网易人工智能 网易云 163yun.com IT大咖说 Boxue 博学教育

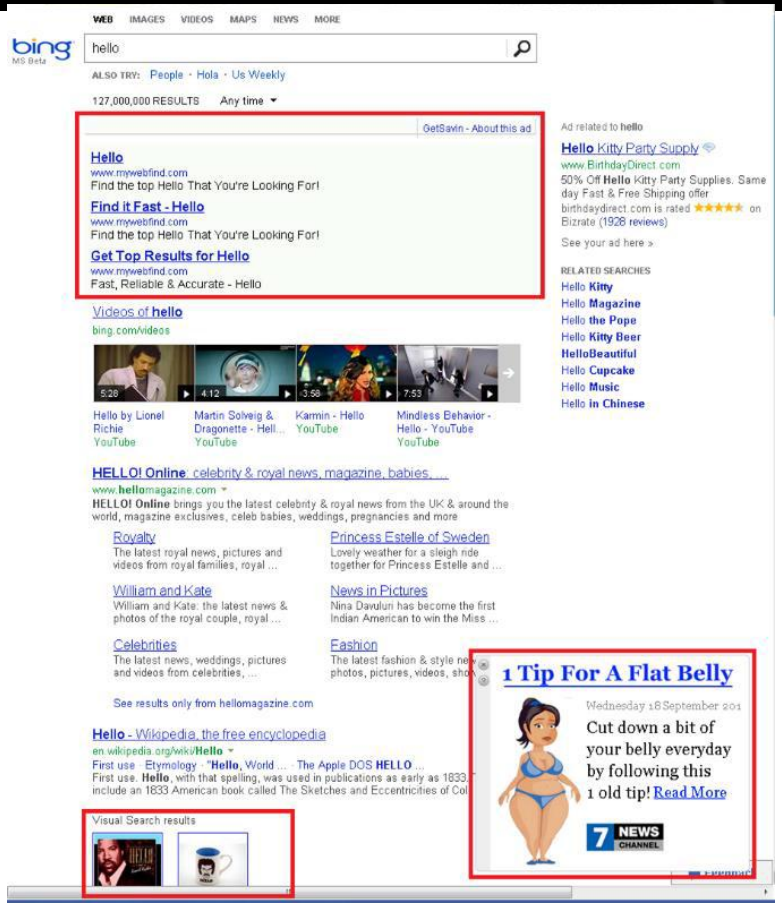
网易博学实践日
大数据和人工智能技术大会

2017.08.12

大规模线上实验与机器学习

京东 推荐应用科学部 熊熹

JD.京东 JD.COM

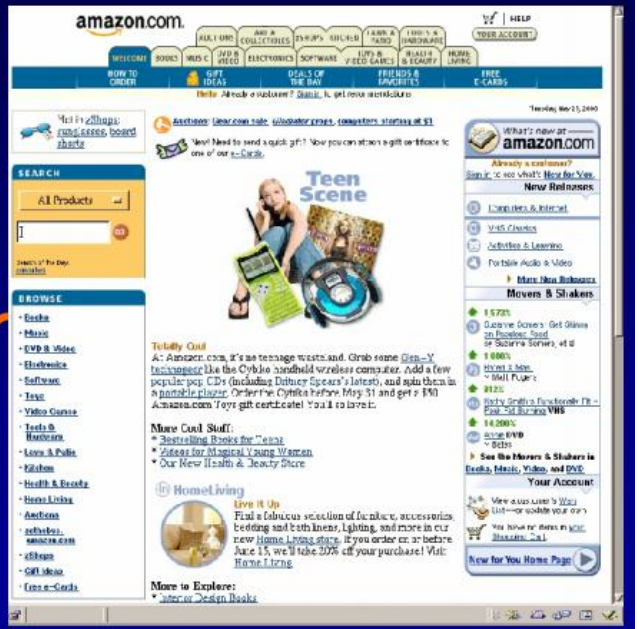


页面越简洁越好？

Bing Search advertising revenue grew 47%

页面越简洁越好？

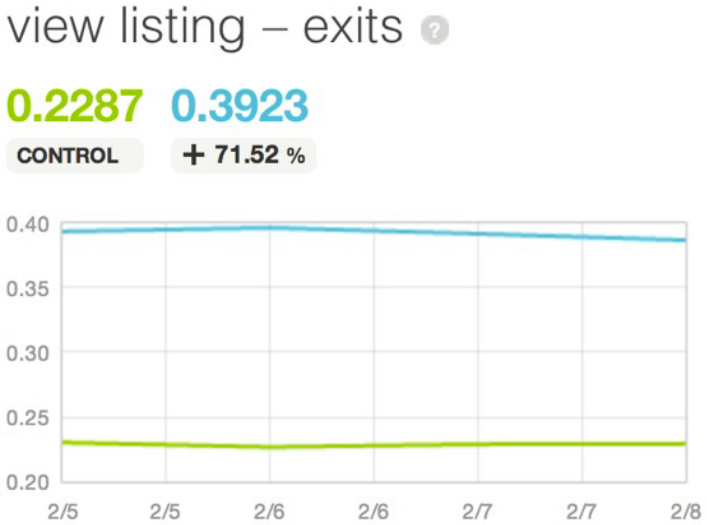
- Simpler is better, right?



- No. Orders significantly down

2008年，MSN在Hotmail上实验了点击链接新开一个tab功能，邮箱人均打开量提升8.9%
2011年，MSN在其搜索页面重复这个实验，同样带来了5%的提升，这是MSN历史上最成功的实验！

In esty



新开一个页面好不好？

In esty search

无限下拉页面？

In jd 推荐

购物车推荐改无线下来模式，推荐引入订单和gmv带来50%+的提升

无限下拉页面？

- 同样的策略，在A成功，不代表在B成功
- 同样的策略，以前成功，不代表现在成功（反之亦然）
- 有的时候我们甚至不知道为什么会成功或者失败
- 相关≠因果，A/B实验是最好的因果检验工具
- Hippos(HIghest Person’s oPinionS) are dangerous!
- AB test everything if possible!

实验设计

