实验结果汇总

符号：

Etime：Estimate Time

Dtime: Dwell Time

Sat: Satisfaction

Midsat: Medium Satisfaction

Unsat: Unsatisfaction

**实验1.1 研究temporal relevance，satisfaction单个变量对于感知时间的影响。**

基本假设：在high temporal relevance条件下的用户，比low temporal relevance条件下的用户倾向于估计更长的时间。

实验：我们在satisfaction设置相同的情况下，比较同一个task上high temporal relevance和low temporal relevance条件下的差值Avg(Etime-Dtime)以及差值与Avg((Etime-Dtime)/Dtime)。

图1在不同满意度条件下，High/Low Temporal Relevance 条件下用户估计时间的情况

图2：在不同的满意度条件下， High/Low Temporal Relevance 条件下用户估计时间的比较偏差率

图中，横轴是task的序号，每一个点上是5个用户的用户的平均值。其中，task11—14是time critical 的query。直观看来，

1. High temporal Relevance条件下，用户倾向于估计的时间更长，但是这个情况在不同的满意度设置下，在不同的task上并不一致。图1中

在Sat条件下，11,13,14,4,10,17,20 与假设不一致；

在Midsat条件下,11,13,1,2,5，15,9,10,20与假设不一致；

在Unsat条件下，14,6,7,8,16与假设不一致；

1. 从差值和偏差率来看，high条件下的值大体上比low条件下的值大，这从一定程度上说明了high条件下用户在Dtime的基础上倾向于估计更长的时间。不过这个现象并不十分明显，可能的原因是没有考虑到user的因素，这里的数据都是所有满足条件的user的平均值，但每个user并不是把所有的task都做过一遍。
2. 在图1和图2所示的实验中，暂时看不清Sat控制的效果。我们讨论到一种情况，可能会导致sat的控制失效，比如说有一个结果，在第7、第8位可以找到一个relevance的结果，但是reverse serp之后反而到了前面。
3. 上面的实验结果事实上认为用户是同质的，即不同的用户对时间估计的偏好是差不多的，这个肯定是有问题的，在后面有一个相关的实验。

**实验1.2 研究在不同的Temporal Relevance 设置下，不同的Satisfaction条件对用户估计时间的影响。**

**基本假设：用户在越不满意的情况下倾向于估计更长的时间。**

实验：我们在temporal relevance设置相同的情况下，比较同一个task上sat，midsat和unsat条件下Etime和Dtime的差值，以及差值与Dtime的比值（偏差率）。

从差值和偏差率来看，sat，midsat和unsat三种条件下看不出明显的大小关系，我们的假设并不成立。可能的原因有两方面，

第一，我们对于sat，midsat和unsat的设置失效。实验中我们是颠倒搜索结果的排序来进行设置，但由于我们的系统本身对搜狗的搜索结果进行了过滤，而且有一些task在最靠前的位置反而并不能找到好的结果，这使得sat，midsat和unsat的标准并不可靠；

第二，在实验这种环境下，用户对于task的满意度的理解和日常的搜索环境可能是不一样的，有可能用户花了很大功夫找到了一个很满意的答案，他就会觉得很满意，我们给用户设定的信息需求不一定能反映他的真实心理，他更多只是把这当成一个任务。

总结：这一部分主要研究temporal relevance，satisfaction单个变量对于时间感知的影响。就satisfaction来说，并没有找到任何有价值的证据来验证我们的假设，而temporal relevance方面，实验数据一定程度上说明了high条件下用户在Dtime的基础上倾向于估计更长的时间，不过这个现象目前还并不十分明显，可能的原因是没有考虑到user的个性化因素。

**实验2: 研究task本身是否具有一定的属性？比如会倾向于让用户估计时间更长或者更短。**

实验：在整个实验中，每个task会被做30次，其中high temporal relevance和low temporal relevance条件下各会被做15次，我们比较两种条件下task的Etime比Dtime长的比例。

图3. 在不同的设置下，Etime > Dtime的比例

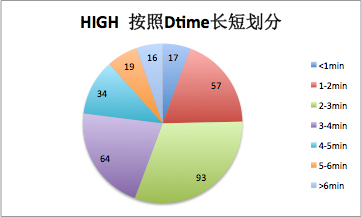
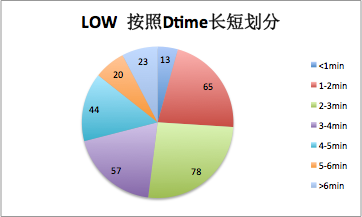
可以看出，对于同一个task来说，high temporal relevance条件下，Etime>Dtime的用户（实验样本）比例更高，这说明high temporal relevance条件下用户更倾向于把时间估计得比真实时间更长。

