搜索引擎的评价指标

王树森

http://wangshusen.github.io/

搜索引擎的评价指标

- 北极星指标：
  - 用户规模、留存率。
  - 单个策略不容易提升规模和留存。
- 中间指标：
  - 用户的点击等行为，反映搜索质量的好坏。
  - 做 A/B 测试，中间指标很容易显著。
- 人工体验评估：
  - 人工评估搜索体验，考察 GSB、DCG 等指标。

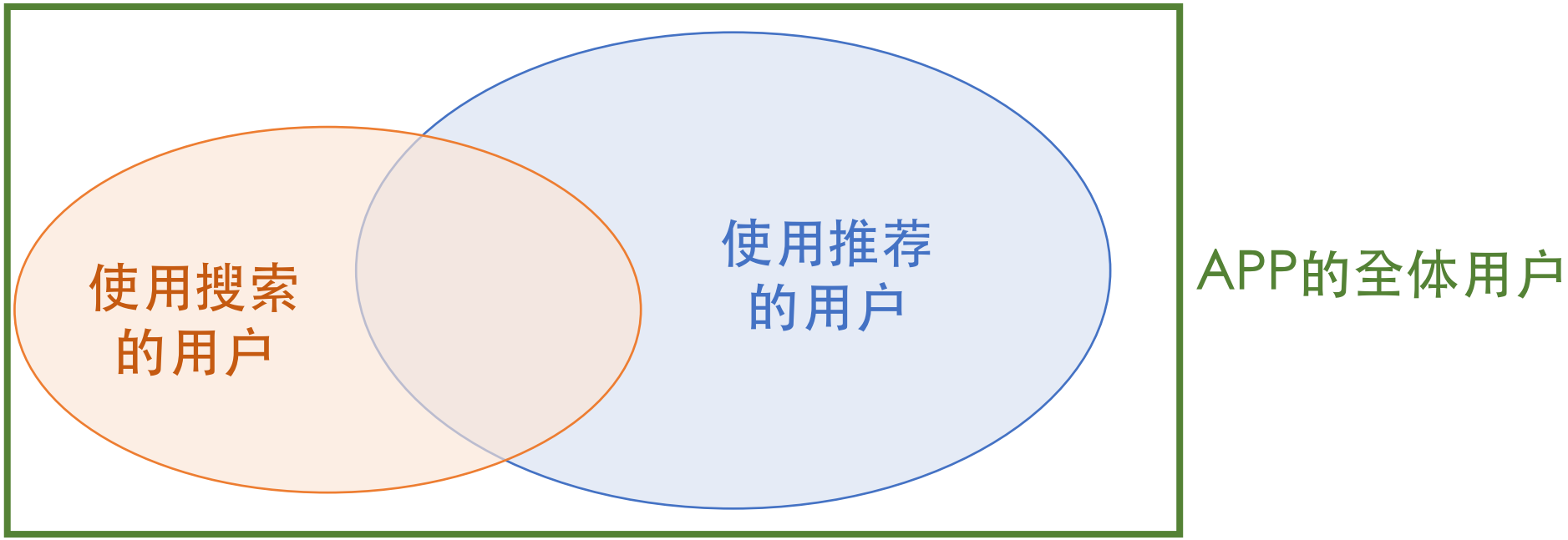
北极星指标：用户规模 & 留存

用户规模

- 日活用户数（Daily Active User，DAU）。

用户规模

- 日活用户数（Daily Active User，DAU）。
- 搜索日活（Search DAU），推荐日活（Feed DAU）。



用户规模

- 日活用户数（Daily Active User，DAU）。
- 搜索日活（Search DAU），推荐日活（Feed DAU）。
- 搜索渗透率 = Search DAU / DAU。搜索体验越好，用户越喜欢用搜索功能，则搜索渗透率越高。
- 提升搜索日活、搜索渗透率的方法：
  - 搜索的体验优化，可以提升搜索留存，从而提升搜索日活。
  - 产品设计的改动，从推荐等渠道向搜索导流，提升搜索渗透率，从而提升搜索日活。

用户留存

APP 的次 7 日内留存（次 7 留）

- Feb 1 有 1 亿用户使用 APP。
- 这 1 亿人中，有 8 千万在 Feb 2~8 使用 APP 至少一次。
- Feb 1 的次 7 留 = 8 千万 / 1 亿 = 80%。

用户留存

- 常用的留存指标：次 1 留、次 7 留、次 30 留。
- 次 n 留随 n 单调递增：

$$\text{次 1 留} \leq \text{次 7 留} \leq \text{次 30 留}$$

- APP 次 n 留、搜索次 n 留、推荐次 n 留。
- 现在更流行 LT7 和 LT30 留存指标。

中间指标：点击等行为

点击率 & 有点比

- 文档点击率
  - 搜索结果页上，文档被用户看到，算作曝光。
  - 文档点击率 = 总点击数 / 总曝光数。
- 有点比（查询词点击率）
  - 搜索结果页上，用户点击任意一篇文档，则本次搜索算有点击。
  - 有点比 = 有点击的搜索次数 / 总搜索次数。
- 首屏有点比
  - 点击发生在首屏，本次搜索算有点击。
  - 首屏有点比 ≤ 有点比。

