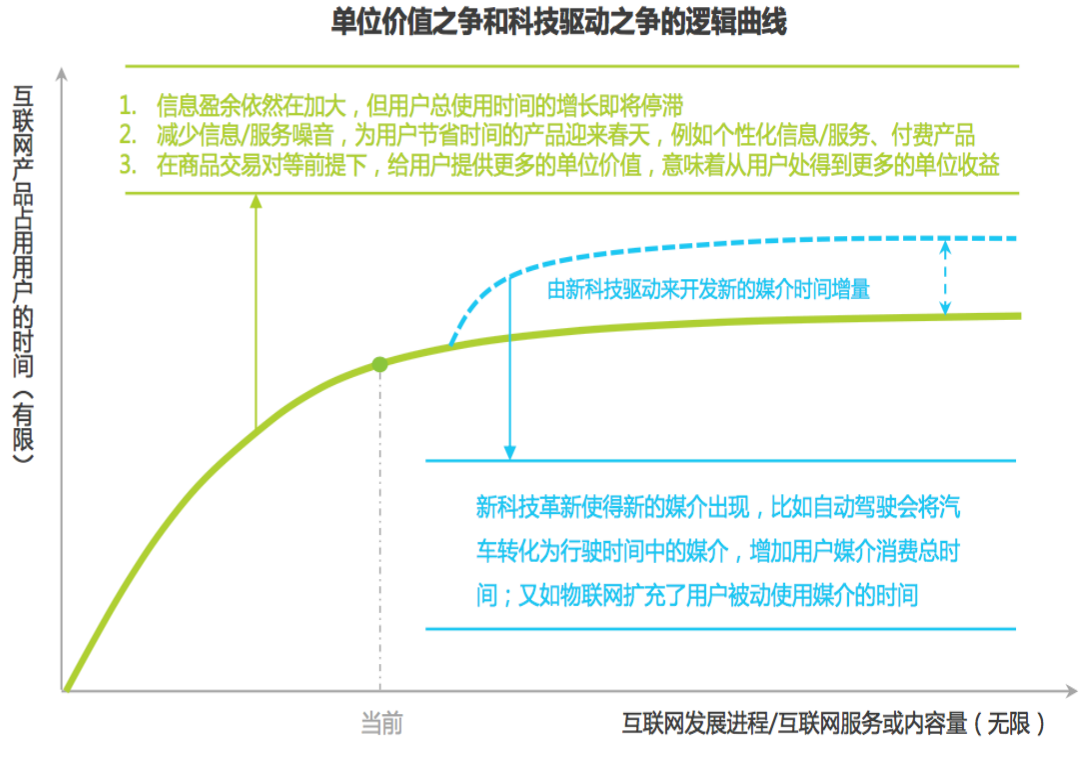
但是用户增长开始趋缓，移动互联网进入下半场



艾瑞

存量改革，由粗放到精细化运营，由总时间之争到单位价值提升



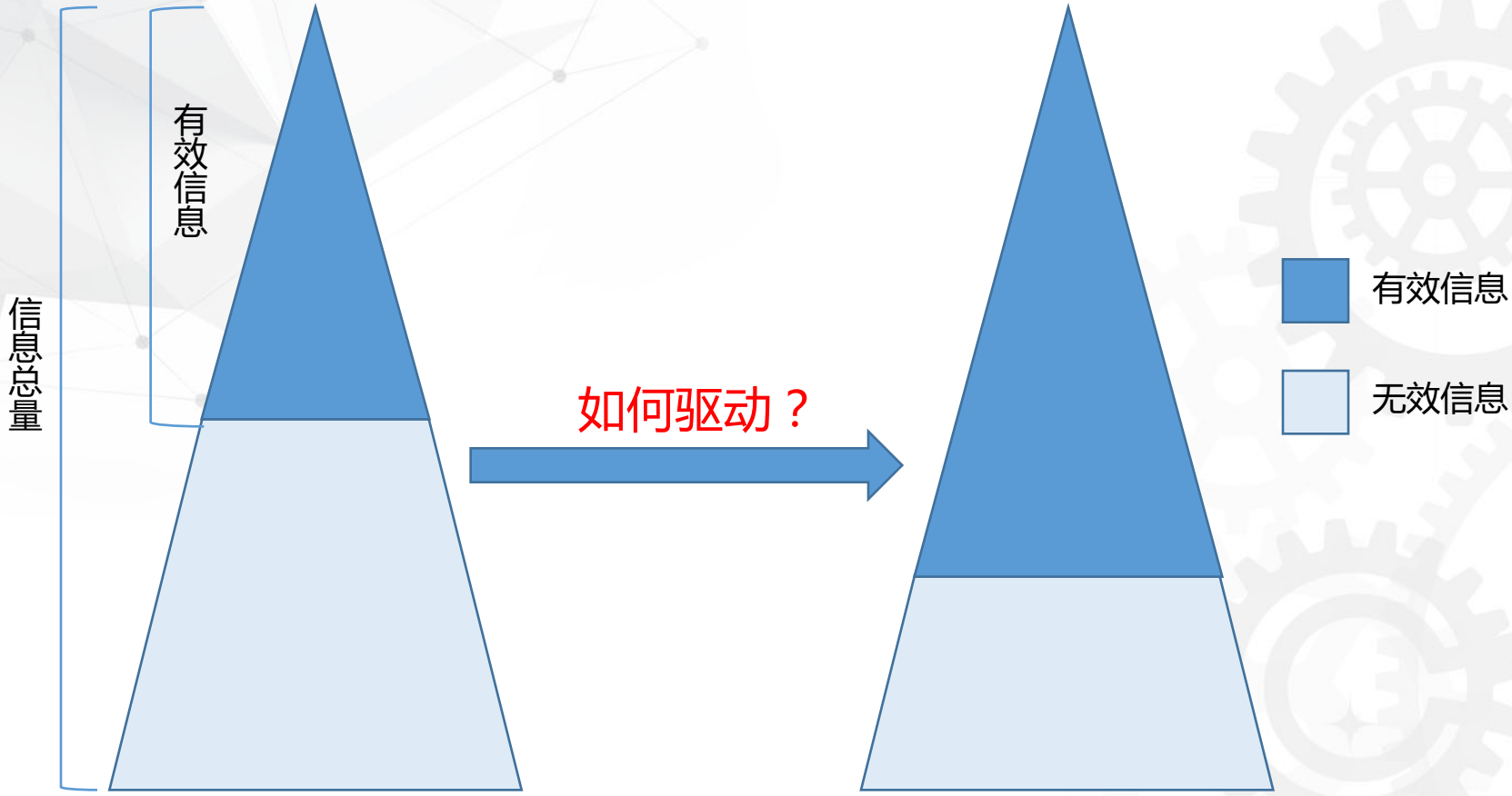


艾瑞

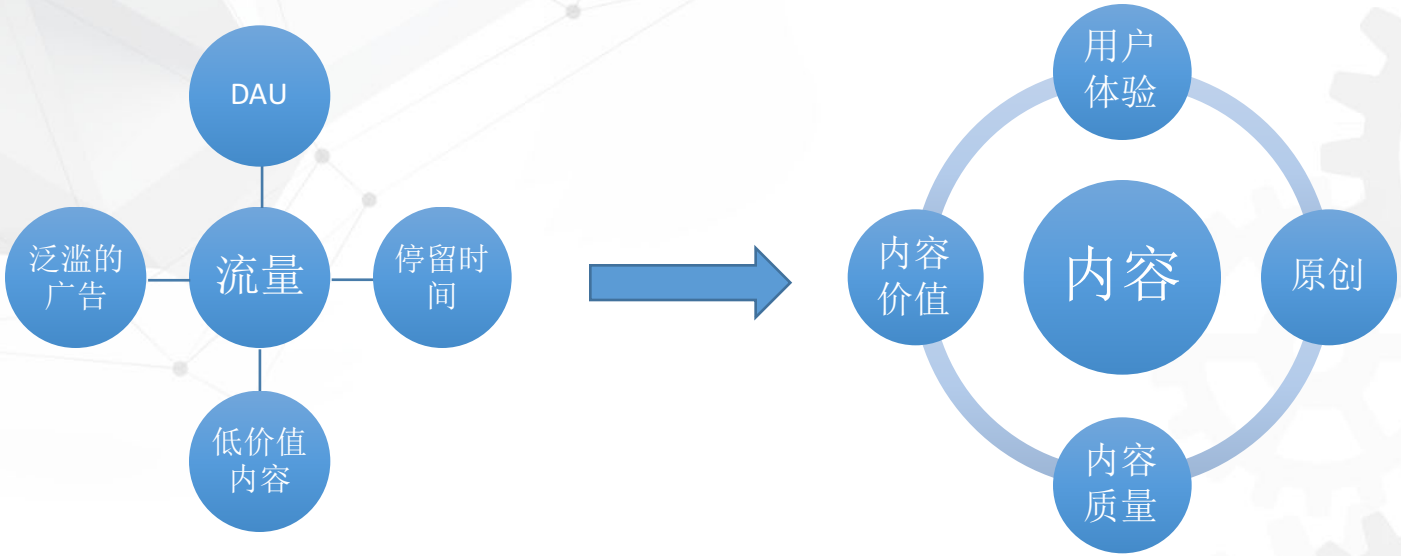
单纯以流量为基础的商业模式导致大量无效信息的产生

标题党 震惊体 美女蟒蛇图

驱动有效信息的有效传递



驱动商业模式从流量向内容转变



