1. **query-click-query**



任务1中，查询-点击数量280条，查询-点击-查询数量15条，占比5.3%

任务2中，查询-点击数量318条，查询-点击-查询数量25条，占比7.9%





任务1的查询-点击-查询数据中，后续跟有点击的有11条，占比73%

任务2的查询-点击-查询数据中，后续跟有点击的有18条，占比72%

任务1中查询-点击-查询的后续点击平均数量为19/15 = 1.3条

任务2中查询-点击-查询的后续点击平均数量为46/25 = 1.8条

从总体来看

任务1查询总数为172条，后续跟有点击的有131条，占比76%

任务2查询总数为145条，后续跟有点击的有118条，占比81%

任务1查询平均点击数量为299/173 = 1.7条

任务2查询平均点击数量为338/145 = 2.3条

1. **粒度差**



16组数据中

2组时间树的粒度差和小于线性日志的粒度差和

4组相等

10组时间树的粒度差和大于线性日志的粒度差和。

1. **重构链**

（从重构链可以覆盖更多的查询词，从查询推荐的角度，可以单独做一组实验）

交叉验证



重构链条数

16组数据中有2组时间树上重构链条数少于线性日志

其余14组均多于线性日志



重构链上平均节点个数

16组数据中有1组时间树上重构链上平均节点个数少于线性日志

其余15组均多于线性日志

1. **聚类怎么能合理**

（看看能不能断开）（相当于找到了一个合理的做法把树上不好的结构调整好？）

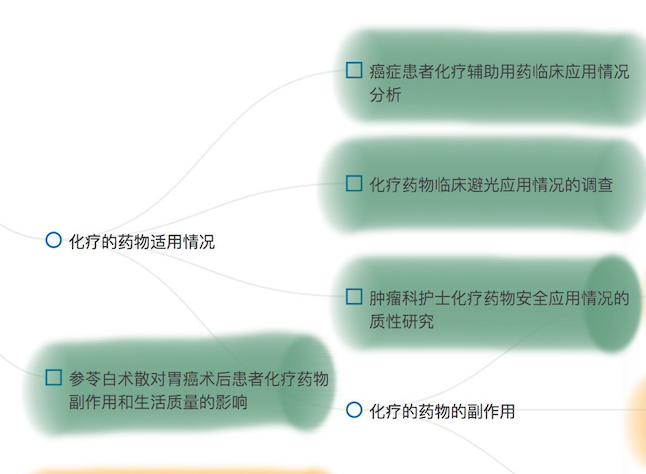
断开：结合状态迁移和粒度

发生了状态迁移，同粒度的情况下，发生处的查询应该与其前一个查询的同粒度最源查询并列。一个查询的最源查询是该查询的祖先中上一次发生状态迁移之后的查询，同粒度最源查询就是……。如果是中粒度到细粒度的状态迁移，就不做断开了，后续看情况再进行分析处理。如果是细粒度到中粒度的状态迁移，该节点就应当与离他最近的中粒度祖先节点的最源查询节点并列。

从点击切题性的角度，不切题的点击经常扎堆儿出现，大致观察了一下，可以统计一下比如不切题的点击出现的情况下，不切题点击占比高于百分之50的比例有多少，或者计算一下不切题点击出现的条件下不切题点击的平均占比。

这样我们就可以说

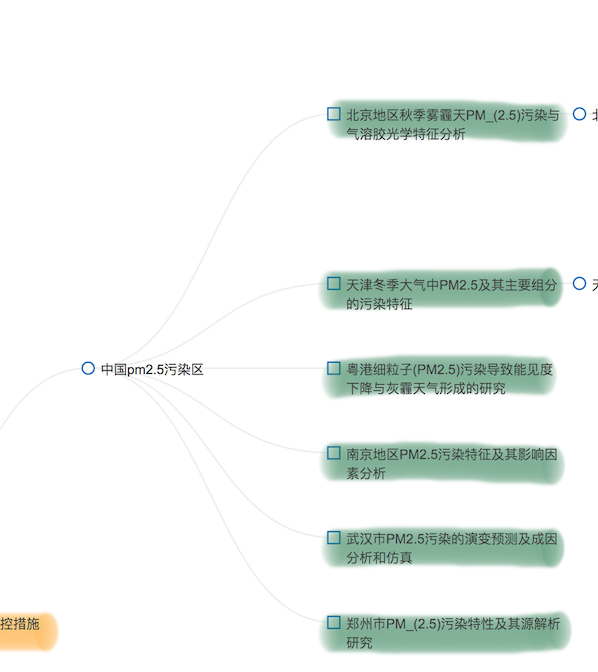
1. 如果出现大量不切题的点击，那么很大可能这个查询是不好的查询。



1. 如果一个查询的点击大部分都是切题的，它们往往都属于同一个主题，即同一个子任务。



1. 还有另外一种可能，就是单独看他们不是切题点击，但是合起来，却是一个子任务。



但不管是哪种情况，一个点击后面跟着的一堆查询，往往可以算作一类内容，可能是属于同一个主题，也可能都是不同的主题即跟查询词无关的各种发散的东西。

这种结构和主题的对应关系在一棵极小子树上肯定是存在的。那么我们有理由猜想，这种结构和主题的对应关系，在更大的子树上也是存在的。

于是我们提出基于位置的主题划分方法。

然后是方法。

然后验证。

然后结论。

另一种思路，算是一种补充，我们发现，有时用户进行的任务会有反复，即图上会有出现重复搜索的过程，并且在任务范围内用户之间的相似查询也会有很多，但是主题的识别却不容易做，那么我们可以基于上述的结构和主题之间的对应关系，我们可以先划分成几块，然后把主题相似的子树拼成大树。以此为基础未来可以做查询推荐。

1. **状态迁移**

每个图中发生状态迁移的查询数与总查询数的比例：



时间线性日志发生状态迁移的查询数与总查询数的比例：



状态迁移与粒度变化之间的关系：



时间线性日志状态迁移与粒度变化之间的关系：



1. **粒度上升**

大部分粒度上升伴随着状态迁移：



但是线性日志上也是这样：



1. **点击粒度**

各粒度点击所占百分比：



点击粒度与查询粒度的关系：

1. 各粒度查询产生各粒度点击的情况：



1. 各粒度点击所来自的源查询的粒度情况：



1. **点击切题性**

切题性百分比：



点击切题性与其源查询粒度的关系：

1. 各粒度查询产生切题／不切题点击的情况：



1. 切题／不切题点击的源查询粒度的情况：



点击切题性与其自身粒度的关系：

1. 切题／不切题点击各粒度分布情况：



1. 各粒度点击的切题／不切题分布情况：



1. **查询粒度，点击粒度，点击切题性之间的关系**

列出所有组合，比例最高的是中粒度查询产生切题细粒度点击，其次是细粒度查询产生切题细粒度点击。

