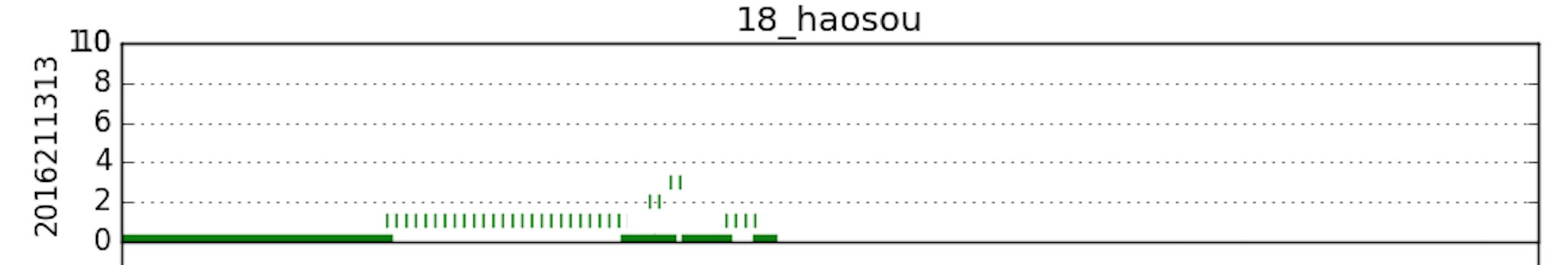
2016年12月10日

本周工作：

首先对于张帆的用户实验中来自不同的文档和不同的源的用户行为进行了可视化分析。如下图所示：



表示用户2016211313上在topic 18（跛脚）上haosou结果页上的行为，横轴表示0至300秒的，纵轴表示用户点击在SERP（y=0，实线）相应Rank的Landing Page（y=r，虚线）上的停留时间，可以看到用户首先在SERP上停留了比较久的时间，然后花费比较久阅读第一条结果，进而快速检验了第二条和第三条结果页，最后回访了第一条结果后结束搜索。

通过对不同用户的行为集中进行分析，可以得到以下几点初步结论：

1. 对于一些通过SERP就可以获得结果（跛脚这个topic为例），用户在SERP上停留的时间是相对比较长的，对于一些需要用户点击后才可以获得信息的话题，用户大量时间都花费在Landing Page上。
2. 用户点击的顺序基本上是和结果排序一致的，回访的行为非常少。

和刘老师讨论后，重新整理了一下评价指标的思路：

对于每一个结果，用户需要检验的结果页面信息的长度L

L=L\_result+p\_r\*L\_lp

其中p\_r=f(query, r) 是用户需要进一步检验结果的概率，通过标注得到，L\_lp是landing page的长度；

衰减函数d=f(accumulate L)，是用户在不同的深度上停止的概率，可以通过张帆的用户实验拟合出来。

这样的指标的贡献有以下两点：

1. 它通过用户浏览的深度，天然地考虑了用户获得information gain的成本；
2. 它考虑到了移动搜索环境下，用户不需要点击就可以获得information gain的情况；

目前需要进行的是针对一些特定的query进行标注，在我的50个query的集合上，这样的结果不是特别多。

其他的工作：

1. WWW准备了中文和英文的queryset，目前的情况是英文的subtask参加的人比较多，中文的还比较少。中文的baseline还在准备中，英文的在indri上统计了各个指标top10，50，100和1000 的平均分数，发现有一些query确实太差，计划丢弃。现在导航类的查询大概占到30%左右。
2. 参加了NTCIR 2nd TO Meeting；
3. 帮助熊辰炎抓取网页；
4. 学习Deep learning book；
5. 和搜狗同事联络SogouT技术细节，得到回复如下：

抽取方法：

1、  数据源：将索引数据过滤低质、外文、违法数据，其中外文只能初步过滤

2、  文本处理：提取html源码转为utf8编码，目前只支持ASCII，GBK，GB2312，UTF-8，ISO-8895-1

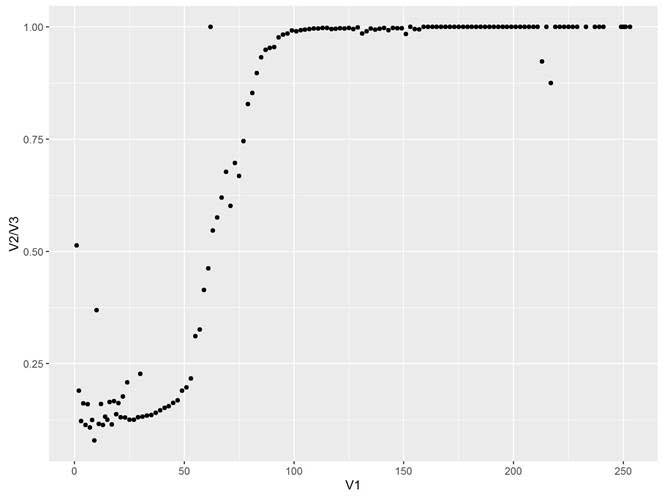
3、  随机抽样：影响一条url是否被抽取到的因素有pr（pagerank）和同站点已抽取url数量。pr是正向影响，同站点抽取数量是逆向影响，然后每次随机抽一批，不够10亿条从剩下的url中继续抽取，直到达到10亿条为止。

4、  格式整理，需要的其他数据merge

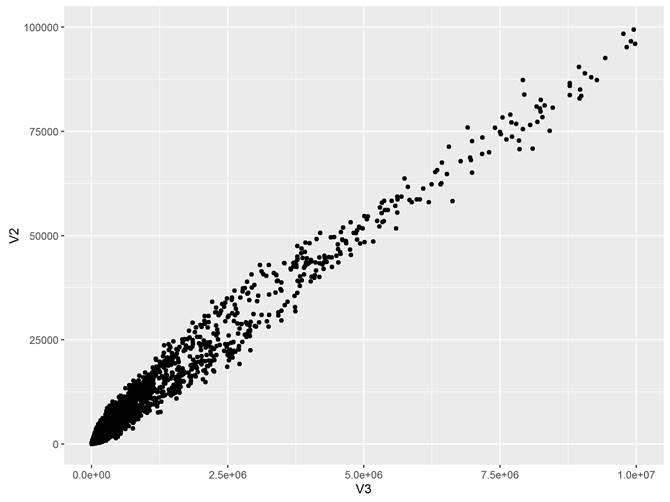
抽取结果分布：

Sogout网页库抽取网页统计情况

（一）X：pr，Y：抽样比例



（二）第一次抽样数据分析，第二次在余下数据中抽样，重复10次，X：站点url数量，Y：抽样数量



（三）X：站点url数量，Y：抽样比例