流量模式下，商业价值=DAU*停留时间

内容模式下，商业价值=内容价值*商业关联性

构建以内容为基础的商业模式，
驱动高质量内容的生产和分发

流量模式和内容模式下广告差异

广告位置：预先固定

采买方式：用户画像

广告量：极多

相关性：差

用户体验：差

VS

广告位置：动态生成

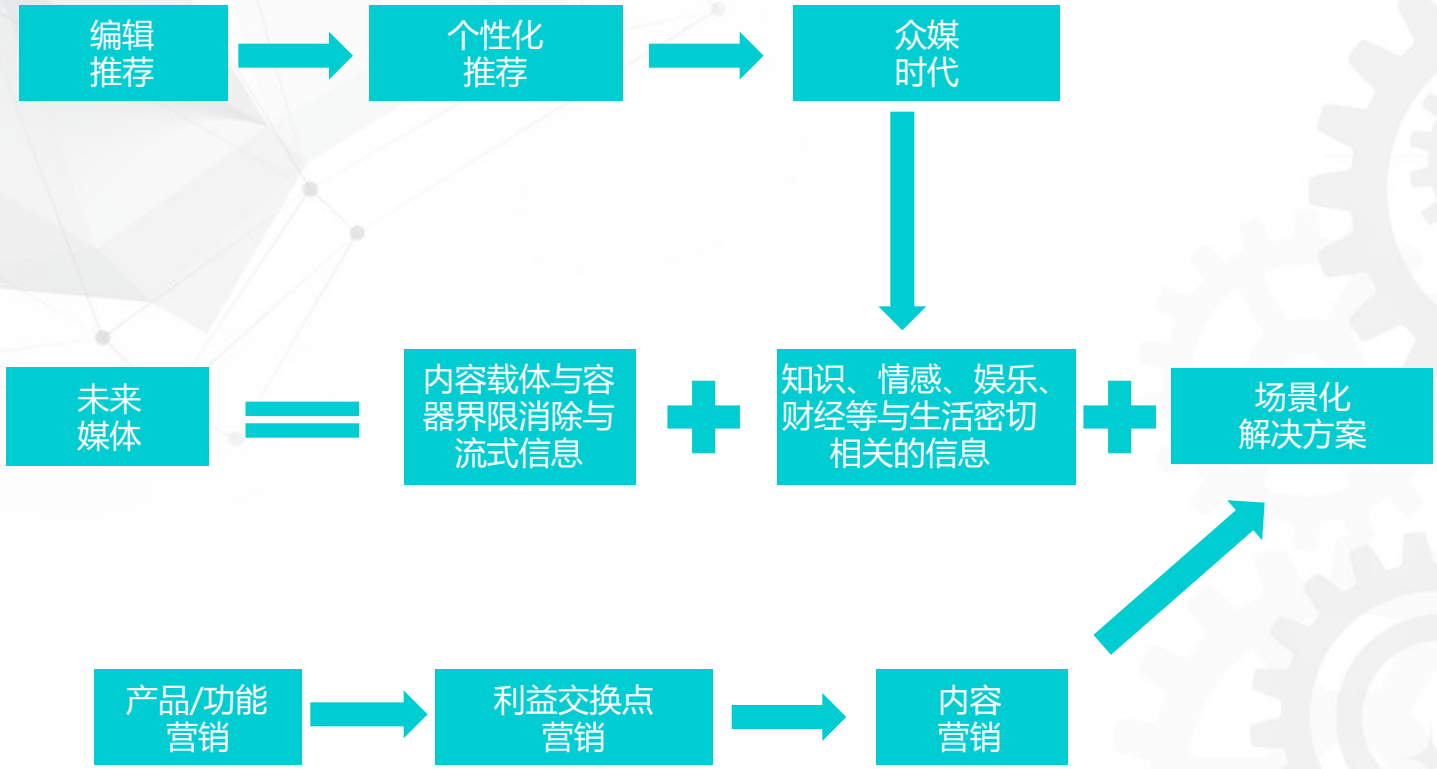
采买方式：商业语义

广告量：可控

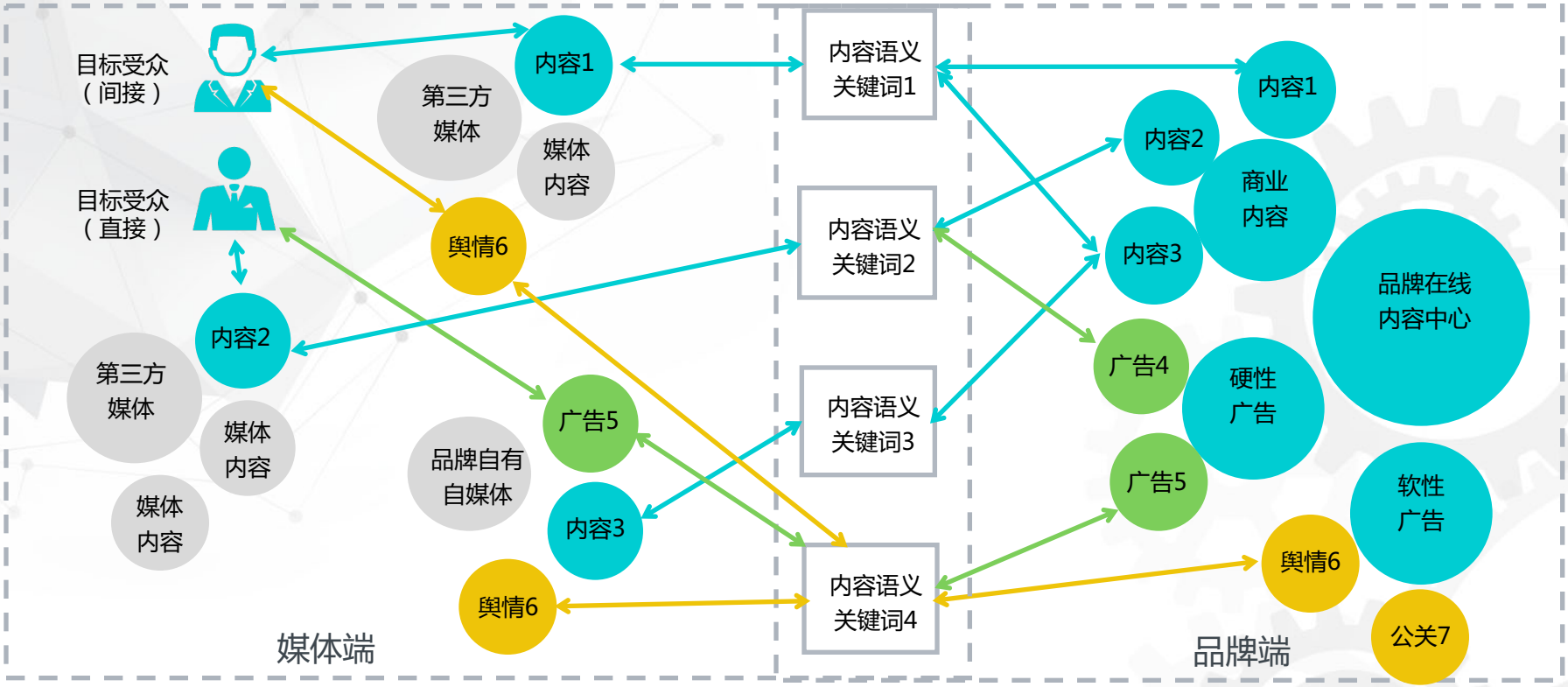
相关性：好

用户体验：好

