以 Facebook 和 Google为例的国外

信息茧房案例研究

Information Cocoons on Facebook and Google

新浪新闻 Sina News

新榜研究院 INSTITUTE OF NEWRANK

目录 CONTENTS

1.0

概念解析

1.1 信息茧房的实质是信息消费个人化、消费内容重复化、消费群体割裂化

互联网为我们快速获取目标信息创造了便利，却犹如硬币的两面性，同时也可能于无形之中诱导我们进入信息密闭的空间。2006年，桑斯坦首次提出“信息茧房”的概念，随着互联网和信息技术的发展，网络用户获取的信息越来越个性化，信息茧房现象也趋于明显。

信息茧房概念发展

信息窄化
2003年
桑斯坦

用户获得自己喜欢的信息，而拒绝接收自己不喜欢的信息，导致接触的信息越来越窄化、信息空间密闭化。

信息茧房
2006年
桑斯坦

用户往往只会注意喜欢并选择的领域相关的信息，久而久之会陷入信息茧房中。

信息茧房实质和内涵

1. **信息消费个人化**
   用户可以为自己打造一份“个人日报”，只接触自己喜欢的信息，排除不需要的信息。
2. **回音室效应**
   在一个相对封闭的环境上，一些意见相近的声音不断重复，形成对事物片面的认知，并坚持认为这是事实的全部。
3. **网络空间巴尔干化**
   网络如现实空间的分割将网民分裂成有着不同利益目标、价值信仰、兴趣取向的繁多群类，将不同群体割裂开来。

资料来源：凯斯·R·桑斯坦《信息乌托邦》、《信息共和国》

1.2 Facebook和Google基于海量数据提供个性化服务

Facebook本质是：基于用户群组和好友的“志同道和”的偏好以及用户偏好，将好友或群组用户感兴趣的信息内容推荐给用户；Facebook的服务是从人与人之间的关系开始的。

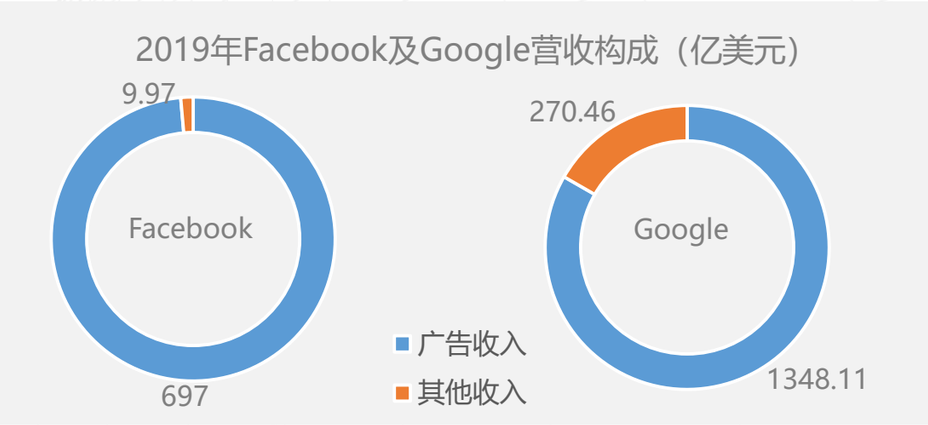
Google本质是：将搜索结果中的链接向上或向下移动或添加到用户的Google搜索结果中；Google的服务是从信息之间的关系开始的。

基于积累的海量数据和算法实现个性化信息服务

- 对于用户而言，Facebook及Google提供了精准高效的个性化信息服务和推送。

用户在默认同意用自己个人数据去交换便利、高效且免费的服务，其中包括Facebook社交媒体服务以及Google搜索服务；Facebook社交媒体服务通过为用户提供人与人间关系的服务的同时，获取用户相关数据，并推送个性化的信息及广告。Google在提供搜索服务的同时也会获取用户数据，并推送个性化信息和广告。

- 对于广告商而言，数据和算法是Facebook及Google精准寻找可能的买家的关键。此外，广告收入是Facebook及Google主要的营收来源。



资料来源：Alphabet 及Facebook2019年报

2.0

Facebook

2.1 Facebook存在信息茧房，且有四个具体表现

信息茧房
具体表现

01 用户信息获取渠道受社交媒体上的朋友、群组及他们分享的内容的影响

02 用户信息获取还取决于信息流排名算法如何对这些文章进行分类并匹配个人的兴趣标签

03 用户发布的内容之间存在很大的两极分化

04 用户在网络中主要按其意识形态聚类

Facebook信息茧房的具体表现（1）

Facebook上信息茧房有四个明显的表现：

用户信息获取渠道由传统媒体渐变成社交媒体：用户消费的信息内容取决于他们的朋友/群组好友是谁以及这些朋友分享什么信息

- 用户在选择成为Facebook好友的对象偏向于与自身生活方式、政治派别和同质化的群体；此外，高度同质化用户组成的群组也倾向于分享用户偏好的信息，导致信息窄化和人群极化。
- 用户偏向于从同质化人群组成的群组中获取信息，截止2019年Facebook上群组数量超过1000万个。

用户信息获取方式由主动变为被动：用户在Facebook上消费的媒体不仅取决于他们的朋友分享什么，还取决于信息流排名算法如何对这些文章进行分类并匹配个人的兴趣标签

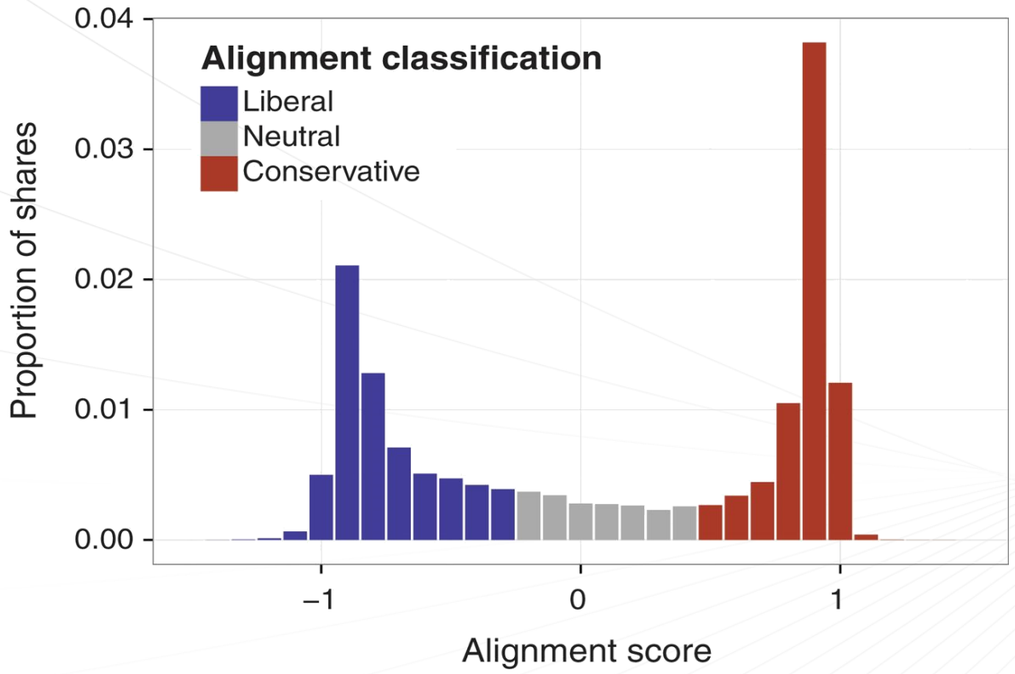
- Facebook从以下维度：用户位置、手机和电脑（型号）、互动的广告类型、是否结婚、生日、工作单位等生活方式获取用户信息并分类，并利用其信息流算法提供个性化服务和广告。

