Query-level Satisfaction 与Session-level Satisfaction 相关性的情况

我们实验数据上的结果				
Measure	Correlation with Sastisfaction			
	Pearson	p-value	Kendall's tau	p-value
Search Outcome(sCG)	0.653	1.34E-43	0.440	2.37E-34
Search Effort (# queries)	-0.435	1.77E-17	-0.303	3.26E-17
Search Outcome / Effort (sCG / #queries)	**0.676**	1.21E-47	**0.488**	6.83E-42
sDCG	0.497	4.47E-23	0.330	4.89E-20
nsCG	**0.676**	1.21E-47	**0.488**	6.83E-42
nsDCG	0. 636	7.6E-41	0.451	4.5E-36
Jiepu Jiang的结果				
	Correlation with Satisfaction			
Search Outcome (sCG)	0.27		0.22	
Search Effort (# queries)	-0.24		-0.23	
Search Outcome / Effort (sCG / #queries)	**0.77**		**0.59**	
sDCG (Järvelin et al [18])	0.41		0.29	
nsCG	**0.77**		**0.59**	
nsDCG (Kanoulas et al. [22])	0.75		0.57	

每一个用户完成12个task，每一个task对应于一个session，在每一个session中，用户对提交的每一个查询标注了结果的满意度，同时对每一个session标注了整体的满意度。

对所有的用户对query标注的满意度，将其转化为标准分数；对于用户对session上标注的数据，同样将其转化为标准分数，这两个序列做相关性的分析，计算Pearson 系数和Kendall's tau， 对比我们的结果和Jiepu Jiang的结果如下：

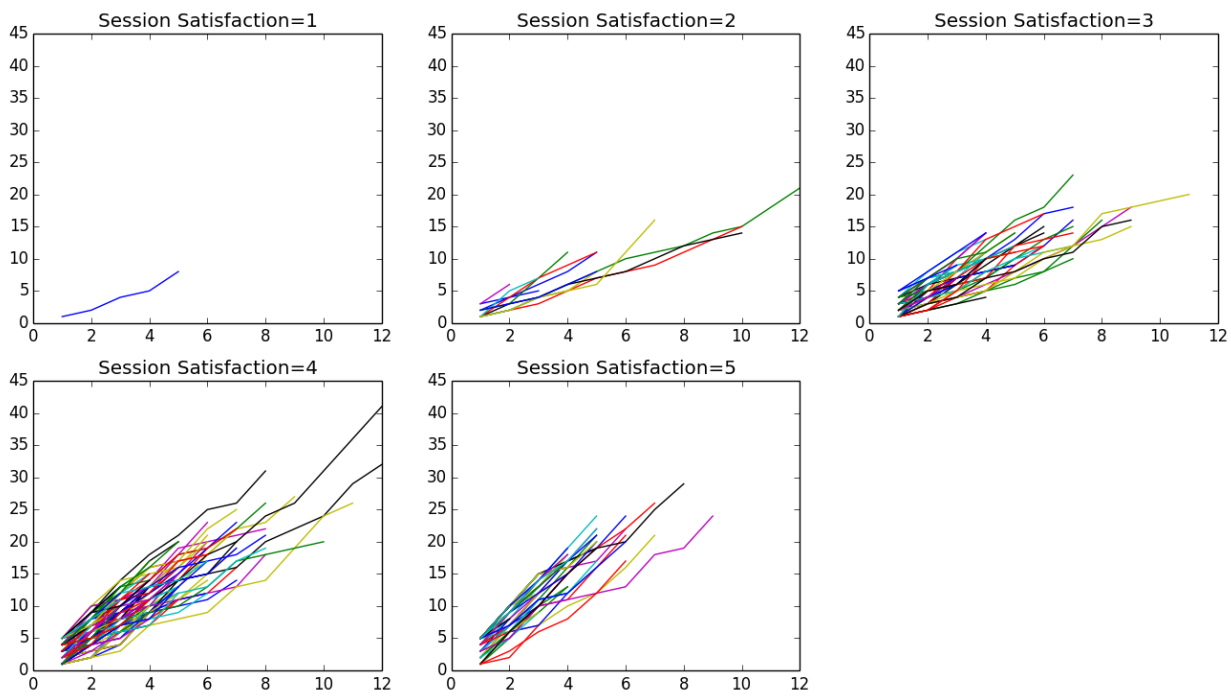
对于我们的实验和Jiepu Jiang的实验对比，在Search Outcome(sCG) 和Search Outcome/Effort (sCG/#queries) 上的趋势是一样的，都是Search Outcome要好一些。sDCG要稍微差一点，这个趋势也是正确的。