人口红利殆尽，精细化、高质量的内容营销成为趋势

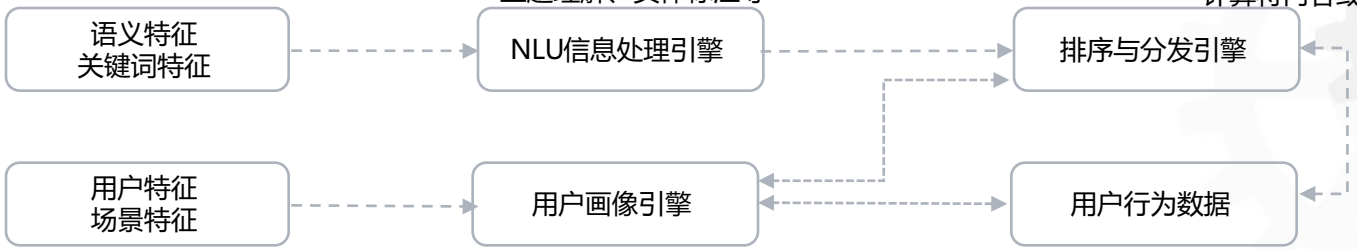


使用逆向搜索技术，实现基于商业语义的原生广告系统



信息处理引擎：将文档或广告进行分词、概念和主题理解、实体标注等

排序与分发引擎：通过相关性、质量、热度等计算将内容或广告推荐给用户

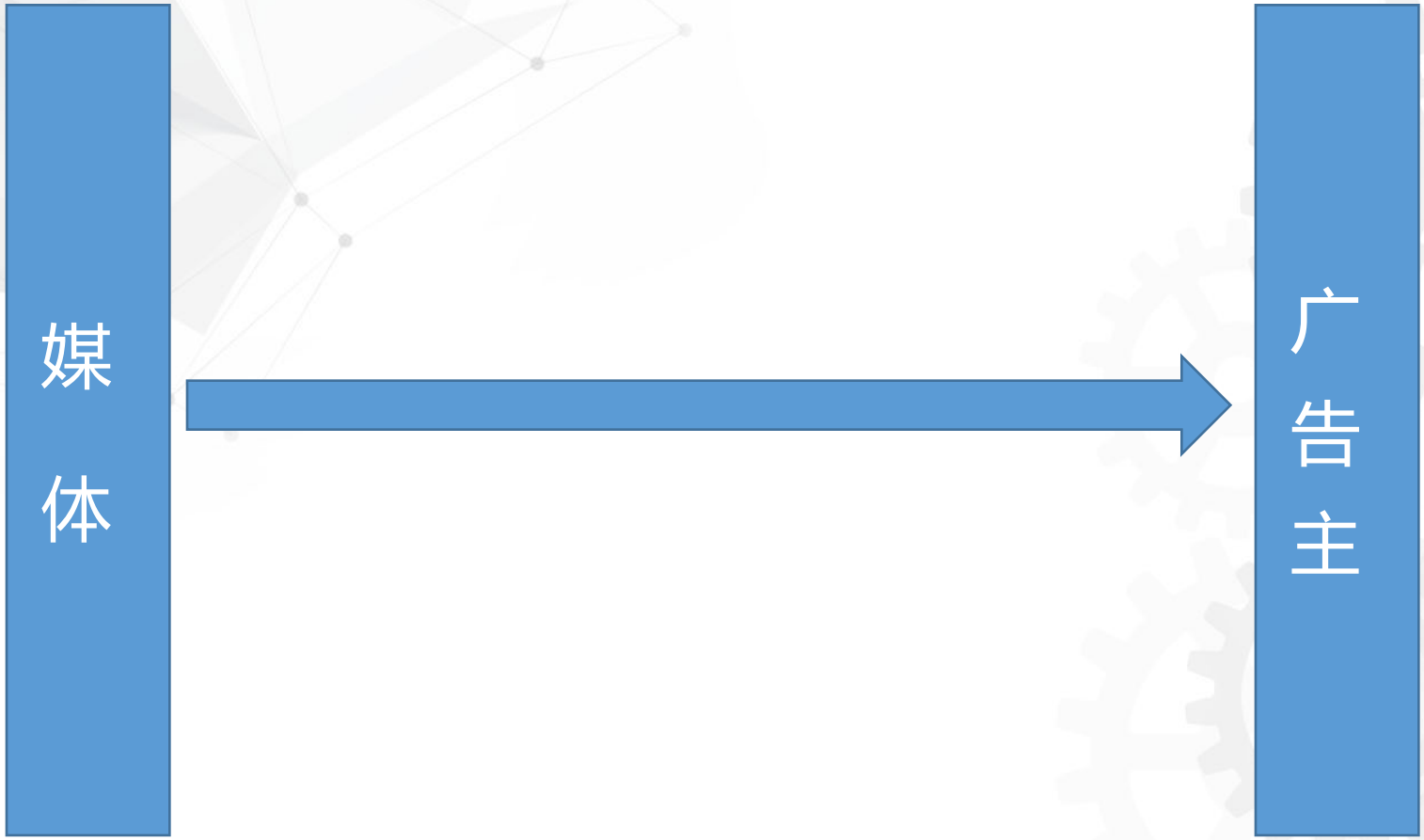


用户画像引擎：基于实时的用户行为数据，通过机器学习与特征聚类构建用户兴趣图谱