**实验3：Dtime 对Etime的影响。**

基本假设：dwell time的长短会对etime的估计产生影响，比如dwell time短的时候，影响比较大。

实验：我们分别在high temporal relevance（后简称HIGH）和low temporal relevance（后简称LOW）下，分析dwell time（Dtime）和估计时间偏差率的关系，看看在这两种情况下Dtime是不是都会对偏差率产生影响。如果有影响的话，会产生什么样的影响。

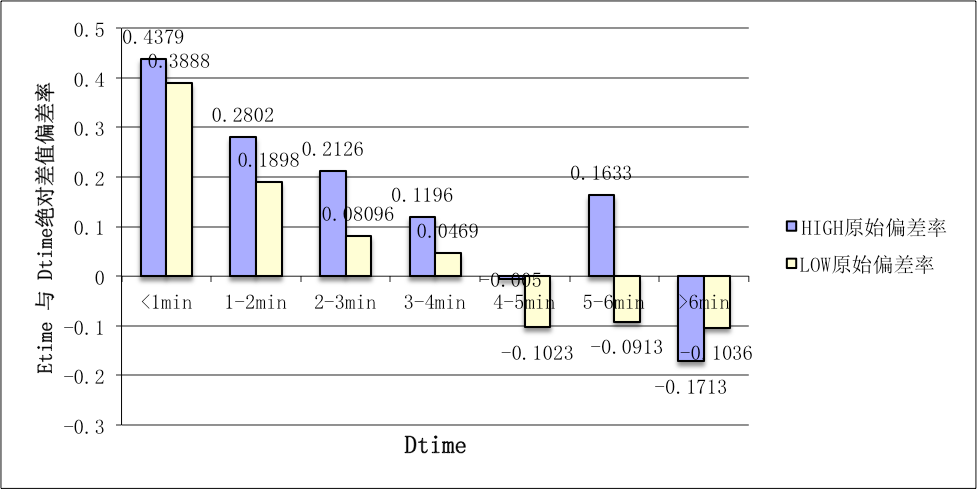
1. 首先，根据Dtime的长短，每间隔一分钟划分一个区间，共划分出了7个区间。我们先统计HIGH、LOW各300条结果的Dtime在各个区间的分布情况，如下两图所示。很容易发现，用户完成实验的Dtime主要集中在1-5min的区间上（均>80%）。



1. 在根据Dtime划分出的每个小区间中，计算Etime与Dtime二者差值偏差率的平均值，即（Etime - Dtime）/ Dtime 的平均值。这不仅可以让我们看到估计的偏差率，还可以看到用户是更容易将时间估计长还是估计短。比较不同区间之间，以及HIGH、LOW两种设置下的偏差率。

绘制出的图像如下图所示。只考虑Dtime分布集中的1-5min区间，可以明显看出随着Dtime增长，偏差率逐渐减小，即时间越长估计越准确。而且可以看出随着Dtime增长，用户倾向于将时间估计短。