值得一提的是，我们这部分实验和Jiepu Jiang的实验并不一样，他的实验中采用标注人员标注了每一个query的result quality，我们用的是用户标注的满意度。但是得到的结果是接近的。

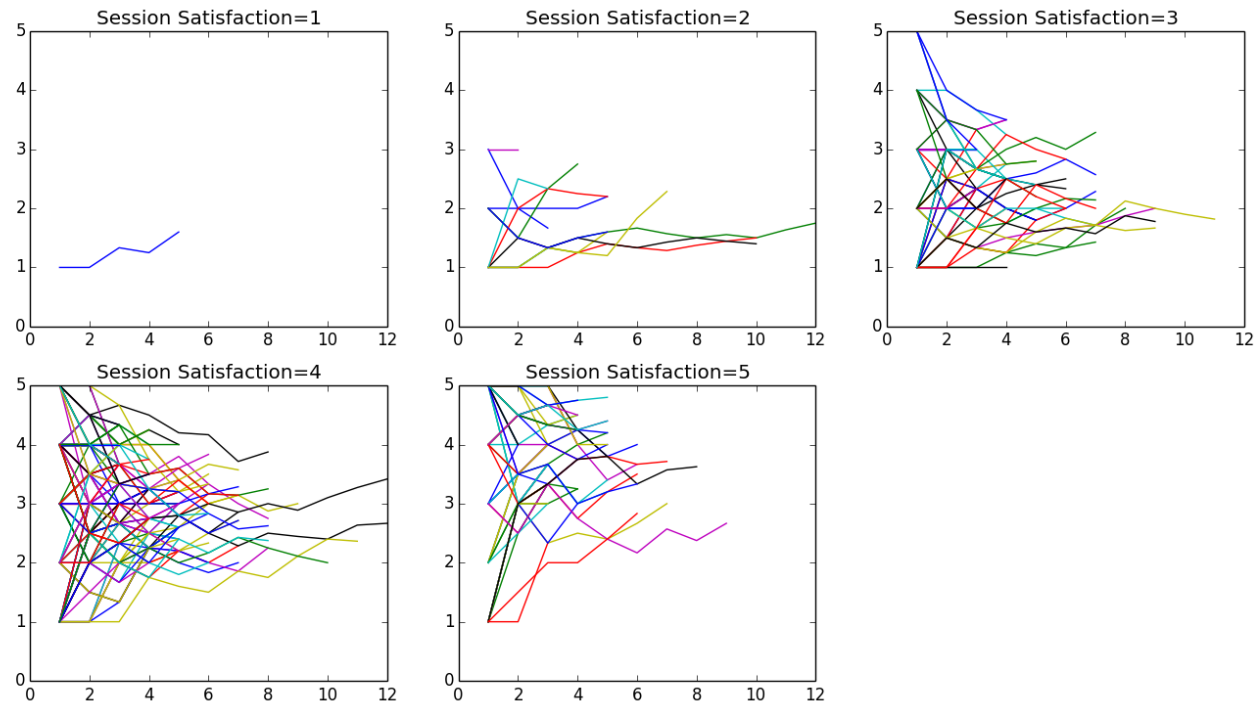
第二部分的实验，我们分析了sCG 和 sCG/#queries 两个指标随着query的数量增加的时候的变化，这里对query的满意度进行了一个预处理，在每一个人的维度上，将所有的query level/session level 的satisfaction归一化为z-score，