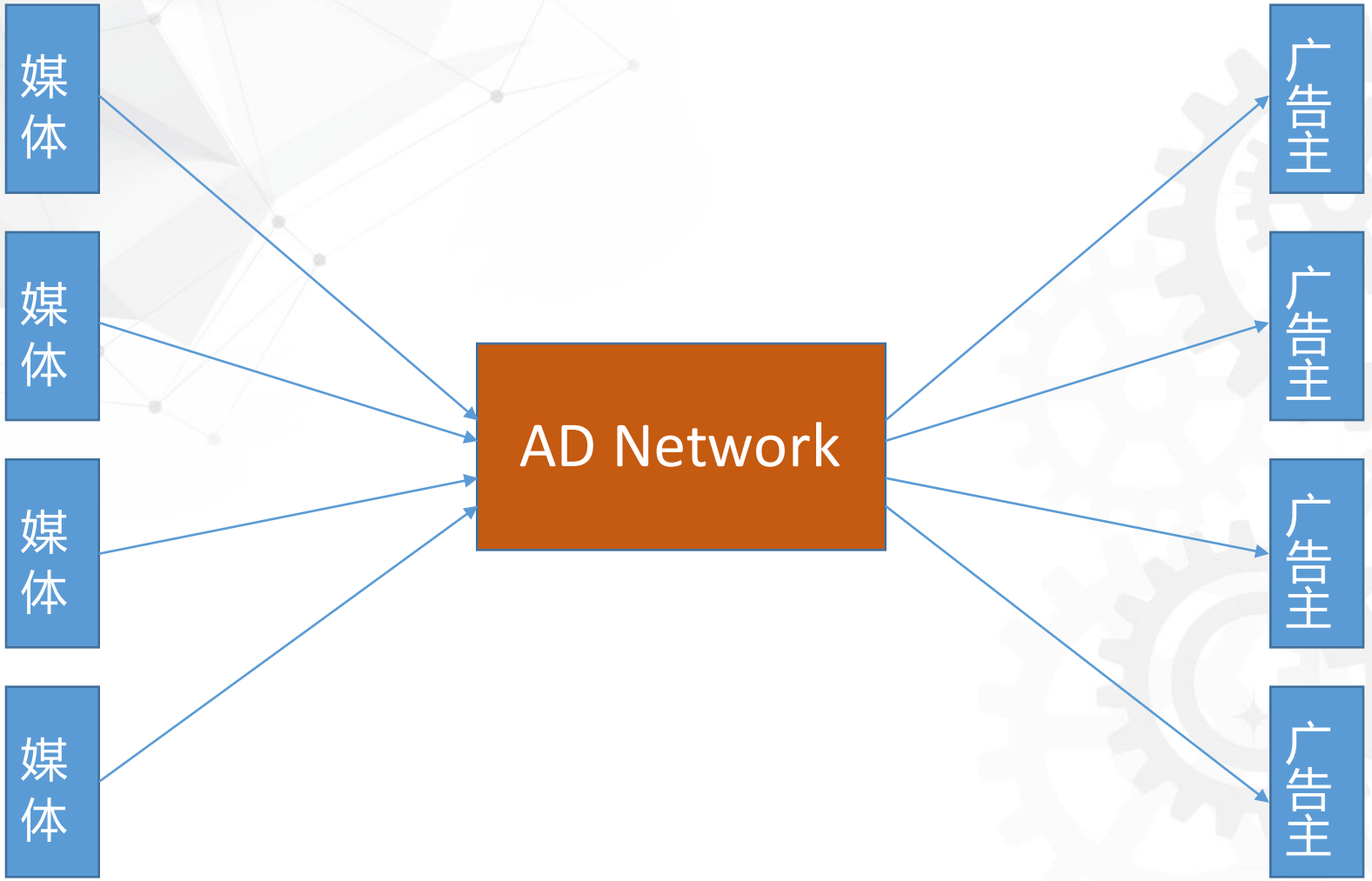
内容语义原生广告介绍

内容语义原生广告核心技术

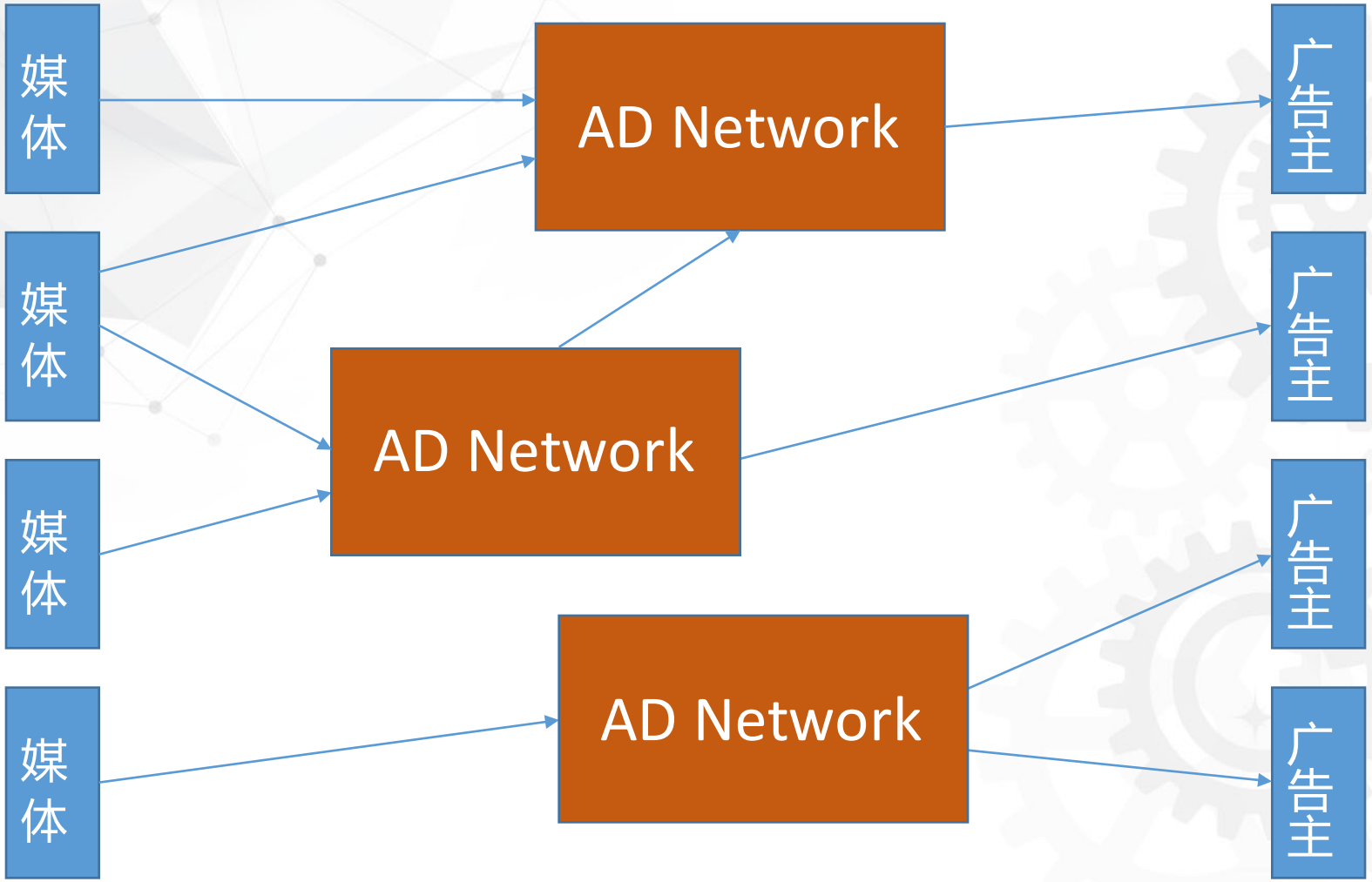
广告生态的演化过程



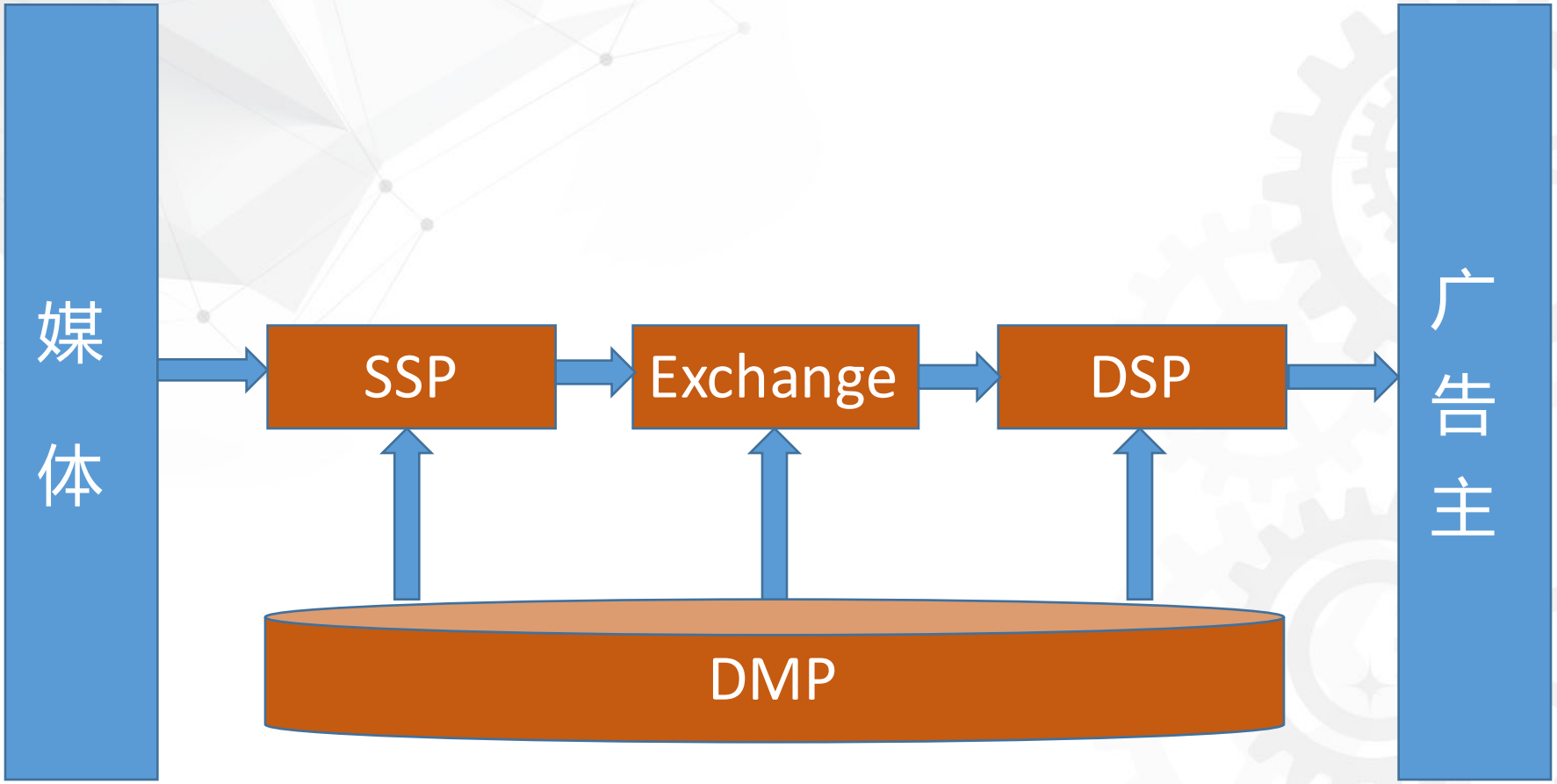
广告生态的演化过程



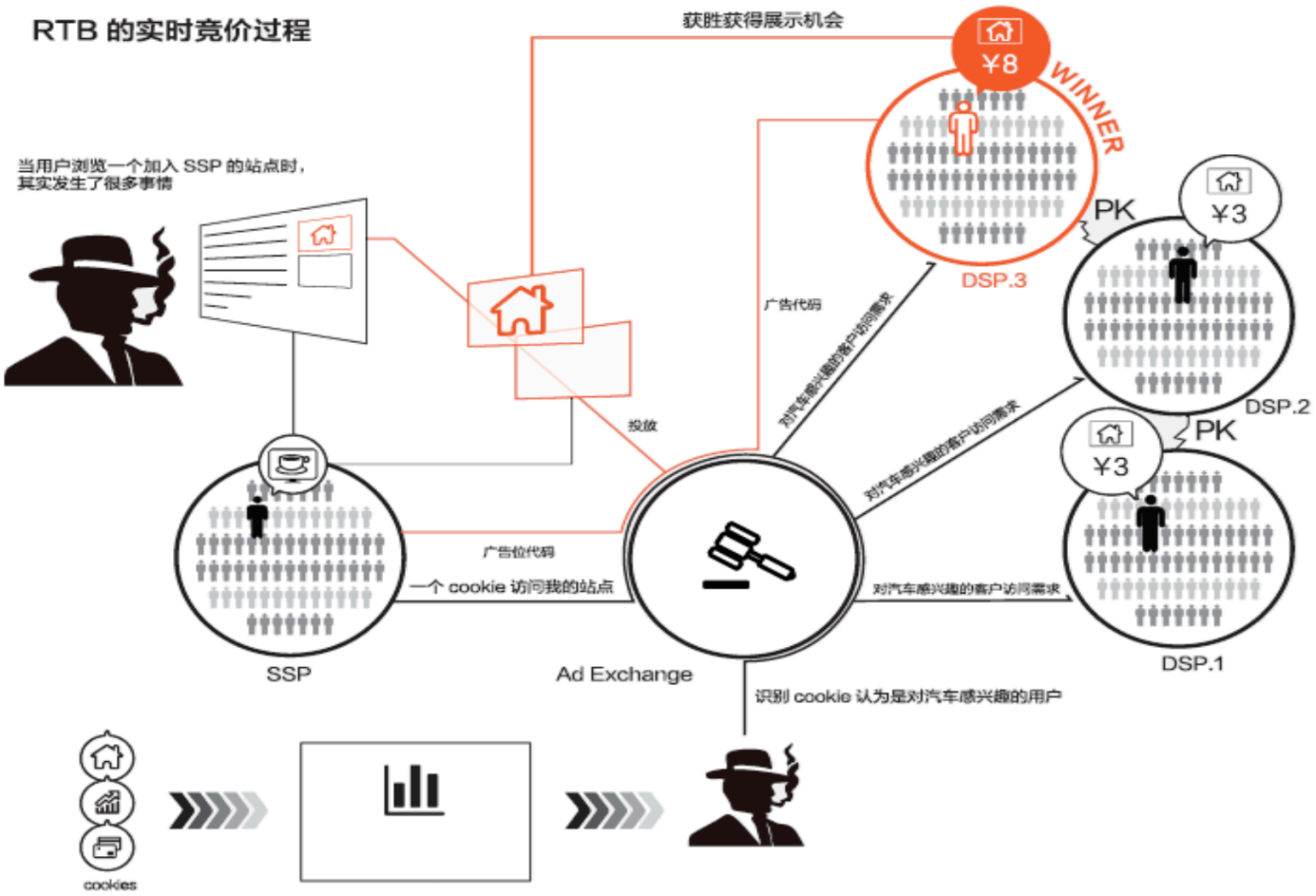
广告生态的演化过程



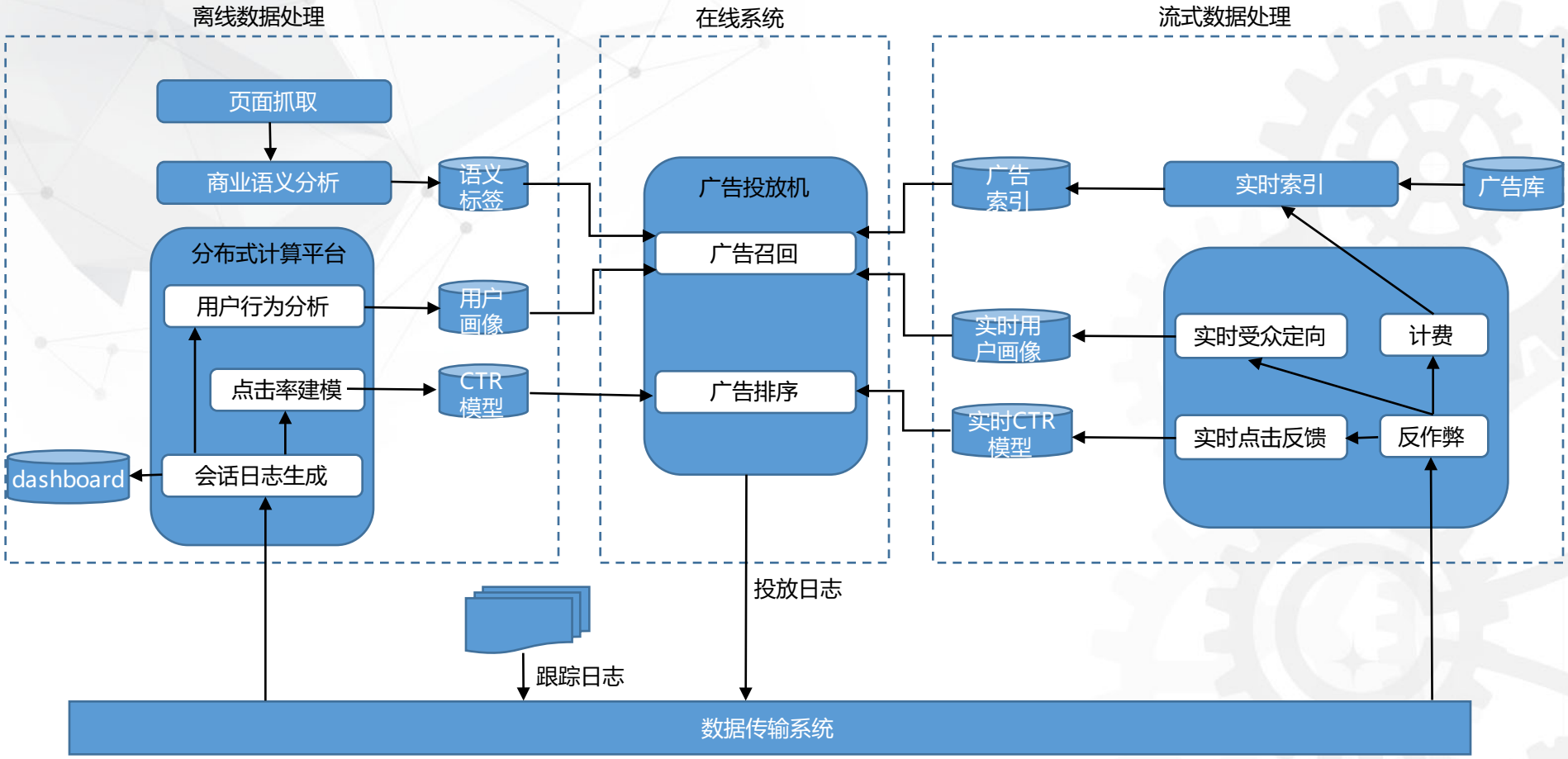
广告生态的演化过程