- 实验对象
- 实验因素
- 实验效应

AB 测试

- 分流
- 单因素
- 指标（metrics）

A/B实验核心三要素——分流

- 随机分流
- 通过AA实验验证多层实验的正交性，并计算最少实验流量

➢ Given the skewness of the metric, defined as $s = \frac{E[X-E(X)]^3}{[Var(X)]^{3/2}}$

a lower bound on the number of samples is $355 \times s^2$.

A/B实验核心三要素——分流

- 均等原则

辛普森谬论——女生比男生更容易被录取？

法学院

性别	录取	拒收	总数	录取比例
男生	8	45	53	15.1%
女生	51	101	152	33.6%
合计	59	146	205	

商学院

性别	录取	拒收	总数	录取比例
男生	201	50	251	80.1%
女生	92	9	101	91.1%
合计	293	59	352	

两学院的数据汇总：

性别	录取	拒收	总数	录取比例
男生	209	95	304	68.8%
女生	143	110	253	56.5%
合计	352	205	557	

- 均等原则

进行AB测试的样本量尽量均等，在逐步放量过程中避免大小对比

辛普森谬论——在京东也一样出现！

有一个算法优化实验，测试实验A 5%流量，对比实验B线上大流量（约95%），实验对比时间是2016.12.10~2016.12.20

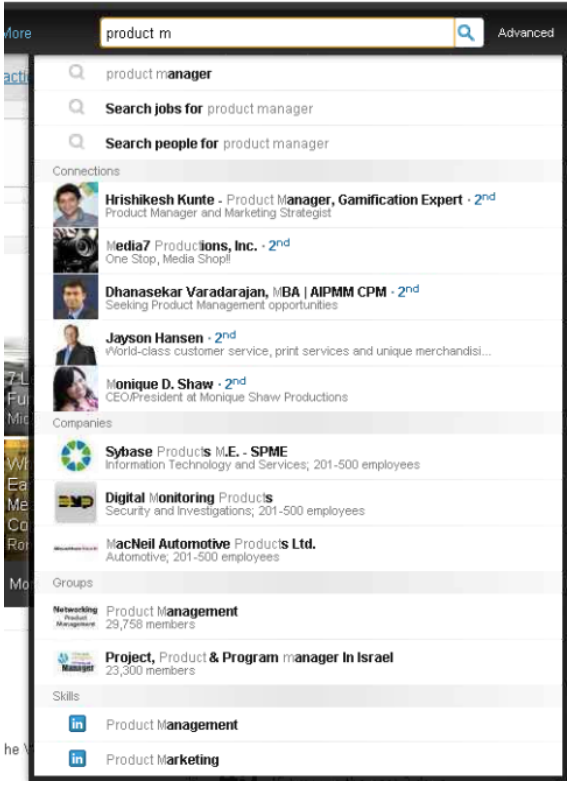
逐步放量的过程中，A每一天都比B好，但是平均意义上比B差！

Why and how?