首点位置

- 平均首点位置：
  - 一次搜索之后，记录第一次点击发生的位置。
  - 如果没有点击，或者首点位置大于阈值 x，则首点位置取 x。
  - 对所有搜索的首点位置取平均。
- 平均首点位置小，说明符合用户需求的文档排名靠前，用户体验好。
- 优化搜索排序，通常会同时改善有点比、首屏有点比、平均首点位置。三者与留存指标强相关。

主动换词率

- 如果用户搜到需要的文档，通常不会换查询词。
  - 例：女性用户搜“机械键盘”，结果大多是黑色的，不符合用户喜好（个性化差）。用户会换词为“机械键盘 女性”。
  - 例：搜“权利的游戏”，搜索引擎没能自动纠错，搜到的文档很少、质量不好。用户会换词为“权力的游戏”。

主动换词率

- 如果用户搜到需要的文档，通常不会换查询词。
- 一定时间间隔内，搜的两个查询词相似（比如编辑距离小），则被认定为换词。
- 主动换词 vs 被动换词
  - 被动换词，比如搜索建议“您是不是想搜 权力的游戏”，用户点击建议。
  - 主动换词，原因是没有找到满意的结果，说明搜索结果不好。

交互指标

- 用户点击文档进入详情页，可能会点赞、收藏、转发、关注、评论。
- 交互通常表明用户对文档非常感兴趣（强度大于点击），因此可以作为中间指标（类似于有点比、首点位置、换词率）。
- 交互行为稀疏（每百次点击，只有 10 次点赞、2 次收藏），单个交互率波动很大，而且在 A/B 测试中不容易显著。
- 取各种交互率的加权和作为总体交互指标，权重取决于交互率与留存的关联强弱。

中间目标 → 留存目标

- 体验优化的策略往往同时改善多种中间指标：有点比、首屏有点比、平均首点位置、主动换词率、交互指标。
- 单个体验优化的策略很难在短期内显著提升留存指标。（通常微弱上涨，不具有统计显著性。）
- 上述中间指标与留存有很强的关联。长期持续改善中间指标，留存指标会稳定上涨。

人工体验评估

Side by Side 评估

- 随机抽一批搜索日志，取其中查询词、用户画像、场景。运行新旧两种策略，得到两个搜索结果页（文档列表）。
  - 固定查询词、用户、场景，搜索结果的差异只来自于策略。
  - 随机抽样搜索日志时，需要覆盖高频、中频、低频查询词。

Side by Side 评估

- 随机抽一批搜索日志，取其中查询词、用户画像、场景。运行新旧两种策略，得到两个搜索结果页（文档列表）。
- 对于一条查询词，人工评估两个列表，分别对应新旧两种策略。
  - 基于查询词、用户画像、搜索场景，判断左右两个列表谁更好。
  - 盲评，即新策略出现在左、右的概率都是 50%。
  - 不是判断具体哪篇文档更好，而是判断哪个列表整体更好。

Side by Side 评估

- 随机抽一批搜索日志，取其中查询词、用户画像、场景。运行新旧两种策略，得到两个搜索结果页（文档列表）。
- 对于一条查询词，人工评估两个列表，分别对应新旧两种策略。
- 使用 GSB 作为评价指标。
  - 如果新策略更优，记作 Good (G)。
  - 如果两者持平，记作 Same (S)。
  - 如果旧策略更优，记作 Bad (B)。
  - 例：评 300 条查询词，GSB 为 50: 220: 30。

Side by Side 评估

Session 信息： 查询词、用户画像、场景信息

哪边更优： 左 相同 右

排名	左	右	差异
1	文档	文档	无差异
2	文档	文档	排序上升2位
3	文档	文档	有差异
⋮	⋮	⋮	⋮

Side by Side 评估

Session 信息： 查询词、用户画像、场景信息

哪边更优： 左 相同 右

排名	左	右	差异
1	文档	文档	无差异
2	文档	文档	排序上升2位
3	文档	文档	有差异
⋮	⋮	⋮	⋮

Side by Side 评估

Session 信息： 查询词、用户画像、场景信息

哪边更优： 左 相同 右

排名	左	右	差异
1	文档	文档	无差异
2	文档	文档	排序上升2位
3	文档	文档	有差异
⋮	⋮	⋮	⋮

月度评估

- 每个月随机抽取一批搜索日志，每条搜索日志包含查询词 q、用户 u、场景 c、排名前 k 的文档 $d_1, \cdots, d_k$。
  - 随机抽样搜索日志时，需要覆盖高频、中频、低频查询词。
  - 文档数量 k 取决于平均下滑深度，比如 $k = 20$。