Facebook信息茧房的具体表现（2）

Facebook上信息茧房有四个明显的表现：

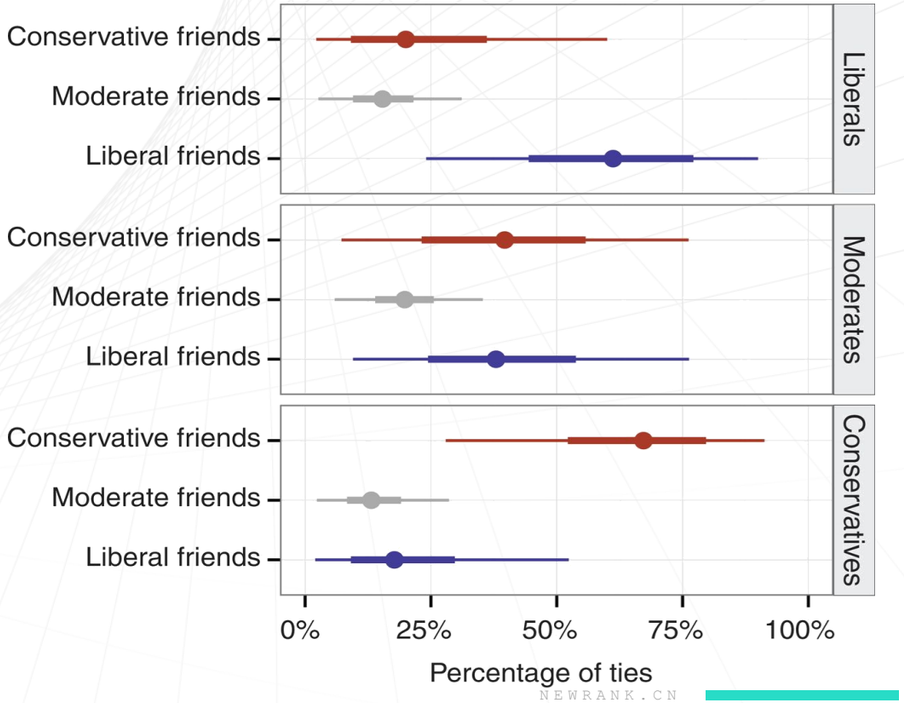
用户发布的内容之间存在很大的两极分化，用户消费及生产的信息与自己的意识形态相似的人群保持一致

对不同意识形态的用户所分享的内容类型研究发现：保守派用户倾向于分享保守思想的内容，自由派倾向于分享自由思想的内容。



用户在网络中的主要按其意识形态聚类，但也有许多跨越意识形态派别的连接。

对不同意识形态的用户所关注的好友所属意识形态（政党派别）研究发现：保守派用户与自由派用户倾向于与自己相同意识形态的群体成为好友。中间派用户好友中保守派用户及自由派用户大致一样多。



资料来源：Science

2.2 Facebook信息茧房形成原因分析

Facebook信息茧房形成因素众多，其中一个主要因素是Facebook对信息内容相关的算法。另外的因素，包括用户主动寻找消费内容、群组内分享的内容等等
Facebook的新算法被认为是基于Vickrey-Clarke-Groves（维克里-克拉克-格罗夫斯机制）算法，该算法“作为封闭式拍卖运作”。

自2017年围绕社交网络数据爆发争议以来，Facebook一直努力提高在"新闻源"内容排名上的透明度。
Facebook算法根据用户有积极反应的可能性，对可以在用户新闻源上显示的所有可用帖子进行排名。
Facebook现在对好友发布的内容进行排名和优先级，而不是发布者，Facebook进行排名和优先级是为了带给用户"有意义的交互"的内容。

有意义的交互

被动交互：

阅览时长、故事类型、发布时间和其他非活动指标。

主动交互：

包括赞、分享、评论和其他活动事件，这些活动会促使参与。如：评论、回复、点赞、分享

Facebook 对新闻消息内容进行排名的算法基于四个因素：

Inventory：

所有可供显示的员额的清单

库存代表可以在Facebook的新闻源上显示给用户的所有内容的库存。 这包括从朋友和发布者发布的所有内容。

Signals：

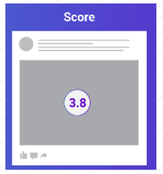
解析Facebook每个帖子代表的是什么信号

信号代表Facebook可以收集的关于一条内容的信息。信号是您可以控制的唯一因素。Facebook分析解析用户发布内容的维度包括：内容类型，发布者，年龄，目的等。

Predictions：

预测用户将如何对每个帖子做出反应

预测表示用户的行为，以及预测用户与内容片进行正向交互的可能性；预测会考虑真实的参与度，例如评论，喜欢和真实个人资料中的分享。



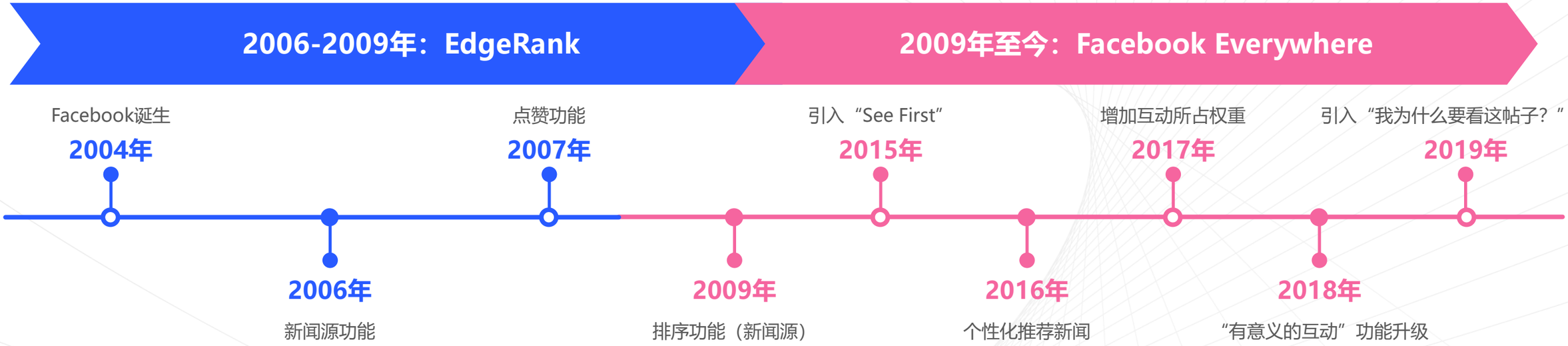
Scores：

根据考虑的所有因素分配给内容的最终分数

Facebook权衡每个预测并得出一个数字来表示他们认为故事对你有多重要。每次访问Facebook时，此过程都会发生在用户的NewsFeed中的每个故事中。

2.3 算法发展历程：EdgeRank到Facebook Everywhere的个性化服务

Facebook信息流推荐算法经历了两个阶段：第一阶段（2006-2009年）：用户数量相对少，以简单的好友亲密度、信息发布量和时间为主要参考依据提供信息服务。第二阶段（2009年至今）：因用户数量剧增，Facebook基于社交媒体打造信息服务生态，将内容导入至社交媒体平台，优化其个性化服务机制。



