数据科学导论 答辩展示

Sogou 用户日志分析

郭骏宇　肖桐　金佳熠

清华大学

2022 年 6 月 6 日

目录

研究背景

- 随着信息量的不断累积以及用户需求的日益多样性，搜索引擎给出的搜索结果经常存在不确定性，给用户的使用带来不便。因此，如何优化搜索策略，提高搜索效率，便成为了各个搜索引擎所面临的主要问题。为解决这个问题，分析用户的搜索行为是十分必要的。
- 对搜狗引擎日志数据集进行处理和分析，可以评估用户的搜索行为，挖掘用户的查询特征，从而为进一步提升搜索引擎的性能提供依据和指导。

数据预处理--编码转换 & 数据收集

将原始数据编码转换为 csv 格式

⇓

收集常见中文停用词便于后续分词

⇓

收集数字类型、自恋型、网址型搜索词关键词便于后续模糊分类

搜索关键词统计

- 调用 jieba 库对搜索词条进行分词，再根据已有的中文停用词表，去除分词后包含的停用词。
- 将所有的词汇总成一个列表，对列表中的元素进行归并处理，并按照频率从高到低排序，即可得到不同关键词的分布。取列表中前 10 个关键词得到 top 10 关键词。

搜索关键词统计

1	地震	89422
2	救灾物资	69092
3	哄抢	69084
4	汶川	68422
5	原因	62688
6	下载	39600
7	图片	32007
8	视频	29173
9	暗娼	26825
10	名单	20026

图 1: top10 关键词

	Frequency
count	117339.000000
mean	33.018425
std	565.349828
min	1.000000
25%	1.000000
50%	3.000000
75%	10.000000
max	89422.000000

图 2: 关键词

- 根据以上统计数据可以看出，频率的平均值为 33，远远小于 top10 关键词的频率数，且有 25% 的关键词词频为 1，有 75% 的关键词词频小于 10，均属于低频数据。
- 该现象满足 zipf 分布，属于幂律的表现形式之一，即 80% 的搜索频数是由 20% 的关键词所贡献的。若后续想完成对不同搜索词的特征分类，则应重点关注低频词，因为相对而言，高频词在此处容易被区分。

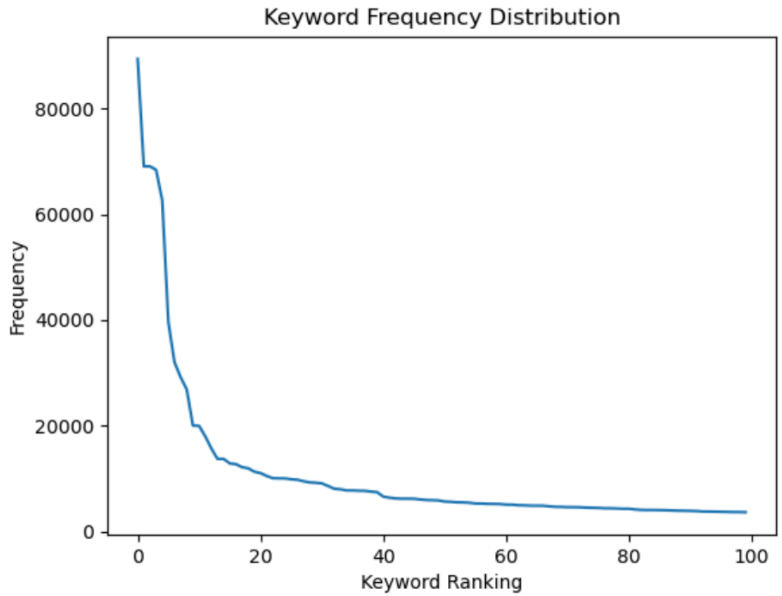


图 3: 词频分布

- 由上述可知，高频词汇已经不需要进一步的分析，故在此处仅考虑频率小于 500 的词，按照其频率所处区间（<10，10-50，50-200，200-500），分别赋予 0-3 四种标签，再进行可视化分析。