我们还观察到，在每个区间内，都是LOW比HIGH偏差率小，即估计得更加准确一些。原因应该是HIGH条件下用户更容易估计出更长的时间。

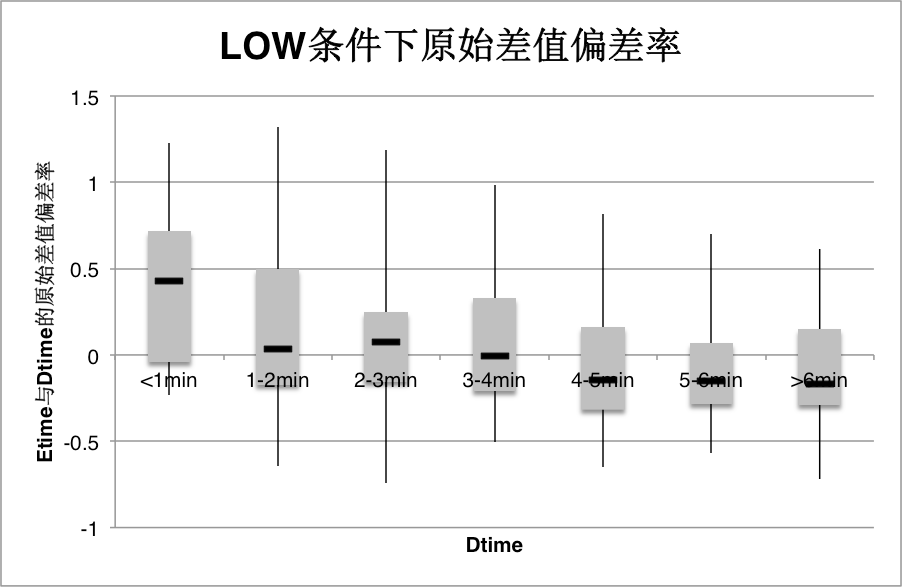
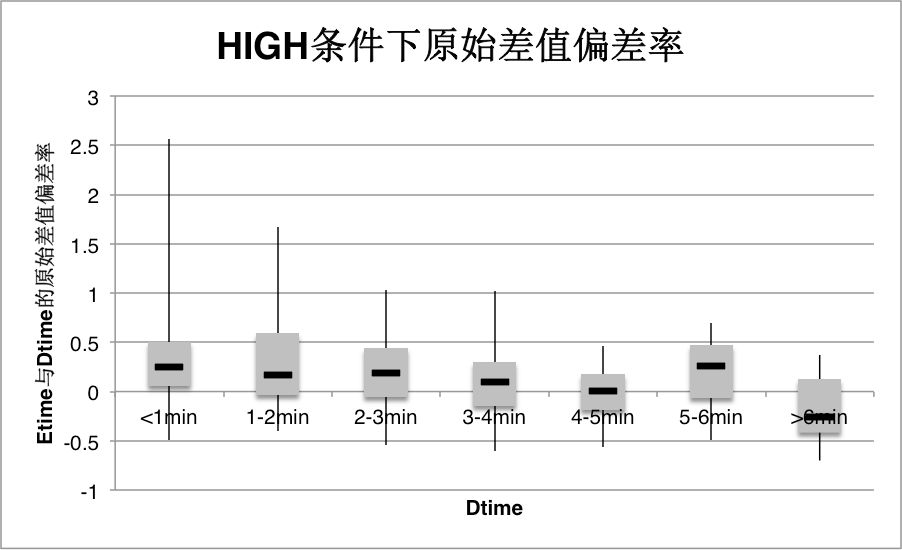


1. 由于平均值并不能反映出每个区间的整体情况，所以接下来采用箱体图来进行进一步的数据展示。

分别看HIGH条件下和LOW条件下的偏差率，在1-5min区间基本可以看出箱体的整体下移趋势。也就是说上面得出的随着Dtime增加，用户的时间感知越来越准确，这一结论是比较可靠的。

