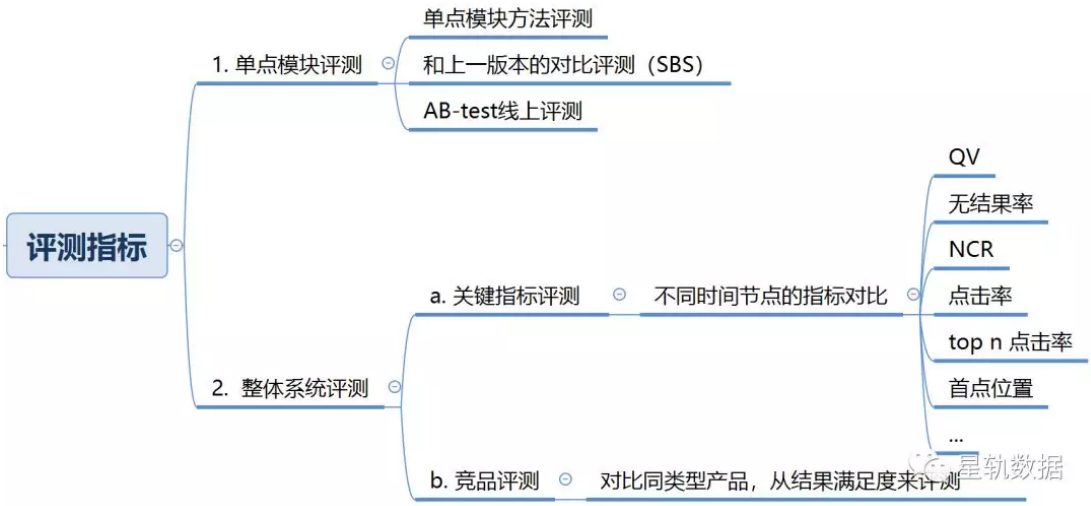
搜索系统的评测方法

原创 XG数据 WePlayData 2019-10-23

搜索系统由多个不同作用的模块组成，多模块相互配合和兜底最终返回用户看到的结果。如何评测搜索系统的好坏是个比较关键的问题。工业界主要从单点模块评测和整体系统评测两个角度分别评测，前者更看中技术指标、后者更看中业务指标。



一、搜索单点优化评测

主要为了判断单个模块优化后，效果提升有多少，可不可以上线？

- **单点模块方法评测**

  每个模块都采用合适的模型方法，此时可通过构建一个标准测试集获取一些模型的指标来评价。比如要取识别query中的实体，可利用在标准测试集上的准确率、召回率、F1来评价。又如做LTR的排序，通常选取MAP和NDCG指标。

- **和上一版本的对比评测**
  - diff率：主要衡量单个模块优化的影响面，比如优化同义词后造成了多少百分比的query和上一版本同义结果不一样；
  - 胜出率：常用的方法称为SBS评测，评测人员对比新旧版本的结果，并标记“好”、“坏”、“一样好”、“一样坏”，胜出率 = （好+（一样好+一样坏）/2）/ALL。一般胜出率大于55%说明影响是正向的。
- **AB-test**

  上面两种评测需要人参与标注或评测，不可避免存在一些主观因素存。AB-test相对是一种比较客观的指标，通过划分不同的流量给新旧版本，然后观察目标指标一定时间（搜索比较看中的是点击率、无点率，首点位置等），最后通过指标的变化来判断优化是正向还是负向。一般先划分小部分流量用于小流量实验，小流量验证有效性后，再全流量上线。

二、系统整体评测

- **关键指标评测**
  根据统计搜索点击日志来生成一些不同时间节点版本的指标对比报表，如QV，无结果率，NCR，点击率，top n点击率，首点位置等。通过观察大盘指标的对比，可以比较清楚掌握业务的发展和问题。
- **竞品评测**
  此外如果有竞品，比如做医疗搜索，可以和是搜狗明医、微信医疗等系统对比。由于不同产品背后的doc资源是不一样的，因此往往从结果满足度来评测，即搜索结果是否满足query的需求。

相关阅读

1. Query理解 - 搜索引擎“更懂你”
2. 搜索引擎新的战场 - 百度、头条、微信
3. 当我们关注舆情系统时，我们忽略了什么?
4. 搜索引擎的两大问题（1） - 召回
5. 搜索引擎的两大问题（2） - 相关性
6. Query词权重方法（1） - 基于语料统计
7. Query词权重方法（2）- 基于点击日志
8. Query词权重方法（3）- 基于有监督学习
9. Query词权重方法（4）- beyond 词粒度
10. Query意图方法（1）- 基于片段意图
11. Query意图方法（2）- 基于文本分类