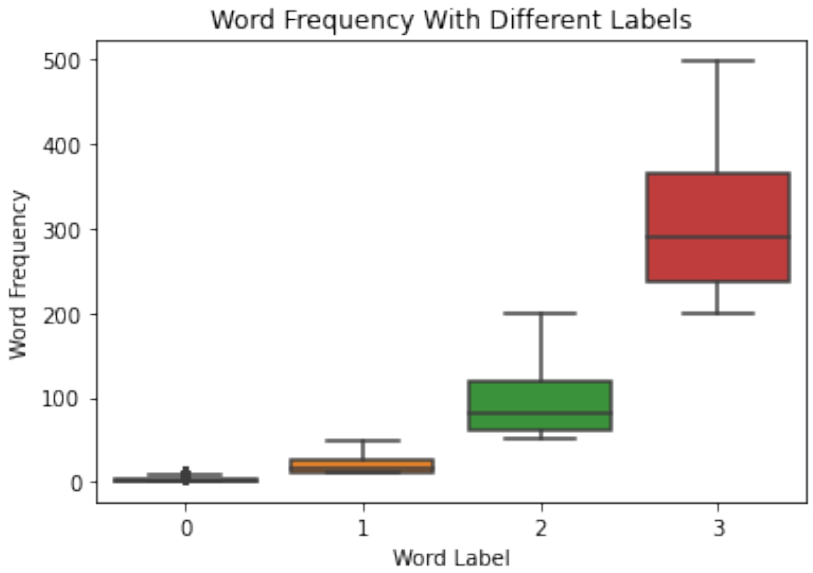


图 4: 词频分布

Word Cloud 可视化

- 在之前得到的搜索词频数列表中，取频数最高的 100 个词，将其作为参数绘制云图，并采用白色背景与云朵形状。

图 5: 词云

Word Cloud 可视化

- 词云中较为明显的词条，有汶川、地震、物资等。结合该日志数据的年份“2008 年”，可以知道是汶川大地震影响了人们的搜索行为。
- 由此可见，搜索关键词会较多地受到高热度事件的影响。大多数搜索关键词都聚焦于年度热点事件。

查询词分析

- 将用户 id 和搜索词条归并为列表，根据列表元素进行同类归并统计相同列表的个数。最后，再将每一个列表拆分成 id 和查询词，得到最终结果。
- 在此处排名最高的词条被一个用户查询了 274 次，而排名第十的只有 107，查询词条和关键词一样都有着相似的 zipf 分布的性质，频数衰减较快。

1	7822241147182134	free+girls	274
2	9165829432475153	人妖摄影	177
3	49180486532951556	babes	160
4	19447614244798927	绳艺kb视频	141
5	7076435807359547	库娃+三围	135
6	900755558064074	liuhecai	123
7	1756178764125793	屋面种植土	118
8	0767168563136269	沈阳家电维修+亨达	112
9	5515612701706876	玄幻小说	111
10	48453866664474126	free+movie	107

图 6: top10 查询词

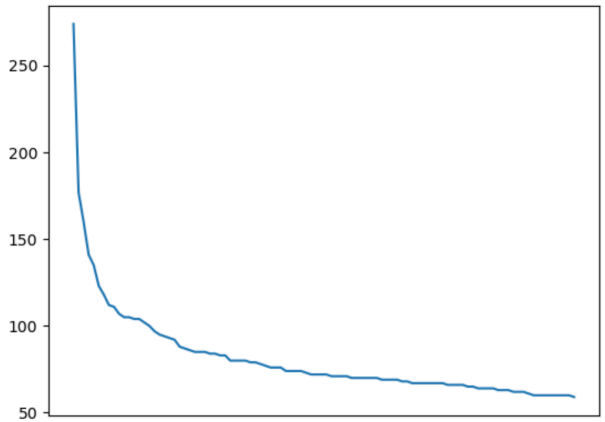


图 7: 查询词分布

查询词分析

- 将用户 id 和搜索词条归并形成列表，根据列表元素进行同类归并统计相同列表的个数。最后，再将每一个列表拆分成 id 和查询词，得到最终结果。
- 在此处排名最高的词条被一个用户查询了 274 次，而排名第十的只有 107，查询词条和关键词一样都有着相似的 zipf 分布的性质，频数衰减较快。