对社交媒体上的每次互动进行个性化排名

信息流推荐的考虑因素有：

- 亲和力：您与某人的友善程度更高（取决于您花费在与之互动并查看其个人资料的时间上），Facebook更有可能向您显示该人的最新动态
- 该类型内容的相对权重，例如：关系状态更新的权重非常高；每个人都喜欢知道谁与谁约会。（许多局外人怀疑重量也是个性化的：不同的人关心不同的内容。）
- 时间顺序：最近发布的帖子权重超过之前的帖子权重。

根据社交图谱（social graph）使整个Web成为“社交”网站的组成部分

到2009年，Facebook已达到3亿用户大关，并且每月增长1000万。

将新闻流植入至社交媒体：

- 在社交图（social graph）和Facebook用户提供的大量信息的基础上，把Facebook的新闻算法引擎放在网络的中心位置。
- 使整个Web成为“社交”网站，并将Facebook风格的个性化功能带到目前缺少它的数百万个站点中。

2.4 Facebook不认为信息茧房存在，只针重大社会问题对采取措施

Facebook CEO 马克·扎克伯格不认为信息茧房存在；

Facebook在Facebook算法、用户或者社群方面的政策变化是基于美国社会层面的问题，而非针对信息茧房。

措施

- Facebook具体措施针对美国比较明显的社会问题，包括严格管理群组内的极端言论、种族歧视等方面的言论等。

风险

- 高度同质化的群体，在接受与其观念相似的（虚假）信息或者接受与其完全相反的真实信息时，都会确认并强化其原有观念。
- 在不同群体形成的网络社区（信息茧房）中引入真实的信息来纠正或“揭穿”虚假信息时，它要么被忽略，要么增强了用户的错误信念，从而进一步强化信息茧房；

资料来源：fastcompany.com、NewScientist

3.0

Google

3.1 Google存在信息茧房，且有三个具体表现

信息茧房
具体表现

01 对不同的搜索者展示结果不同；有些重要的搜索结果会被因人而异地展示

02 个性化服务不仅体现在网页搜索，在新闻和视频信息框中也存在很大的不同

03 个体用户私密浏览（退出Google账号）模式下的搜索结果与正常模式下的情况大致相同

表现1：不同用户间搜索相同词汇所得结果不同

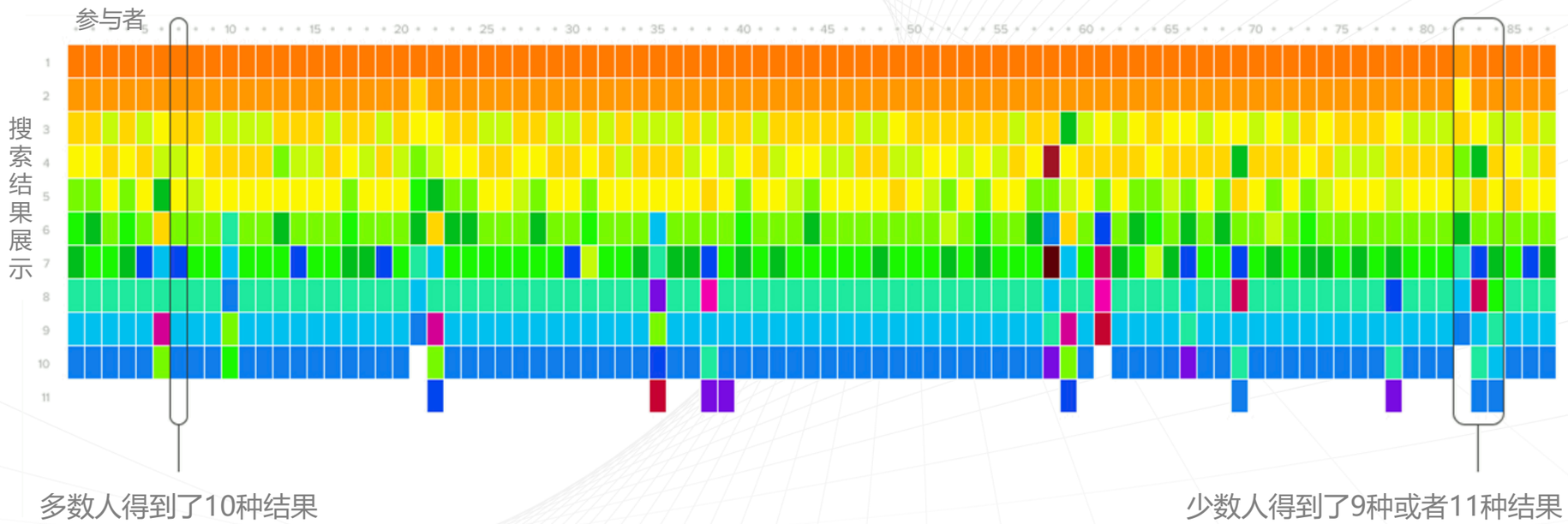
不同用户同时输入相同搜索词时，绝大多数人看到的结果对他们来说都是独一无二的。
某些用户搜索得到一些不寻常的结果；此外，用户并无法知道自己搜索结果的缺失。

在全美范围内随机抽选的87名用户，在处于私密浏览模式下同时Google搜索“枪支控制”（gun control），发现19个域名（链接）以31种方式呈现。

域名（链接）出现频率

procon.org	99%	nytimes.com	91%
wikipedia.org	99%	npr.org	17%
justfacts.com	100%	smithsonianmag.com	6%
texastribune.org	99%	nraila.org	2%
chicagotribune.com	76%	newyorker.com	3%
huffingtonpost.com	100%	twitter.com	3%
allsides.com	67%	politico.com	2%
nbcnews.com	41%	washingtonpost.com	1%
cfr.org	99%	localdomain.com	1%
propublica.org	99%		

谷歌信息茧房域名(链接)变化



资料来源：DuckDuckGo

表现2：用户搜索新闻及视频，其结果存在着明显的个性化差异

不同用户在同一地点搜索同一词汇，其在新闻咨询框以及视频信息框所展示结果明显不同。

新闻及视频搜索结果分析

新闻资讯框：

76人搜索三个词汇及结果对比：

- “枪支控制”：来自5个来源的3个变化（编辑距离），出现在75人搜索结果中。
- “移民”：来自7个来源的6个变化（编辑距离），所有人搜索结果都不同。
- “疫苗接种”：来自3个来源的2个变化（编辑距离），出现在2人搜索结果中。