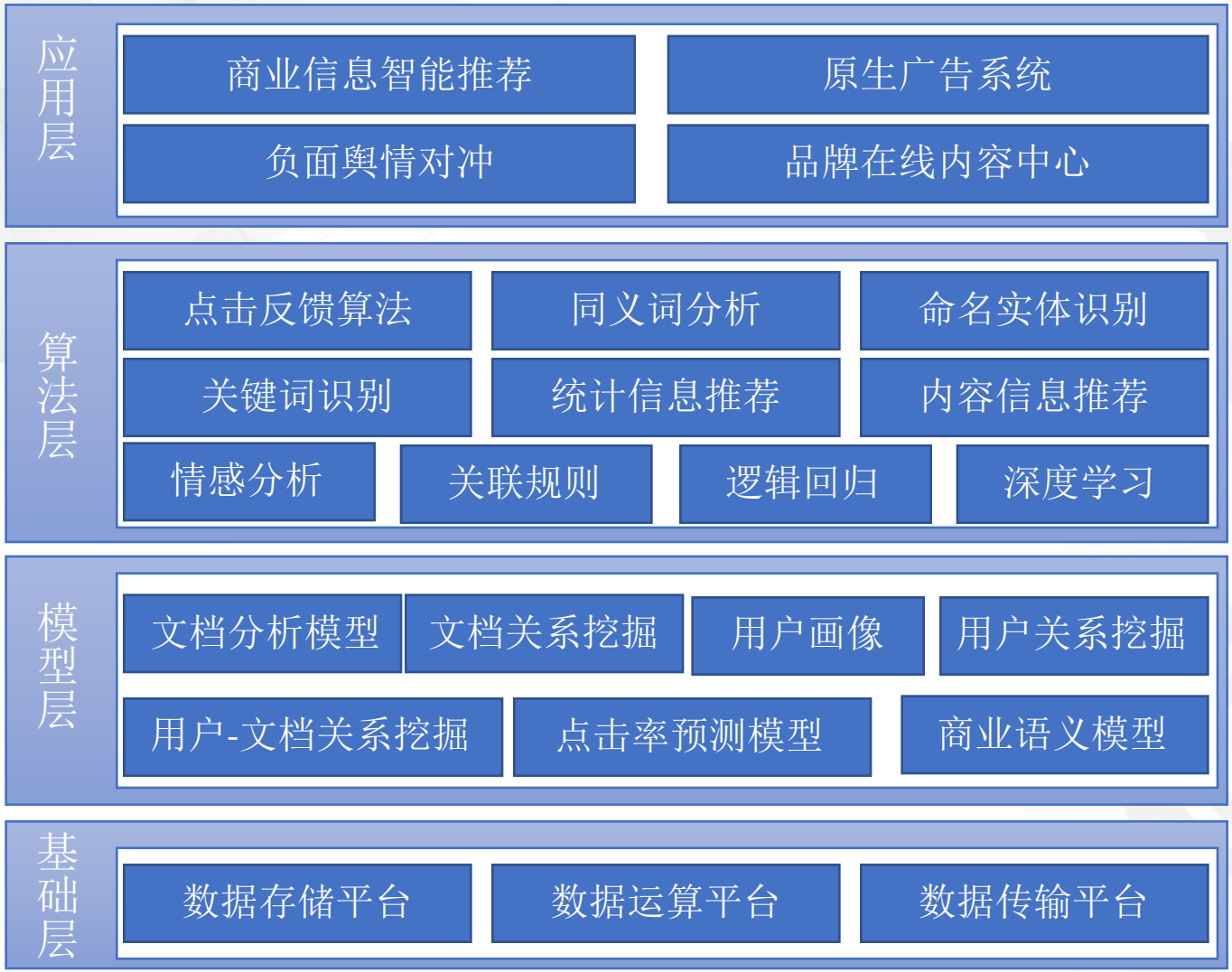
广告生态的演化过程


RTB 的实时竞价过程
获胜获得展示机会
¥8
WINNER
当用户浏览一个加入 SSP 的站点时，
其实发生了很多事情
DSP.3
PK
¥3
DSP.2
PK
¥3
DSP.1
广告代码
投放
对汽车感兴趣的客户访问需求
对汽车感兴趣的客户访问需求
对汽车感兴趣的客户访问需求
广告位代码
一个 cookie 访问我的站点
SSP
Ad Exchange
识别 cookie 认为是对汽车感兴趣的用户
cookies