A/B实验核心三要素——单因素

- 实验因素，越简单越好
- 永远一次只测试一个因素
- 一旦确认一个因素的正面影响

马上开始下一个



A/B实验核心三要素——指标（metrics）

If you can't measure it, you can't improve it. - *Peter Drucker*

- 引入假设检验，计算置信区间
- 区分随机因素和真实影响
- 重大提升出现的概率不到1/500（data by microsoft），而且在一个成熟系统中，基本上都被做过了；大多数的增长都是缓慢的积累，因此方向比努力更重要！

Define metrics for metrics

- Mandatory quality 1：directionality（反例：Distinct Queries per Unique User）
- Mandatory quality 2：sensitivity（反例：PV/UV, request per users；客单价）

- 指标稳定性——AA测试

Metrics of metrics

Sensitivity(可以是确认正向实验的假设检验P值)-Penalty of direction(AA实验的不稳定性）

能算的指标，越多越好

多个指标之间的平衡

- 业界的做法

其他指标没有下降，或者下降不显著的情况下，另一些指标上升，即可认为是一次成功的实验

不允许降低一个指标来提升另一个

- 学界的做法

使用机器学习的方法，学出指标之间的最佳组合，以进行最好的区分（linkedin, google,yandex和Microsoft都在这个领域发过很多篇论文）

观察指标的一点小建议

- *Define*
  - α is the statistical significant level = 0.05
  - β is the type-II error level = 0.2 (80% power)
  - π is the probability that the alternative hypothesis is true, i.e., the experiment is moving metrics
  - TP is True Positive, and SS is a Stat-sig result, then we have Bayes Rule:

$$P(TP|SS) = P(SS|TP) * \frac{P(TP)}{P(SS)} = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}$$

- If we have a prior probability of success of $\pi = 1/3$, which is what we reported is the average across multiple experiments at Microsoft, then the posterior probability for a true positive result given a statistically significant experiment is 89%.
- However, if the probability of success is one in 500, then the posterior probability drops to 3.1%.

Following Tail Lights 别人做过的成功实验，再做一次成功率高

我们成功的实验

- 个性化 VS非个性化 （曾经给我们的搜索引流带来X%+的提升）
- 机器学习VS 规则 （曾经给我们的购物车推荐带来Y%+的提升）
- 深度学习VS 机器学习 （曾经给我们的广告收入带来Z%+的提升）
- 强化学习 （正在给我们的首页用户粘性带来不可估量的提升）
- 更深的数据挖掘，更细致的模型，更广的覆盖，更实时的响应,继续给我们的持续增长带来无限可能

机器学习实践要点

- Y=f(X)
- Y: 深入理解你的metrics ； 调整正负（中）样本权重很重要；加入人工标注
- f:
- X： 特征工程;feature log;
- 持续优化！

算法模型之外的思考

- 响应速度

In a Web 2.0 talk by Marissa Mayer, then at Google, she described an experiment where Google increased the number of search results on the **SERP from ten to thirty** Traffic and revenue from Google searchers in the experimental group **dropped by 20%.** Her explanation? The page took half a second more to generate.

在硅谷，一个工程师可以把核心服务响应速度降低10ms，他这一年的工资都可以躺拿

- 可解释性
  - **debuggability**
  - 业务规则
  - 提升/降低可解释

最后的一点建议

- 一旦有了点子，越早开始试验越好
- 一次只改动一个地方
- 线上线下一致性非常重要
- 如果一个实验一开始就不奏效，除非你查到实现bug，继续等待下去，它一定也会不奏效
- 性能监控越实时越好，产品数据分析t+1天绝对是必要的

AB测试之外

- Online: interleaving, variance reduction,幽灵实验
- Offline： 算法评测指标 ;simulation;人工评测;golden set

CAPITAL OF STATISTICS
PROFESSION, HUMANITY & INTEGRITY

统计之都（Capital of Statistics，简称COS）成立于2006年5月，是一个旨在推广与应用统计学知识的网站和社区。统计之都网站最初由谢益辉创办，现任理事会主席为冯凌秉，现由世界各地的众多志愿者共同管理维护。

我们的口号是：

中国统计学门户网站，免费统计学服务平台

- 主站+论坛
- R会议，数据沙龙，线上沙龙

Email：contact@cos.name
新浪微博：@统计之都
人人网：@统计之都
Twitter：@cos_name
微信公众号：CapStat


THANKS