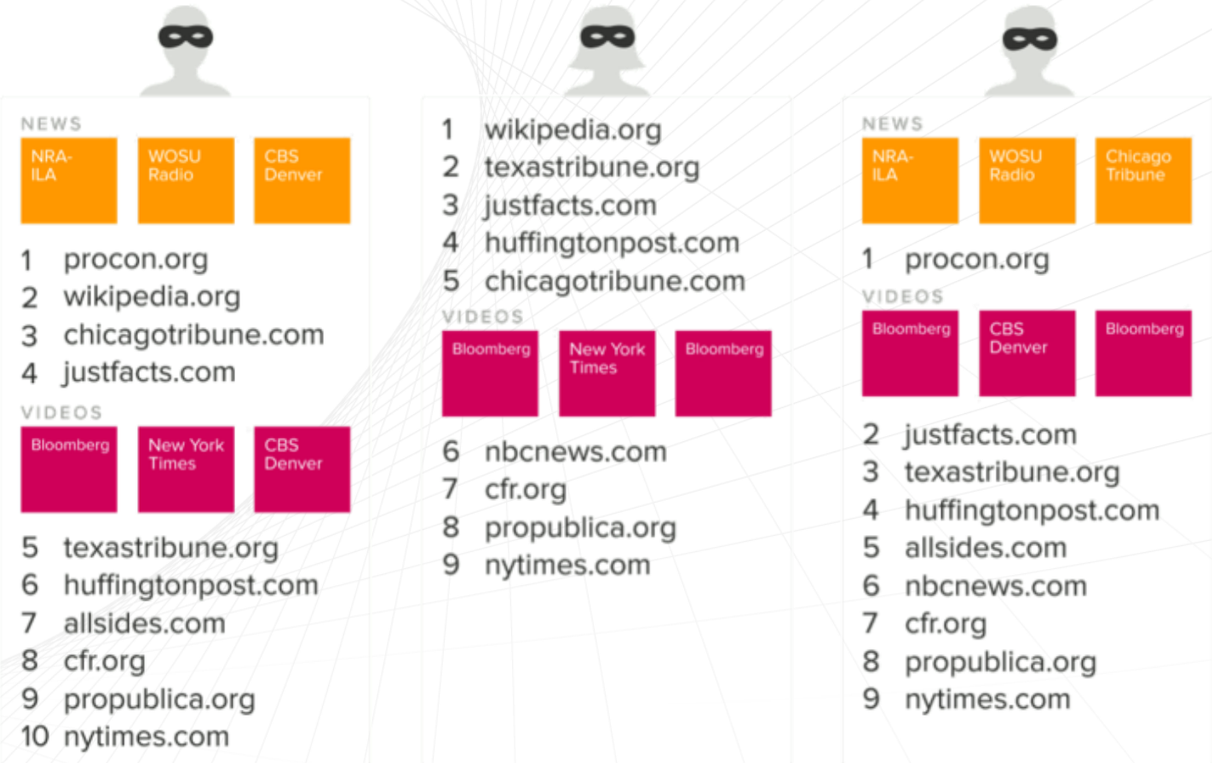
视频信息框：

76人搜索三个词汇及结果对比：

- “枪支控制”：来自7个来源的12个变化（编辑距离），出现在75人搜索结果中。
- “移民”：来自6个来源的6个变化（编辑距离），出现在75人搜索结果中。
- “疫苗”：未在搜索结果中显示。

以上变化（编辑距离）均指莱文斯坦编辑距离

参与实验的三个用户新闻及视频搜索结果



资料来源：DuckDuckGo

表现3：私密浏览模式并退出帐户后，信息茧房仍在起效

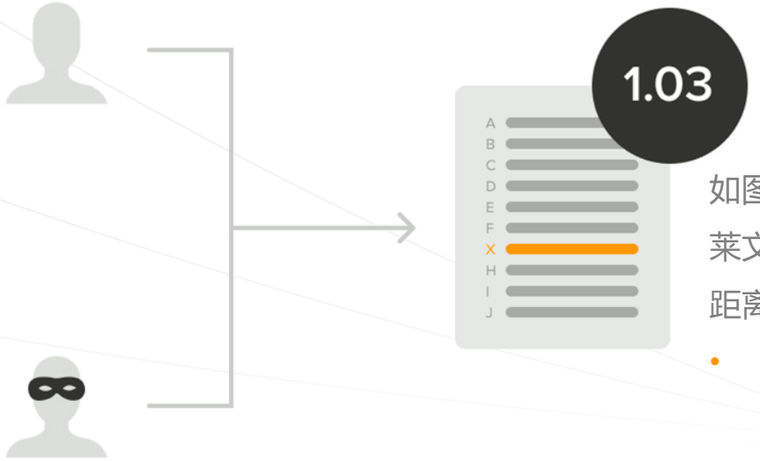
同一用户在私密浏览模式（退出Google账户）下的搜索结果与正常模式下的搜索结果大致相同。