商业语义原生广告总体架构

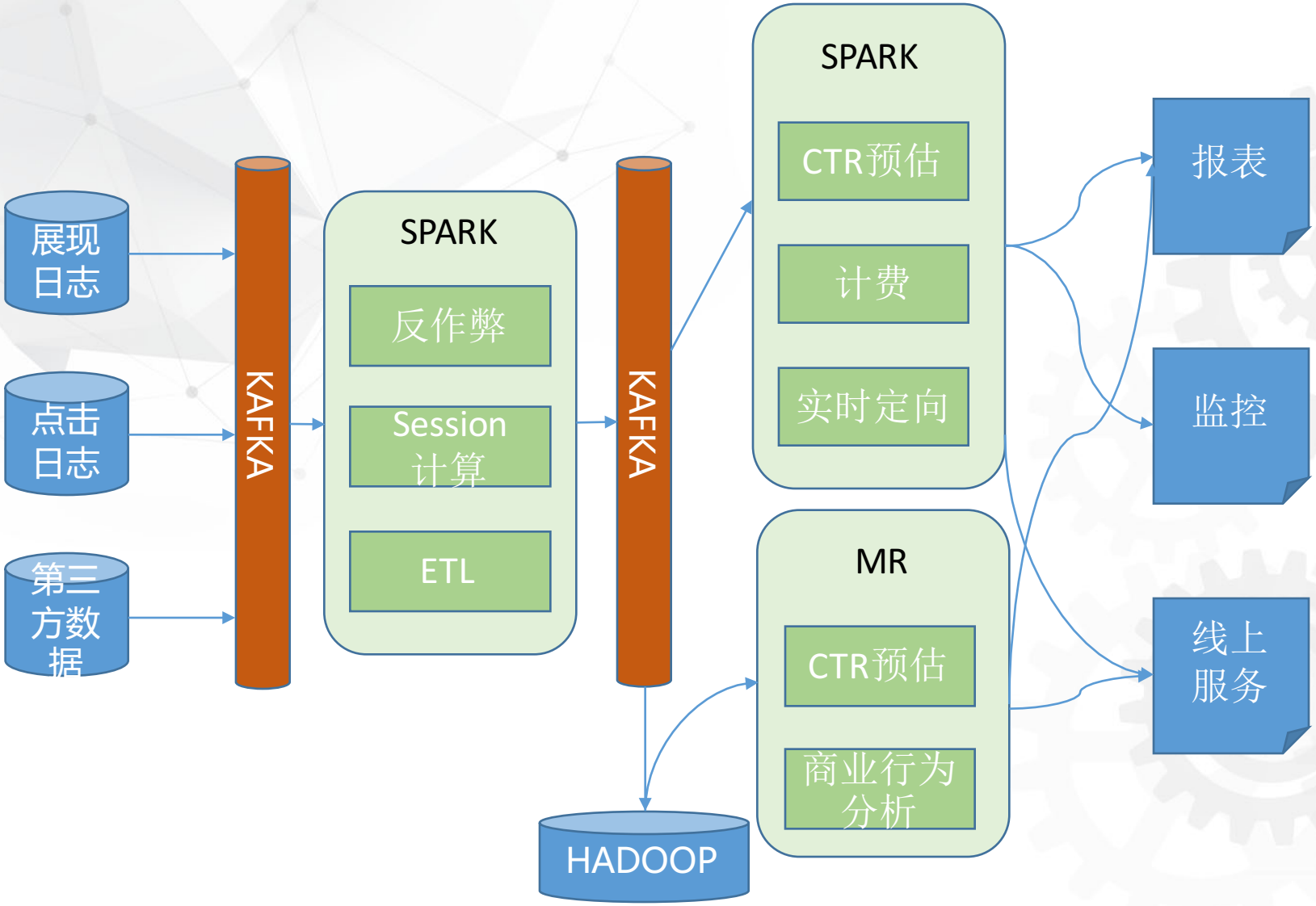


商业语义原生广告技术栈



- 日志系统设计
- 商业价值分析
- 语义匹配算法
- CTR预估

广告日志系统设计


展现日志
点击日志
第三方数据
KAFKA
SPARK
反作弊
Session计算
ETL
KAFKA
SPARK
CTR预估
计费
实时定向
MR
CTR预估
商业行为分析
HADOOP
报表
监控
线上服务

文档多维分析领域

原创 北京现代上海车展阵容 新名图/索纳塔等

2017-04-10 15:29:57 来源：PCauto 作者：陆杰豪 分享 评论[6]

【太平洋汽车网 新车频道】日前我们获悉了北京现代2017上海车展新车阵容，据悉在本届车展上，北京现代将带来新款名图、第九代索纳塔中期改款车型、一款全新SUV车型以及在2017CES上现代汽车所展示的高新科技成果。