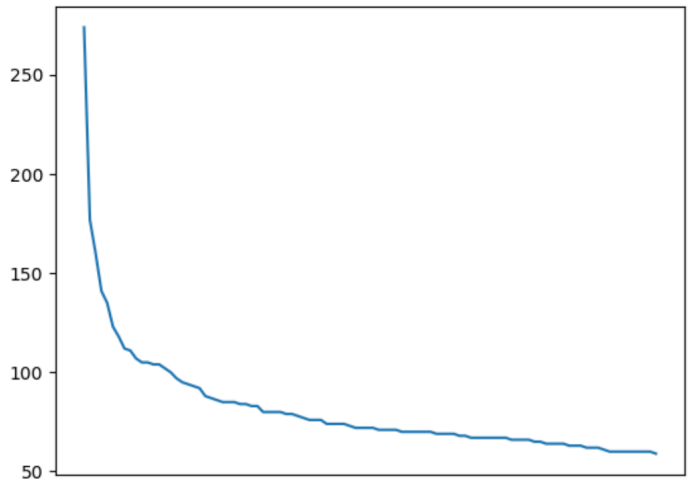


图 8: 词条频数分布

用户点击的顺序号分析

- 首先利用 numpy 中的 mean 函数计算日志数据中 order 列（用户点击的顺序号）的平均值，然后使用 dataframe 中的 value_counts 功能统计每个顺序号出现的频次并排序，取前 10 的顺序号保存至 click_order_top10.txt，最后利用 matplotlib 中的 pie 函数绘制排名前十位的 Order 和频数的数据分布饼状图。
- 用户点击的顺序号可以反映用户找到目标网页的难易程度，即若点击顺序的平均值较大，则用户寻找目标网页需要经过更多的点击。统计日志集中用户点击顺序的平均值为 7.72，说明搜索引擎有时候依然有较大的不确定性。

用户点击的顺序号分析

- 由饼状图可知，顺序越靠后的搜索结果用户点击的频率越低且频数急剧递减，点击第一个和第二个结果的次数占据了一半以上，也有着类似的 Zipf 分布。这与大多数用户的使用习惯相一致，即用户通常会从第一个结果开始依次往后观察，直到得到了满意的结果。

1	511806
2	267076
3	165821
4	116992
5	88327
6	69848
7	56197
8	46796
9	39280
10	33799

图 9: 点击顺序号频数统计

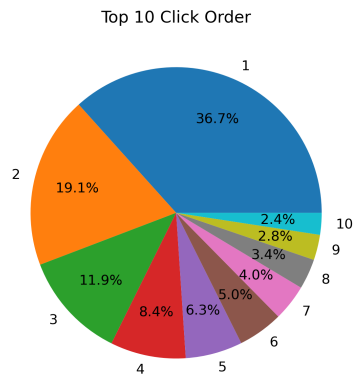


图 10: 点击顺序号频数分布饼状图

搜索时间段分析

- 首先，用 split 函数提取出“访问时间”列中的“小时”，并用其替代原“访问时间”列。其次，用 groupby 函数依据“访问时间”重组 dataframe。接着，得到重组后每一组对应的时间段和用户搜索次数，将时间段和用户搜索次数这两列合并为新的 dataframe，存入 search_time_analysis.txt。最后，画出对应的柱状图，在极大极小值点上标注数值，并存为 search_time_analysis.png