不同用户私密浏览模式下的搜索结果仍有较大不同。

私密浏览和注销不会显著减弱信息茧房，信息茧房依旧在起效。

同一用户不同浏览模式下搜索结果变化

在不同的浏览模式下，用户平均有1个不同的域（即莱文斯坦编辑距离为1），这表明在私有浏览模式下Google信息茧房依旧在起效。

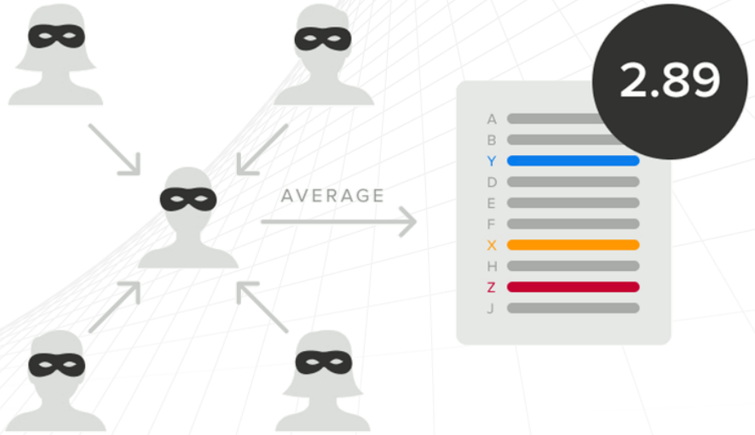


如图中有一个莱文斯坦编辑距离：

- X 替换 G

不同用户私密浏览模式下搜索结果变化

比较两个随机私人浏览模式结果时，平均会有近3个域变化（即莱文斯坦编辑距离为3），这表明在Google信息茧房提供个性化搜索服务。



如图中有三个莱文斯坦编辑距离：

- Y 替换 C
- X 替换 G
- Z 替换 I

资料来源：DuckDuckGo

3.2 Google搜索通过个性化服务造成信息茧房

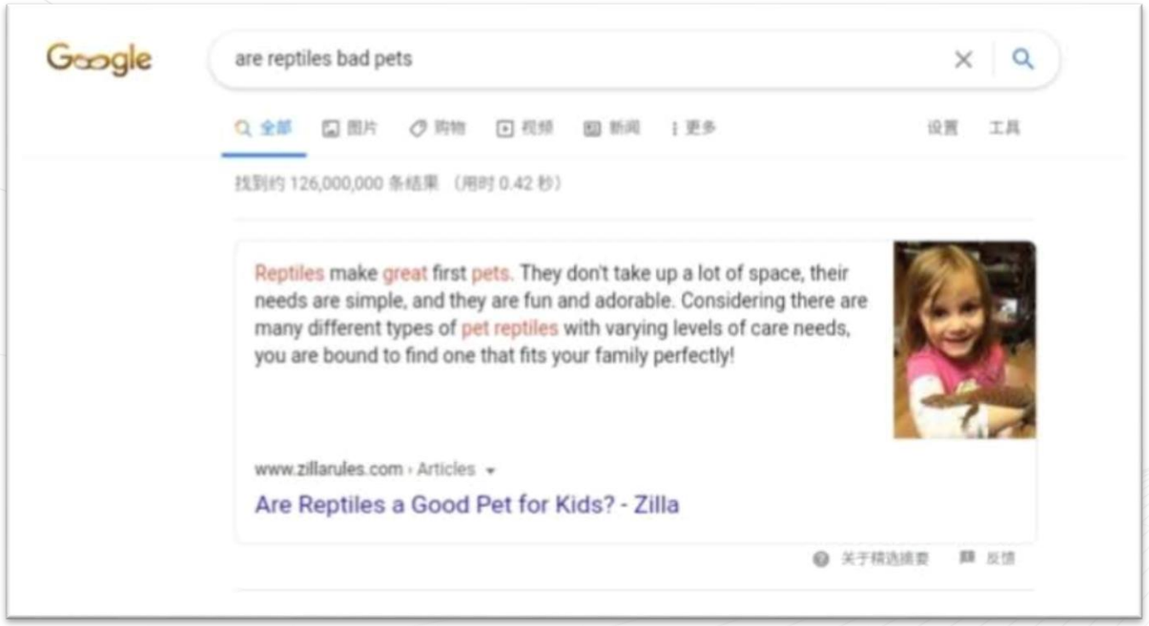
个性化搜索结果不只是基于传统的排名因素（如网页与搜索词或其权限的相关性），还基于搜索引擎在给定时间对用户的信息（如位置、搜索历史记录、人口统计或兴趣）。其目的是增加特定用户的结果的相关性。

Google搜索满足大众化统一基础上的个性化，但是Google搜索明显的缺点是Google搜索会增强用户的预见性偏见，这是造成信息茧房原因之一。

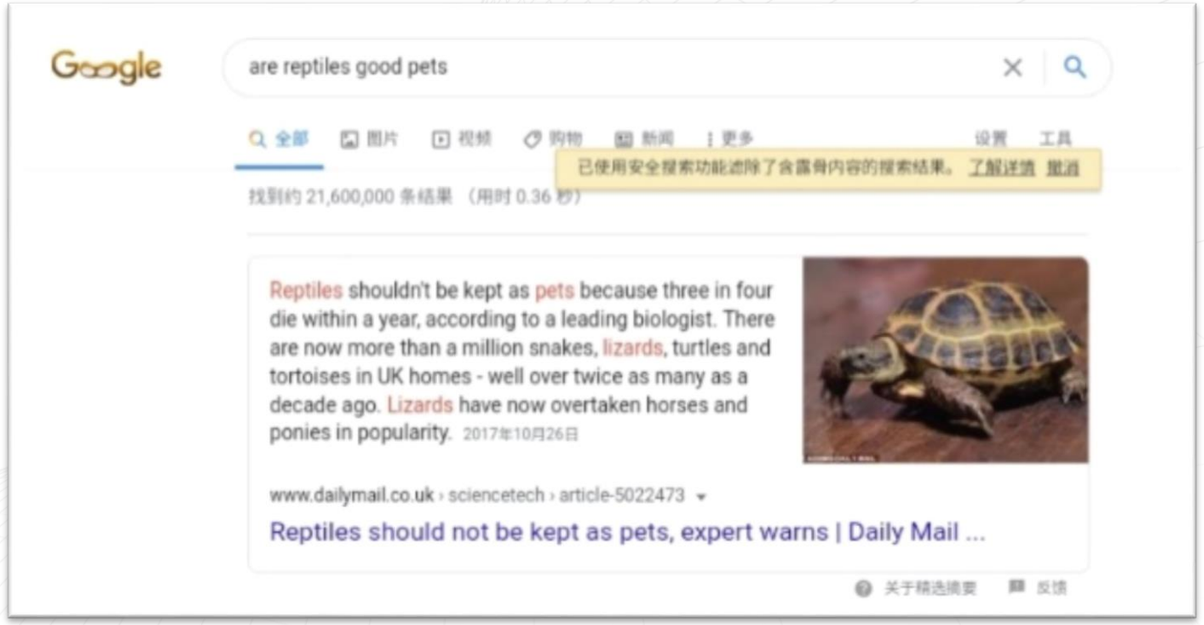
早在2011年，实验显示，超过50%的谷歌搜索结果是个性化搜索，此后这个数字可能上升。

在Google搜索中搜索“提问性语句”时，会带给用户偏见性的结果

搜索“爬行动物是坏的宠物吗”时呈现的结果



搜索“爬行动物是好的宠物”时呈现的结果



资料：MIT Technology Review

在搜索时位置、历史记录等会影响其个性化结果

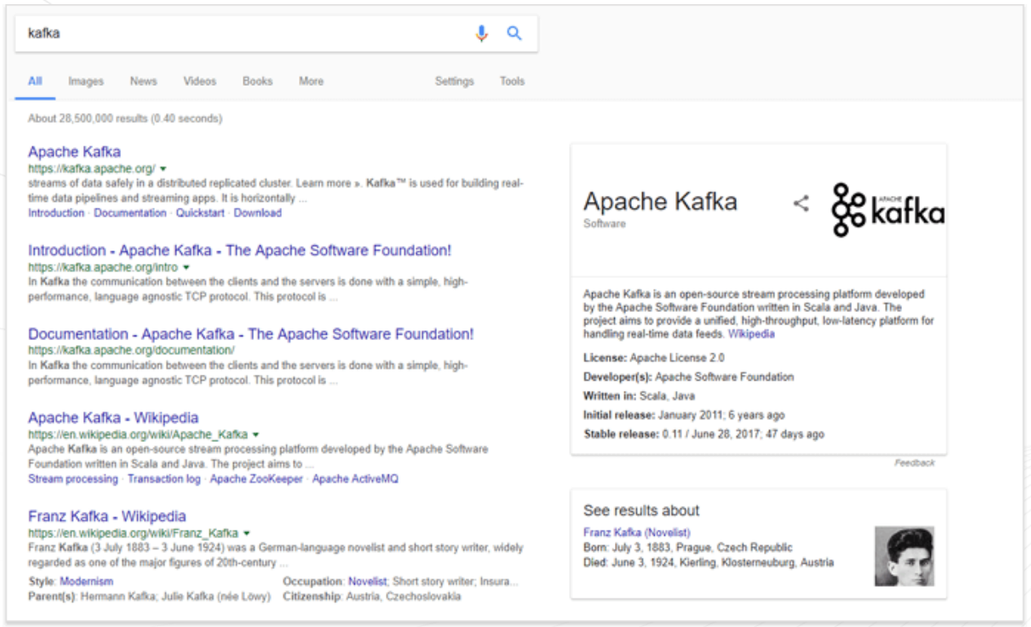
1、用户地理位置：在大多数情况下，位置数据会影响暗示用户正在寻找物理位置的搜索，。同一城市内的搜索者，彼此只有几英里也会看到同一搜索词的不同结果，尤其是在本地包中。