月度评估

- 每个月随机抽取一批搜索日志，每条搜索日志包含查询词 q、用户 u、场景 c、排名前 k 的文档 $d_1, \cdots, d_k$。
- 标注员评估每一篇文档，打分 $\text{score}(q, u, c, d_i)$。
  - 可以单独给相关性、内容质量、或时效性打分。
  - 可以只打一个综合满意度分数。

月度评估

- 每个月随机抽取一批搜索日志，每条搜索日志包含查询词 q、用户 u、场景 c、排名前 k 的文档 $d_1, \cdots, d_k$。
- 标注员评估每一篇文档，打分 $\text{score}(q, u, c, d_i)$。
- 用 DCG 评价一次搜索 $(q, u, c, d_1, \cdots, d_k)$ 结果的好坏：

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{\text{score}(q,u,c,d_i)}{\log_2(i+1)}.$$

月度评估

- 每个月随机抽取一批搜索日志，每条搜索日志包含查询词 q、用户 u、场景 c、排名前 k 的文档 $d_1, \cdots, d_k$。
- 标注员评估每一篇文档，打分 $\text{score}(q, u, c, d_i)$。
- 用 DCG 评价一次搜索 $(q, u, c, d_1, \cdots, d_k)$ 结果的好坏：

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{\text{score}(q,u,c,d_i)}{\log_2(i+1)}.$$

月度评估

- 每个月随机抽取一批搜索日志，每条搜索日志包含查询词 q、用户 u、场景 c、排名前 k 的文档 $d_1, \cdots, d_k$。
- 标注员评估每一篇文档，打分 $\text{score}(q, u, c, d_i)$。
- 用 DCG 评价一次搜索 $(q, u, c, d_1, \cdots, d_k)$ 结果的好坏：

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{\text{score}(q,u,c,d_i)}{\log_2(i+1)}.$$

- 对所有搜索日志，取 DCG 的均值，作为月度评估的结果。
  - 可以是自我对比，是否优于往期的 DCG。
  - 可以与竞对对比，是否优于竞对的 DCG。

总结

北极星指标：用户规模 & 留存

- 用户规模：
  - APP 总体的 DAU、搜索的 DAU。
  - 搜索渗透率（Search DAU / APP DAU）。

北极星指标：用户规模 & 留存

- 用户规模。
- 用户留存：
  - 次 1 留、次 7 留、次 30 留。
  - LT7、LT30。
  - APP 总体的留存、搜索自身的留存。

北极星指标：用户规模 & 留存

- 用户规模。
- 用户留存。
- 规模和留存指标未必适合评价单个策略。
  - 单个策略很难显著提升规模和留存。
  - 规模和留存指标需要很长时间才能显著。

北极星指标：用户规模 & 留存

- 用户规模。
- 用户留存。
- 规模和留存指标未必适合评价单个策略。
- 规模和留存指标更适合作为大盘长期指标观察。
  - 评估整个团队长期的表现（所有策略叠加）。
  - 长期优化搜索体验，规模和留存会稳定提升，反映在 A/B 测试的 holdout 上。

中间指标：用户的点击等行为

- **中间指标**：与规模和留存强关联，且容易在 A/B 测试中显著。
- **有点比**：是否找到至少一篇用户需要的文档。
- **首屏有点比**：是否把用户需要的文档排在首屏。
- **首点位置**：用户需要的文档排名是否靠前。
- **主动换词率**：没搜到用户需要的文档，用户会换词重搜。
- **交互率**：文档是用户非常需要的，那么用户会点赞、收藏、转发、关注……

人工体验评估

- Side by side 评估：以 GSB 作为评价指标，对比新旧两种策略，决策新策略是否可以推全。

人工体验评估

- Side by side 评估：以 GSB 作为评价指标，对比新旧两种策略，决策新策略是否可以推全。（争议比较大！）
  - 评估过于主观，评估标准未必与普通用户体验一致。
  - 结果噪声大，稳定性不如 A/B 测试。
  - 速度慢于 A/B 测试，影响开发迭代效率。
  - 人工成本比较贵。
  - 个性化较难处理，仅凭用户画像难以判断用户真实需求。

人工体验评估

- Side by side 评估：以 GSB 作为评价指标，对比新旧两种策略，决策新策略是否可以推全。（争议比较大！）
- 月度评估：以平均 DCG 作为评价指标，与自己往期做对比、与竞对做对比，判断搜索团队整体水平。
- Side by side 和月度评估的区别：
  - 目的不同：前者决策新策略是否推全，后者判断搜索团队整体水平。
  - 指标不同：前者的指标是GSB，后者的指标是 DCG。
  - 有无争议：前者充满争议，后者没有缺点和争议。

思考题

- **问题**：用于评价用户体验，整个搜索引擎的日均搜索次数是好的指标吗？
- **提示**：
  - 是好的商业目标，它与广告收入正相关。
  - 可以用来对比新旧策略的用户体验吗？
  - 日均搜索次数 = DAU × 人均搜索次数。
  - 人均搜索次数与换词率的关系是什么？

Thank You!

http://wangshusen.github.io/