● 新款名图

据悉，北京现代新款名图将会在4月19日正式开幕的2017上海车展上正式发布并公布售价。

分类：健康

情感：0.1

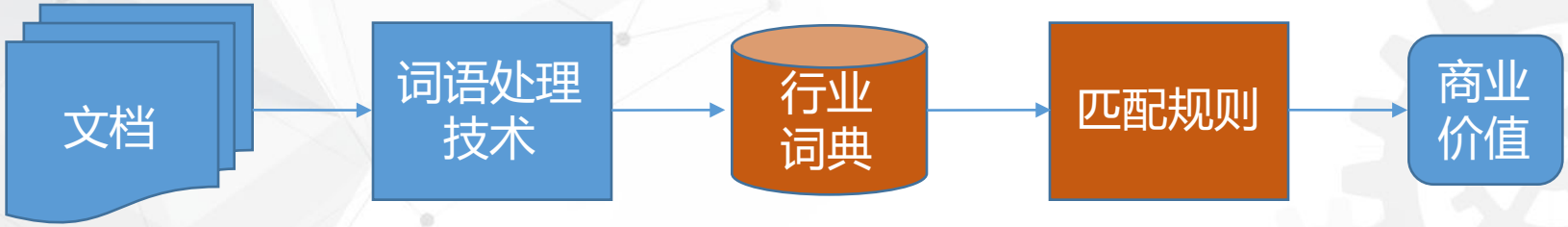
质量：8分

商业价值：8分

时效性：0.7

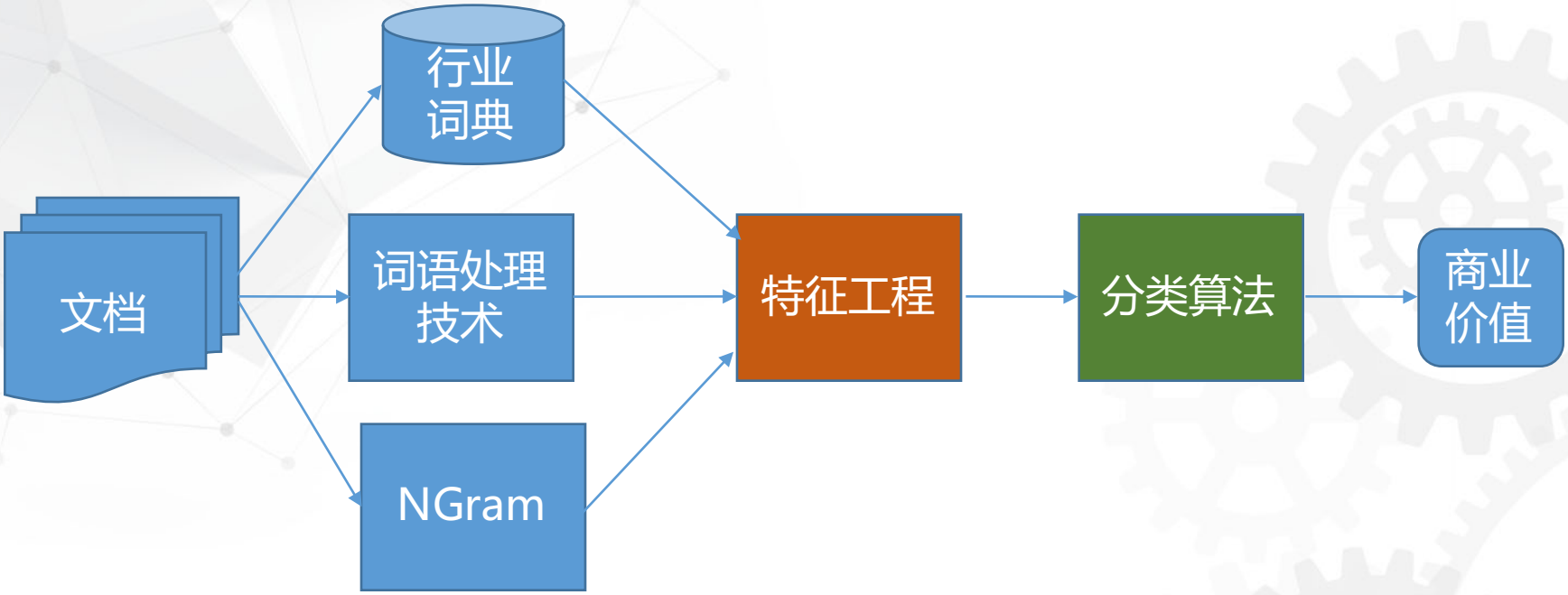
丰富度：0.8

商业价值分析-启发式规则



优点	缺点
• 简单易上手 • 启发式无需标注	• 可扩展性差 • 词典维护工作量大 • 匹配规则非常复杂

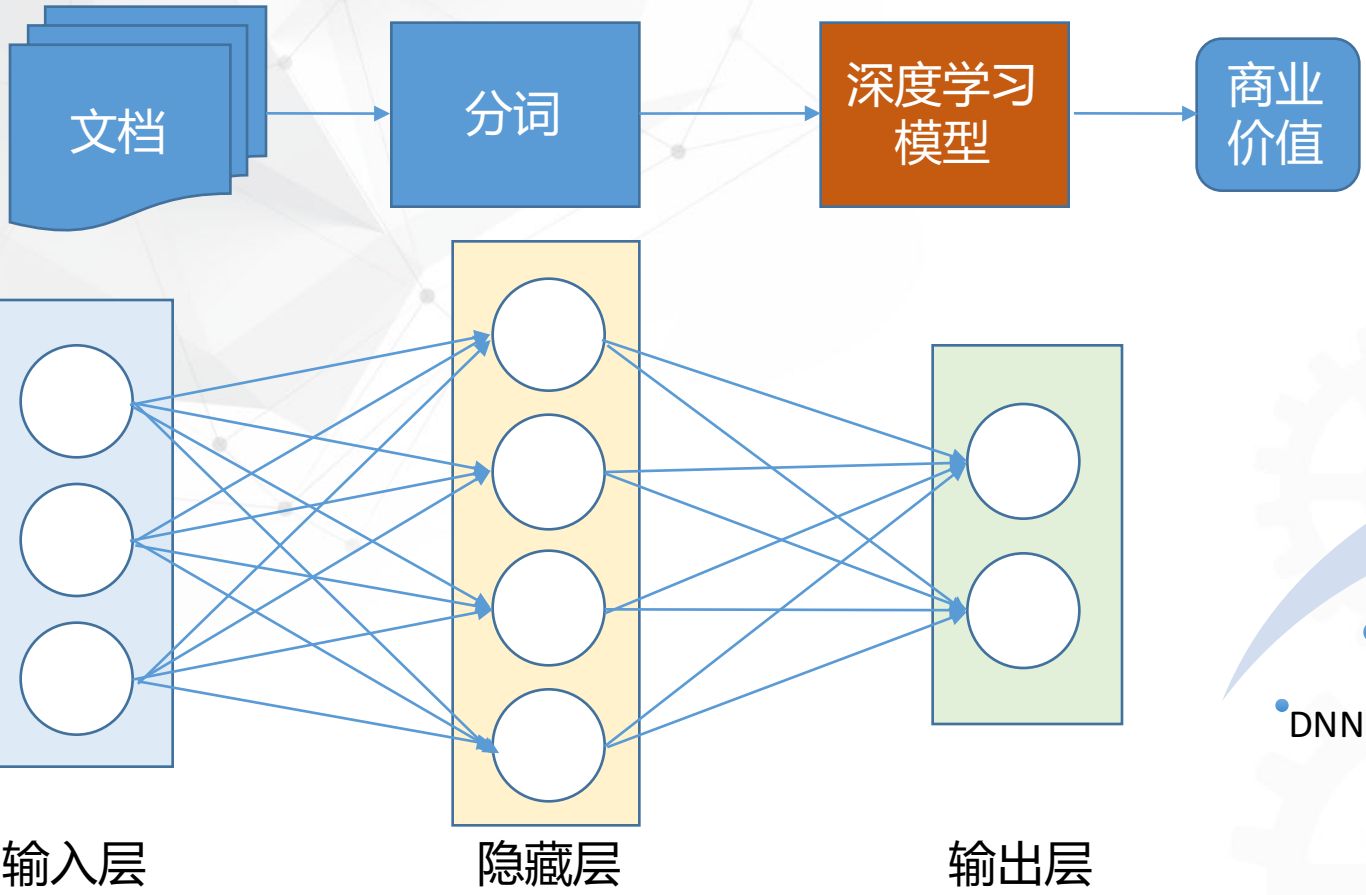
商业价值分析-机器学习

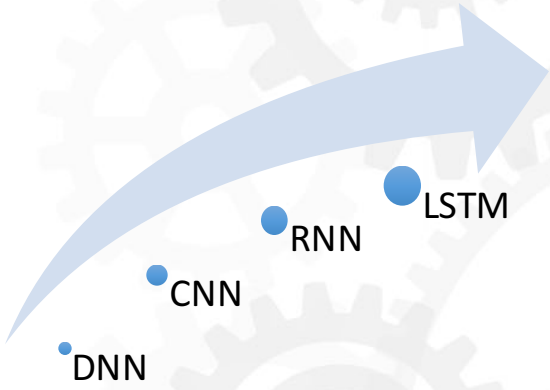


再不需要开发复杂的匹配规则，模型更加柔和。

但效果非常依赖特征工程的质量，需要非常多的特征工程经验。

商业价值分析-深度学习





End to End，深度学习专家要失业？

需要的是大量的标注数据。

商业语义匹配算法 – 基础工作

标题：北京现代上海车展阵容 新名图/索纳塔等

① 分词(基本词粒度) ：北京，现代，上海，车展，阵容，新，名图，索纳塔，等

② 分词（短语粒度）：北京现代，上海，车展，阵容，新名图，索纳塔，等

③ 词性标注：北京n，现代n，日前t，获悉v

④ 实体：书名，人名，地名，NER…

⑤ Term weight：北京现代（0.3），上海（0.2），车展（0.2），阵容（0.07），新名图（0.1），索纳塔（0.1），等（0.03）

商业语义匹配算法 – bm25是商业语义匹配的baseline

命中词的数量
正向贡献

Term Frequency，置信度，并且考虑到了频率饱和度

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

Inverse Document Frequency，
体现单个词的重要性

商业语义匹配算法 – 未命中词处理

归一化命中

置信度为1.0的命中

例子：

- ✓数字格式归一化：1000克->1千克，1000千克->一吨
- ✓词干提取：apples->apple，loved->love
- ✓繁简体转化：汽車->汽车