2、搜索和浏览历史记录：Google会根据每个搜索者浏览历史记录、搜索历史记录和 SERP 点击次数为每位搜索者创建个性化个人资料，然后根据用户的兴趣更改用户查看的搜索结果。

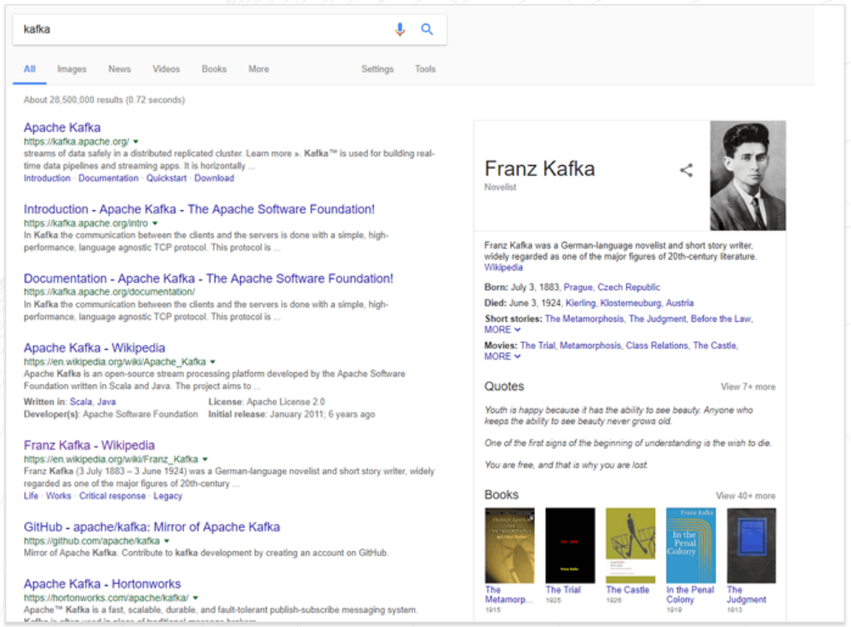
搜索和浏览历史记录：

前后两次搜索“kafka”时，结果变化并不大，但展示面板确实发生了改变。谷歌的搜索个性化算法已经发现，用户更感兴趣的小说家，而不是数据软件服务

第一步：在一个全新的浏览器中， 我做了一个搜索 “kafka” ， 并得到结果如下图结果：



第二步：点击了一些关于小说家弗朗茨·卡夫卡的结果。之后，再次搜索“卡夫卡”，结果如下图：



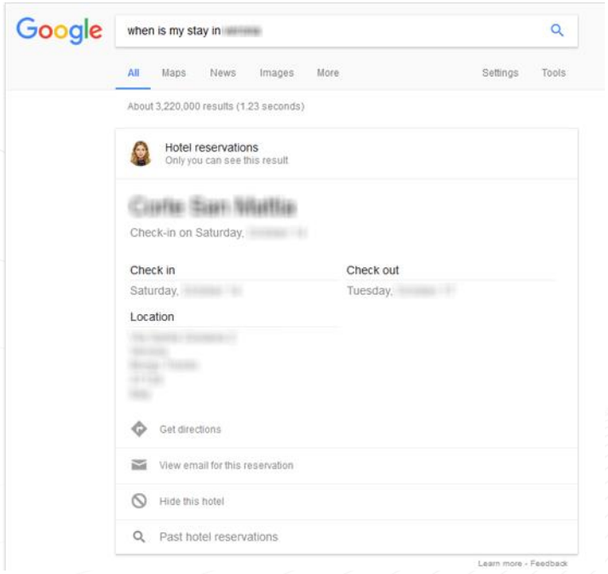
Google账号、使用的设备和Google产品使用情况影响其个性化结果

3、Google账号：Google 可能会将用户的连接中社交媒体帖子添加到搜索结果中，将用户的连接所认可的结果推送到搜索结果的更高版本，并巧妙地将用户获取的搜索结果定制到用户的个人资料中。如果用户有G+个人资料，并且在浏览器中登录了 Google 帐户，则看到的搜索结果可能会显著显示用户的朋友在第1页搜索结果中共享的内容。

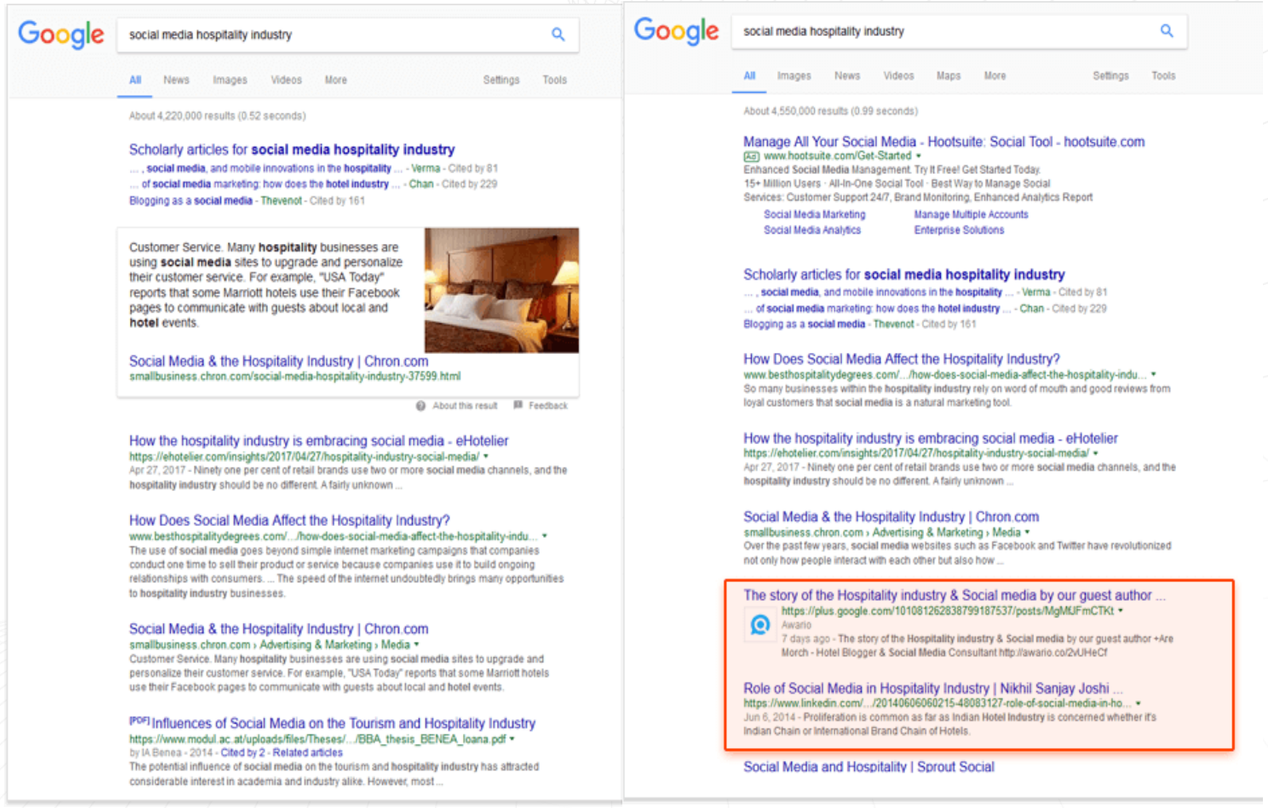
4、用户搜索时所用的设备类型：在移动和桌面设备上，搜索页面的排名不同，搜索查询的解释也不同。移动版本的Google搜索引擎已经变成了单独的搜索引擎，通过不同的因素来排名网页。