搜索时间段分析

由柱状图可以看出：

1. 用户搜索主要集中在早上八点至凌晨十二点，与大众作息时间相符
2. 搜索量的几个峰值出现在中午十一十二点，下午四五点以及晚上八九点，再结合极小值分析可以发现，饭点的用户搜索量会相对较低，而饭前饭后的用户搜索量则会相对较高

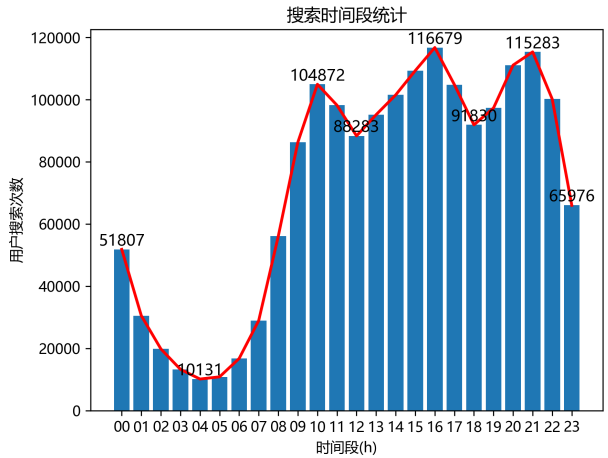


图 11: 搜索时间段分布柱状图

用户平均点击量分析

- 首先，去掉“用户点击的 URL”列，用 groupby 依据“用户 ID”分组。接着，计算每个用户 ID 对应的总点击数和总查询词个数，用总点击数/总查询词个数计算出每个用户 ID 在单个查询词上的平均点击数。其次，分别统计平均点击量 =1，>=3，>=5，>=7，>=10 的用户数和用户占比，将这些数据合并为新的 dataframe，存入 rep_click.csv

- 由表格可以看出，83% 的用户在一个搜索词结果中的平均点击量 <3，即平均而言，大部分用户不会在一个搜索词的结果中较多次点击。这可能说明搜狗引擎得到的结果整体上与用户需求匹配度较高，也可能说明用户会及时更改搜索词以获取更多信息或是得到更贴切的结果。更进一步的分析还需要依赖于对用户搜索词相似度的研究（探究用户更换搜索词的行为）

单次搜索平均点击量	用户数	用户占比
=1	286453	0.5300984312803839
>=3	94691	0.17523136624985888
>=5	28273	0.05232087968214783
>=7	12059	0.02231590167605209
>=10	4720	0.00873464266613864

图 12: 用户平均点击量统计

URL 点击搜索分析

- URL 地址代表用户搜索点击的具体网址，即便是一样的搜索词用户也可以会选择点进不同的词条，即会产生不一样的 URL，在此处对 URL 地址使用 group_by 函数并计算不同 URL 的频数，对其分布进行可视化分析。

	Unnamed: 0	url	Frequency
0	0	news.21cn.com/social/daqian/2008/05/29/4777194...	17229
1	1	news.21cn.com/zhuanti/domestic/08dizhen/2008/0...	10955
2	2	pic.news.mop.com/gs/2008/0528/12985.shtml	9222
3	3	www.tudou.com/programs/view/2F3E6SGHFLA/	8229
4	4	bjyouth.ynet.com/view.jsp?oid=40472396	6994
5	5	www.17tech.com/news/20080531107270.shtml	5360
6	6	www.baidu.com/	3881
7	7	bbs.cdqss.com/thread-59453-1-1.html	2964
8	8	www.taihainet.com/news/military/jslwt/2007-06-...	2391
9	9	news.vnet.cn/photo/292_6.html	2290

图 13: 统计出现次数最多的 URL 地址

- 对 URL 的频数分布进行统计，得到左图结果。
- 由于即使搜索同一个关键词，同一个用户也不会反复点击同一个 URL，所以在此处大部分的 URL 出现频率都是 1 次。但是最高点击次数较多，达到接近 20000 次，这里推测可能是较为热门的帖子，例如关于汶川大地震的新闻报道等帖子。类似于前面对关键词进行分类的标准，在此处对 URL 依旧进行如上的标签分类，在此处高于 10 次的占少数。