上面得出的第二个结论，LOW比HIGH估计得更加准确，这一点只能在1-4min区间中通过中位数看出来，并不十分显著。

还可以明显看出，LOW在每个区间的数据分布范围都比HIGH的要大很多，这比较容易理解，因为HIGH条件有一定的时间参考，个人估计时间的差异性会被削弱一些。

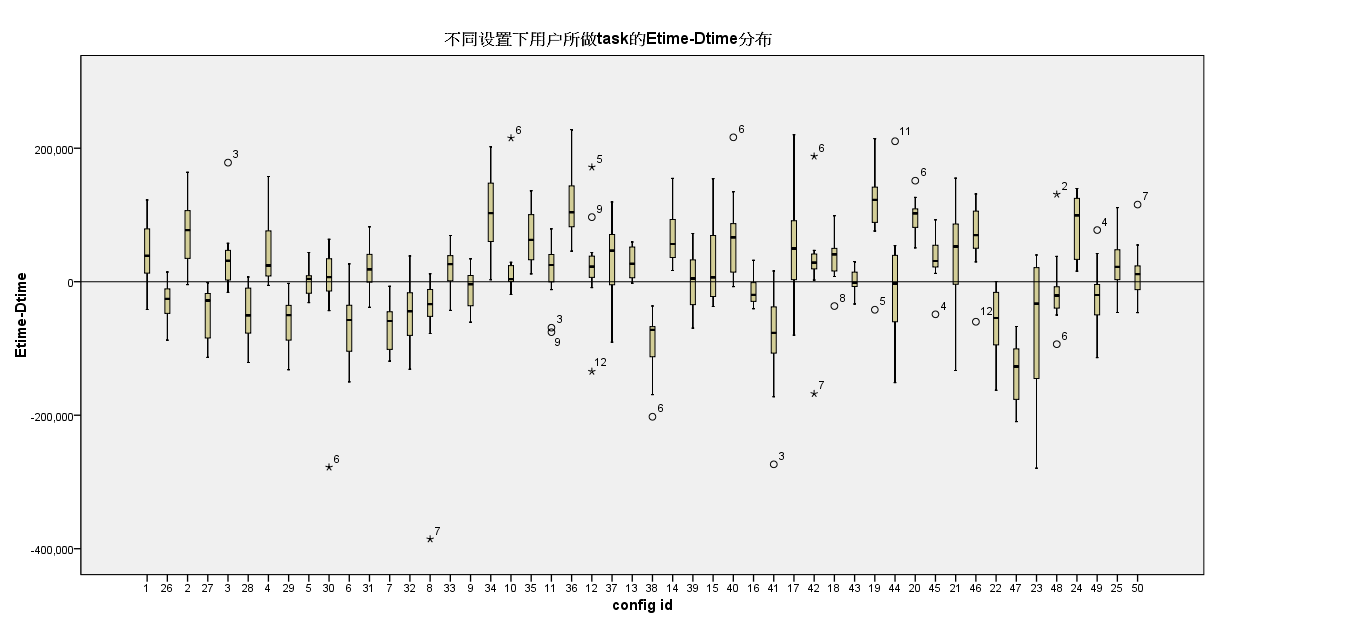


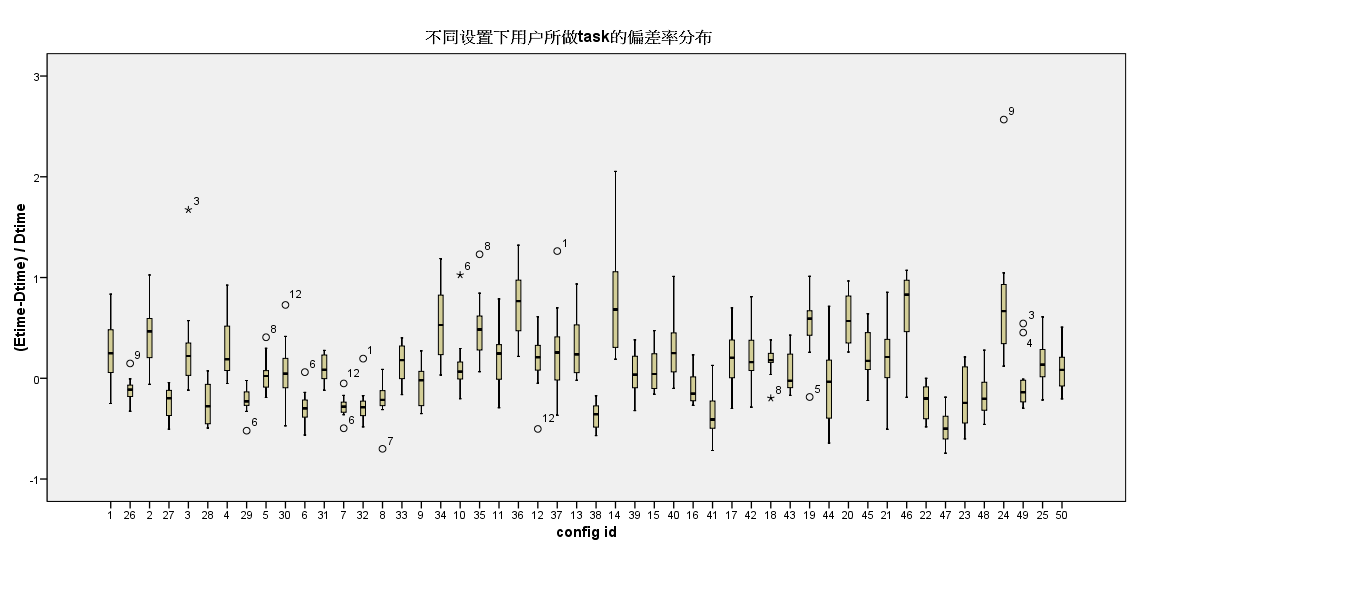
总结：通过实验我们验证了提出的假设，不同的dwell time确实会实验者估计时间的偏差产生影响。用户Dtime集中分布于1-5min，在这个区间内，用户完成任务的实际时间越长，对时间的感知就更准确。并且，HIGH条件会有效抑制用户本身估计时间的个人差异性，数据分布范围更小。

1. 研究不同temporal relevance条件下不同user的时间感知的情况

假设：high temporal relevance条件下的user比low temporal relevance条件下的user倾向于估计更长的时间

实验：我们对其他设置相同，只有temporal relevance设置不同的user（设置1和26, 2和27, …, 25和50）所做的12个task的Etime和Dtime的差值以及差值与Dtime的比值（偏差率）进行对比，看二者是否有明显区别，为了直观看出user所做的12个task的整体差异，用箱线图表示。





图中横轴相邻的两个设置（如1和26）只有temporal relevance设置不同，可以看出基本上两个user估计的12个task的差异是比较明显的，但我们的假设并没有得到验证，因为虽然差异明显，但有的是high条件下差值和偏差率大，有的是low条件下差值和偏差率大，这不能直接说明temporal relevance这个因素对感知时间的影响，这可能跟不同的设置有关，即可能在有些设置下high>low，在有些设置下low>high，这说明设置本身也会对时间感知带来影响，要预测时间感知需要考虑到设置方面存在的变量。不过值得注意的一点是user与user之间估计行为的差异是比较明显的，这说明我们考虑user个性化的因素是很有必要的。