5、用户使用的其他Google产品：Google会根据每个搜索者浏览历史记录、搜索历史记录和搜索结果点击次数为每位搜索者创建个性化个人资料，然后根据用户的兴趣更改用户查看的搜索结果。

Google产品：Google通过其多样化产品矩阵尽可能未用户提供更加精准和智能的个性化搜索服务，即“Google个人搜索”。当搜索“我何时起飞”时，结果展示：航班起飞时间、入住酒店名称和地点，以及入住和离开时间。



Google账号：在未登陆Google账号时的搜索结果与登陆Google账号时的搜索结果对比发现：在登陆Google账号状态时，搜索结果中出现了用户G+和LinkedIn上的信息



3.3 算法发展历程：Pagerank到基于神经网络的精准化搜索

Google搜索算法经历了三个阶段：第一阶段（2007-2017）：基于用户标签提供个性化新闻服务，称其为PageRank；第二阶段（2017-2018）：结合用户实际搜索体验，优化其算法，并提供更符合用户偏好和需求的专业、权威并具的个性化服务；第三阶段（2018年至今）：在干预并控制极端及敏感问题上的搜索服务的基础上（对极端、敏感社会问题及事件相关的搜索拒绝提供服务），搜索服务更加智能化、个性化与精准化。

2007-2017年

Google新闻推出个性化服务PageRanK

Google突出显示热门新闻。在置顶显示后，Google会根据用户兴趣以及点击过的文章，会向用户推荐极具个性化的、只与用户所处地相关且与个人相关的信息，Google将其简称为PageRank（用Google创始人Larry Page的名字）。Google认为新闻消费将非常个人化，非常有针对性，此外个性化新闻会伴随着个性化的广告。

2017-2018年

Google搜索结果兼具专业、权威性和个性化

调整算法时的具体实验操作流程;

1、一个团队会与一小部分实际用户一起对建议的调整进行测试，以了解他们是如何与野生环境以及一组称为“搜索质量评估者”的承包商进行互动的。

2、Google在全球范围内约有10,000个评估员，他们对搜索结果的意见可以帮助Google搜索团队评估是否应进行一次特定的调整。评估者通常会同时看到新旧结果，并确定哪个更好。

3、“更好”不是纯粹的主观用语。它由搜索质量评估者指南的已发布文档定义，该文档描述了评估者应如何判断显示在其结果中的页面。特别注意页面的专业知识，权威性和可信赖性。

2018年至今

Google搜索全方位智能化、个性化

7月份，Google对指南进行了一些重大更改，其中除其他外，要求评估者考虑页面作者的声誉。结果，没有明确作者的页面现在可能会被评为较低质量。

2019年推出了一种基于神经网络的“自然语言”处理技术BERT，禁止存在争议的话题方面，如宗教、种族等敏感话题相关的提问上提供搜索服务，以免强化族裔极化。

资料来源：Google官网

3.4 Google针对信息茧房采取了一些措施，但因信息茧房面临着风险

Google不断优化的搜索算法以及管理措施使得Google得以避免类似Facebook、Twitter和YouTube因创建信息茧房而受到的批评。此外，YouTube和Google虽都在Alphabet旗下，但是Google和YouTube算法独立，由完全分开的两个团队来创建和运营。

措施

- 在重大事件发生时，首先推荐经人工审核编辑过的权威信息；其他信息还是采用个性化推荐。
- Google拥有明确的数据来指导优化其算法：
  - 多数时候Google会听取用户的反馈；
  - 其他时候，算法更改的想法来自公司指令或优先事项，如针对敏感社会问题或者明显的事实错误等。

风险

- 高度个性化可能导致人们对Google失去信任。尽管Google并未对大多数搜索排名进行个性化设置，但由于其收集的数据范围非常广泛，其广告还是非常个性化的。
- Google搜索时明显的错误仍然会出现，有时是算法问题，有时是因为有关社会问题搜索结果存在倾向性和偏见。