同义词命中

置信度较高的命中

例子：

- ✓汽车(0.7)->轿车，小孩（0.9）->孩子

相关词命中

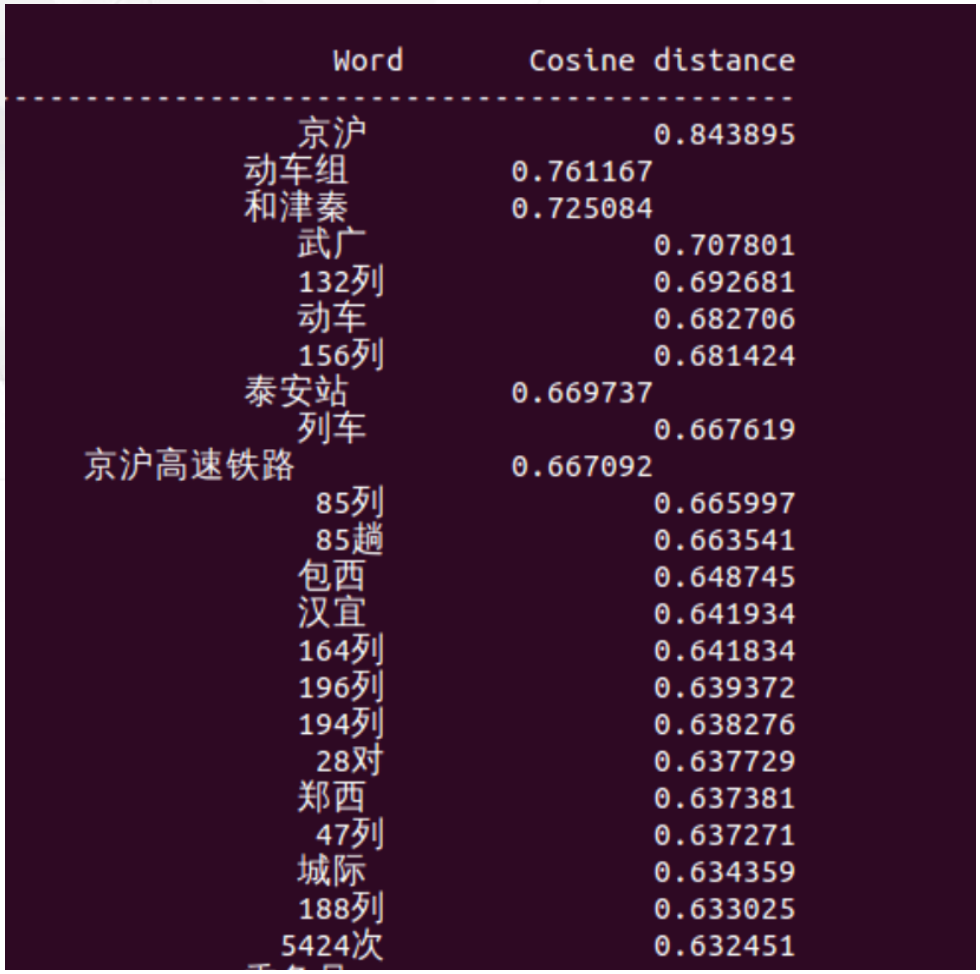
置信度较低的命中

例子：

- ✓现代（0.1）->驾驶，健身(0.05)->滑雪

利用word embedding处理词语之间的“爱/恨”关系

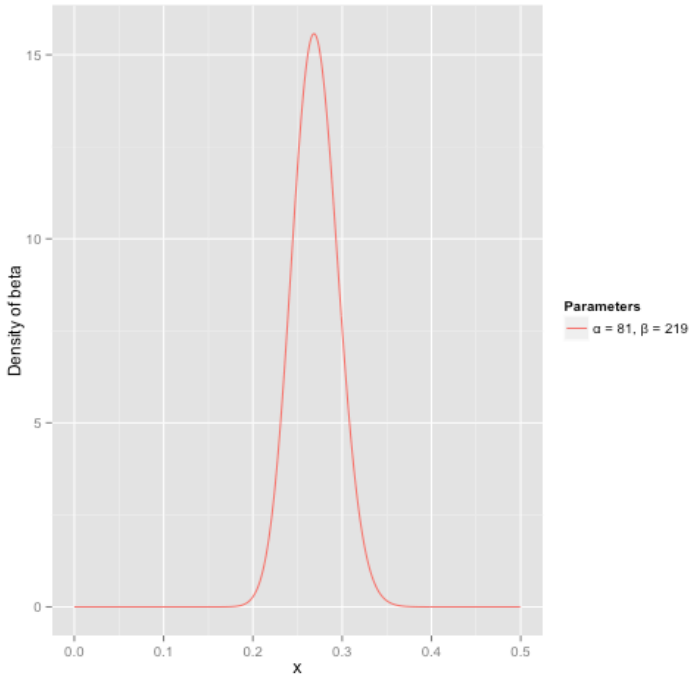
与【高铁】有关的词

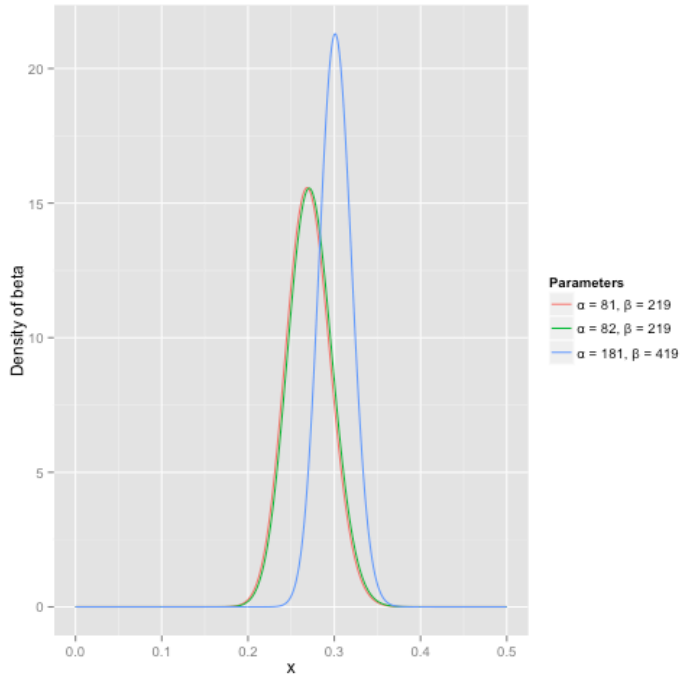


Word	Cosine distance
京沪	0.843895
动车组	0.761167
和津秦	0.725084
武广	0.707801
132列	0.692681
动车	0.682706
156列	0.681424
泰安站	0.669737
列车	0.667619
京沪高速铁路	0.667092
85列	0.665997
85趟	0.663541
包西	0.648745
汉宜	0.641934
164列	0.641834
196列	0.639372
194列	0.638276
28对	0.637729
郑西	0.637381
47列	0.637271
城际	0.634359
188列	0.633025
5424次	0.632451

CTR预估 – 冷启动的处理

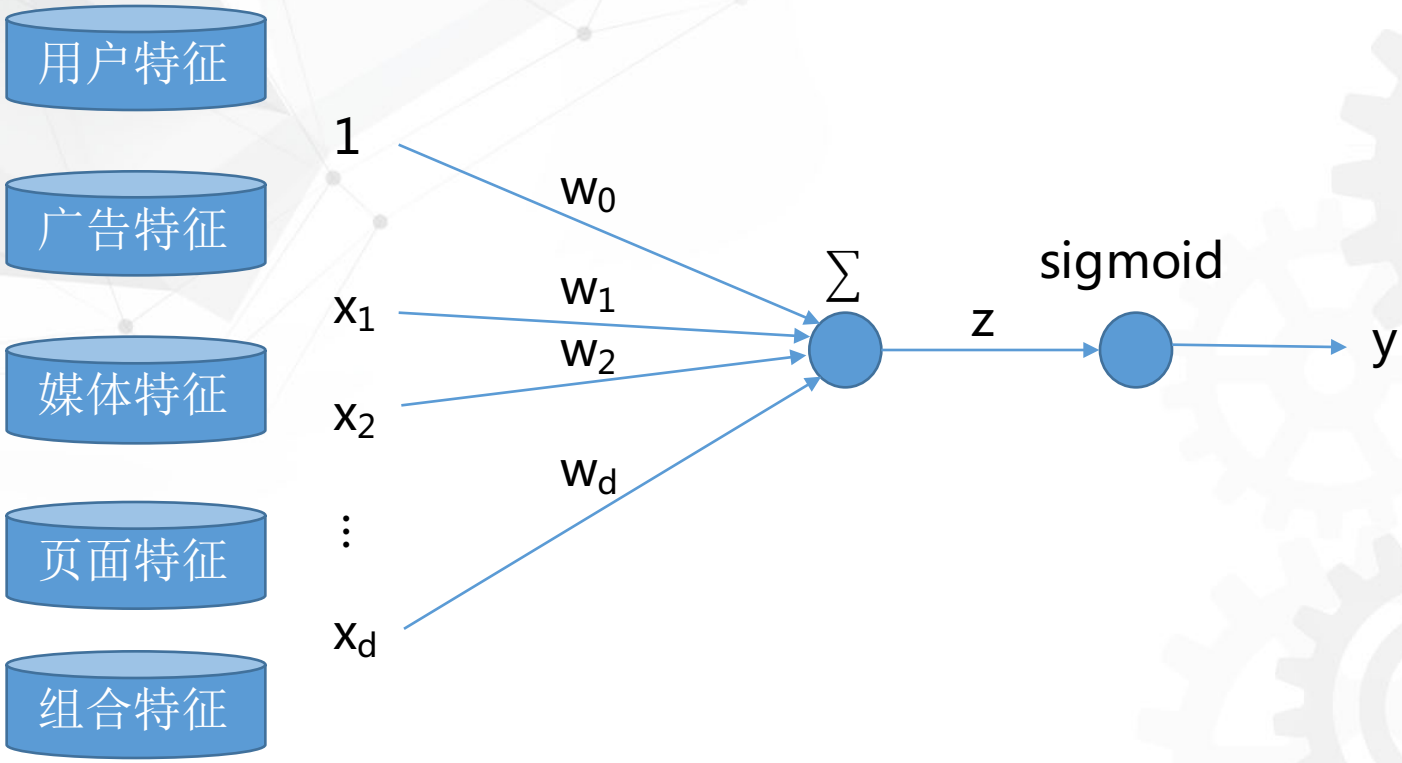
$$Beta(a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \propto \theta^{a-1}(1-\theta)^{b-1}$$





Beta分布建模，解决系统冷启动问题

CTR预估 – logistic regression



THANKS

SequeMedia 盛拓传媒 IT168.com ITPUB ChinaUnix