	Frequency
count	514760.000000
mean	2.037017
std	38.716364
min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
max	17229.000000

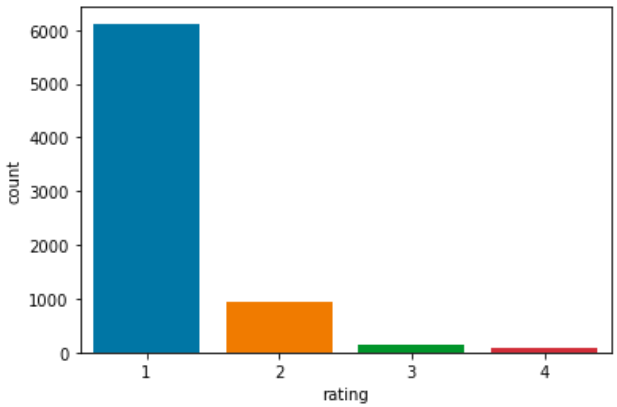
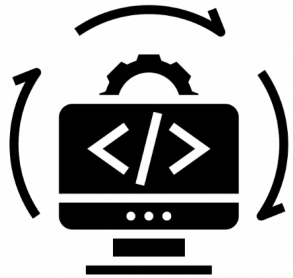


图 14: URL 频数分布统计　　图 15: URL 标签分类结果

程序运行与测试方法

- 将每个子任务单独写进一个 python 源文件中，通过保存为文件的形式实现数据的互通（例如一个程序的输出结果作为另一个的输入）。
- 下载了规模更小的迷你版日志进行调试，这样可以获得更高的调试效率。通过迷你数据集上的测试效果以及运行时间合理预测在大样本上的运行时间，从而选择合适的算法。



不足之处

- 采用 PySpark 可以进一步提高数据处理效率
- 在此处仅有一天的数据量，而传统低频数据分析通常采用更长时间维度的分析
- 后续可以建立用户搜索评价体系，分析搜索词条是否匹配用户需求的程度

未来展望

- 采取对部分数据进行人工标记，利用 Kmeans 方法构造决策树可以对低频搜索词进行特征提取
- 运用用户评价体系中的用户搜索满意度对决策树进行修正，从而得到更加满意的搜索结果。

分工

- 任务 1：金佳熠、肖桐
- 任务 2：金佳熠
- 任务 3：郭骏宇、金佳熠
- 任务 4：金佳熠
- 任务 5：肖桐
- Bonus1（单次搜索点击量分析）：肖桐
- Bonus2（URL 点击搜索分析）：郭骏宇
- 报告撰写：郭骏宇、金佳熠、肖桐
- PPT 制作：郭骏宇、金佳熠、肖桐

实验环境简述

- Python version : 3.9.7
- Pandas version : 1.3.4
- Seaborn version : 0.11.2
- Numpy version : 1.20.3
- Jieba version : 0.42.1
- Wordcloud version : 1.8.1
- Matplotlib version: 3.3.4
- Pillow version : 8.2.0

参考资料

[1] 姚婷, 张敏, 刘奕群, 马少平, 茹立云. 低频查询的用户行为分析和类别研究 [J]. 计算机研究与发展,2012,49(11):2368-2375.

[2] 张磊, 李亚楠, 王斌, 李鹏, 蒋在帆. 网页搜索引擎查询日志的 Session 划分研究 [J]. 中文信息学报,2009,23(02):54-61.

[3] 万飞, 赵溪, 梁循, 潘登, 倪志豪. 基于移动互联网日志的搜索引擎用户行为研究 [J]. 中文信息学报,2014,28(02):144-150.

[4] 刘健, 刘奕群, 马少平, 张敏, 茹立云, 张阔. 搜索引擎用户行为与用户满意度的关联研究 [J]. 中文信息学报,2014,28(01):73-79.

结语

Thank you!