资料来源：Google官网、CNBC

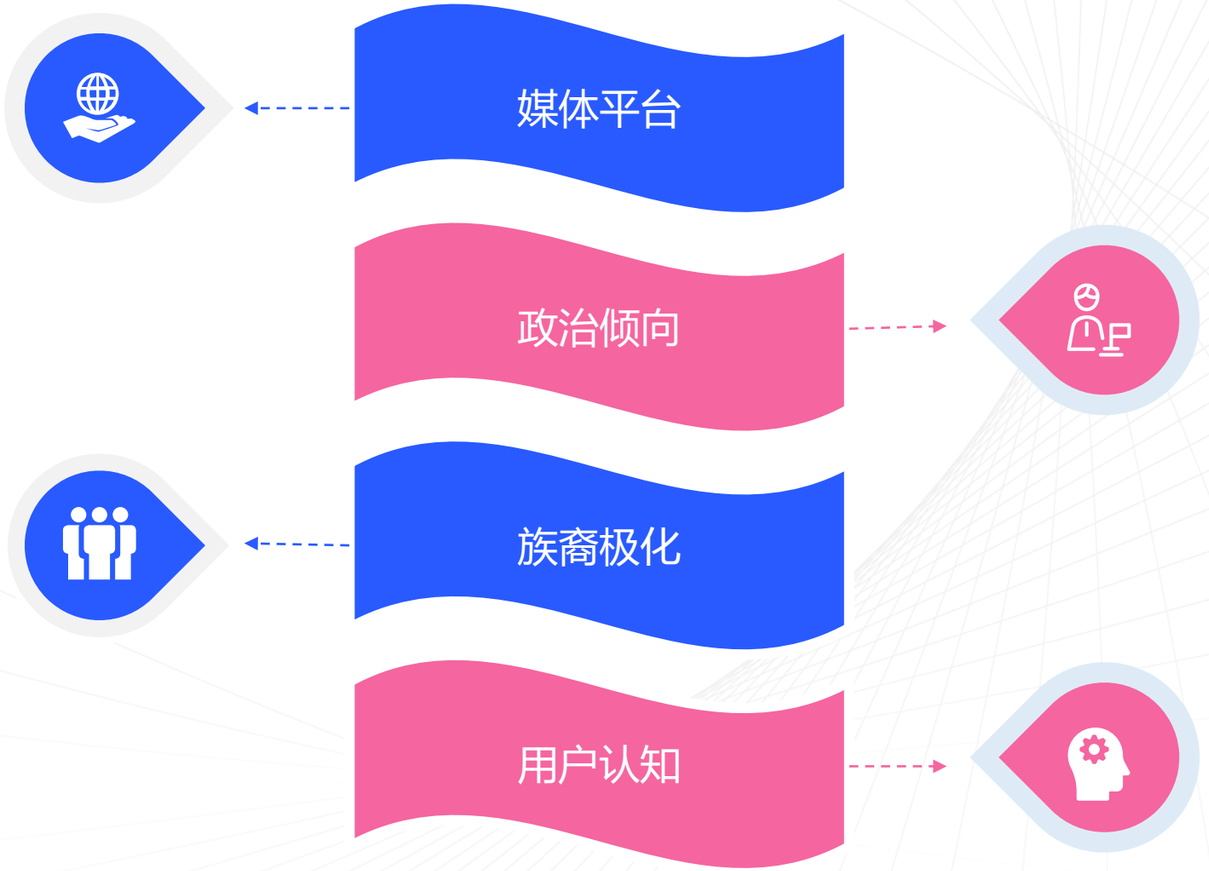
4.0

总结

4.1 宏观因素的长期客观存在性致使信息茧房长期存在

西方国家政治倾向引起的“站队”现象、长期以来的族裔极化等问题都是导致信息茧房形成的重要因素，另外媒体平台以及用户认知偏好等因素也是信息茧房形成的重要驱动因素。

社交媒体和个性化推荐信息平台构建的信息茧房没有建立“地球村”，建立了一个个高度极化、封闭的社区，反而加速了人群的分散，破坏民主形成了在政治上两极化、意识形态上多极化的社会。



2019年Facebook上与36%的新闻与政治相关。在1980年代至1994年之间出生的年轻人在政治上比GenXers或婴儿潮一代更加两极分化。在过去的几十年中，被认为是“极端保守”或“极端自由”的人数也有所增加，中间派人数减少。

2018年调查显示，美国民众科学素质达标率为28%，但是在不同种族之间呈现出明显差别，白人明显高于拉美裔和非洲裔。
在不同群体形成的网络社区（信息茧房）中引入真实的信息来纠正或“揭穿”虚假信息时，它要么被忽略，要么增强了用户的错误信念，从而进一步强化信息茧房。

用户根据其喜好分成几小类，并投其所好接受同质化信息，造成用户认知窄化，长此以往导致用户认知缺失。
用户认知缺失会在信息获取并消费过程中促成并强化信息茧房。

4.2 信息茧房长期存在，平台算法的优化只是平衡信息茧房因与果

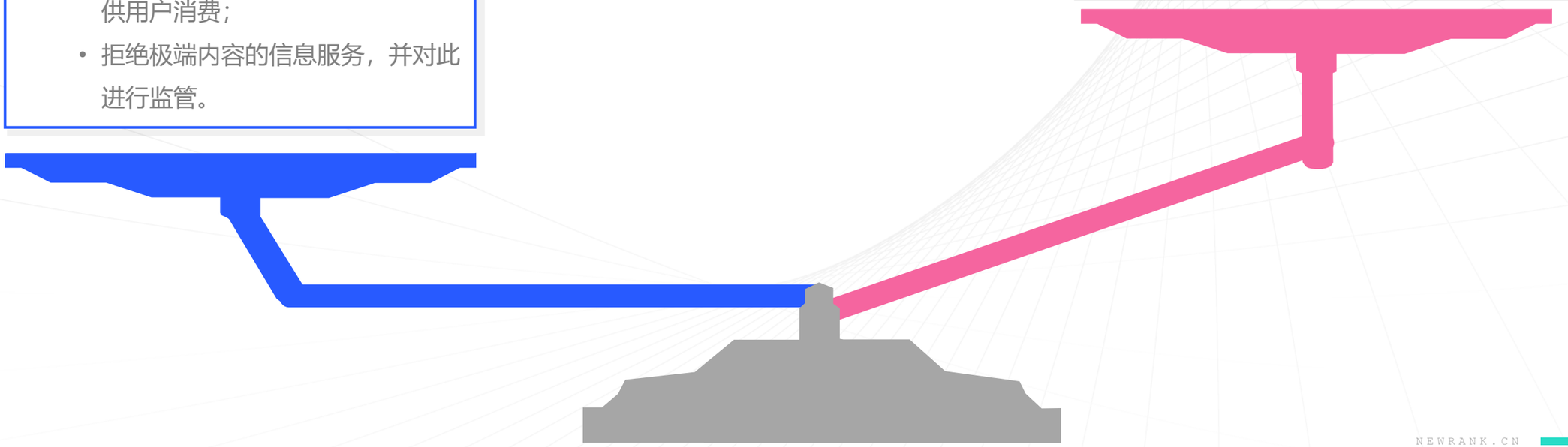
造成信息茧房形成的因素众多，包括信息消费侧——用户，以及信息供给侧——媒体与平台，此外信息茧房的结果会体现在用户身上。

综合众多因素考虑，由于用户数量众多且对其调整优化难度大；因此，打破信息茧房的尝试集中信息供给侧——媒体与平台，尤其集中于平台算法机制方面。

信息茧房将长期存在，且不可消灭；当信息茧房结果明显或者严重时通常媒体和平台采取相应的措施，以减轻信息茧房，在其“因”与“果”间达到一定平衡。

- **通常采取的措施：**
  - 在信息生产方面：筛选提供内容的媒体，及优化内容生产者的管理；
  - 在分发方面：优化平台算法，在个性化服务的基础上提供多样化信息供用户消费；
  - 拒绝极端内容的信息服务，并对此进行监管。

- **信息茧房造成的结果：**
  - 族群极化
  - 个人认知窄化
  - 政治派别出现极化影响社会稳定
  - 虚假信息的泛滥及个人媒介素养下降



THANKS!

新浪新闻 Sina News

新榜研究院 INSTITUTE OF NEWRANK

撰稿人：努尔麦麦提·买合木提、郑志垚
组稿人：张翔 夏维兰