首先按照最终session level 的满意度区分（没有归一化之前，1~5）

sCG

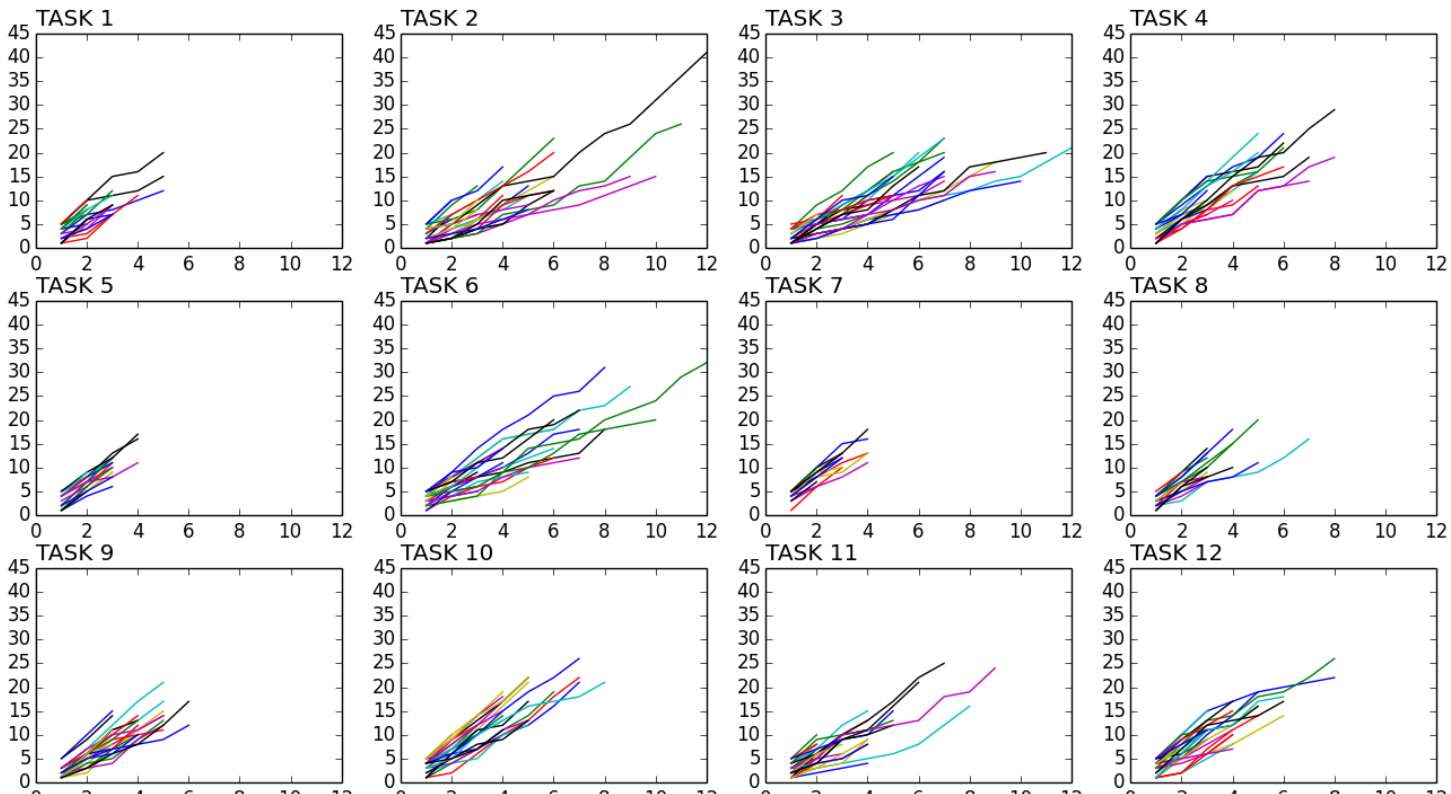


sCG/#queries

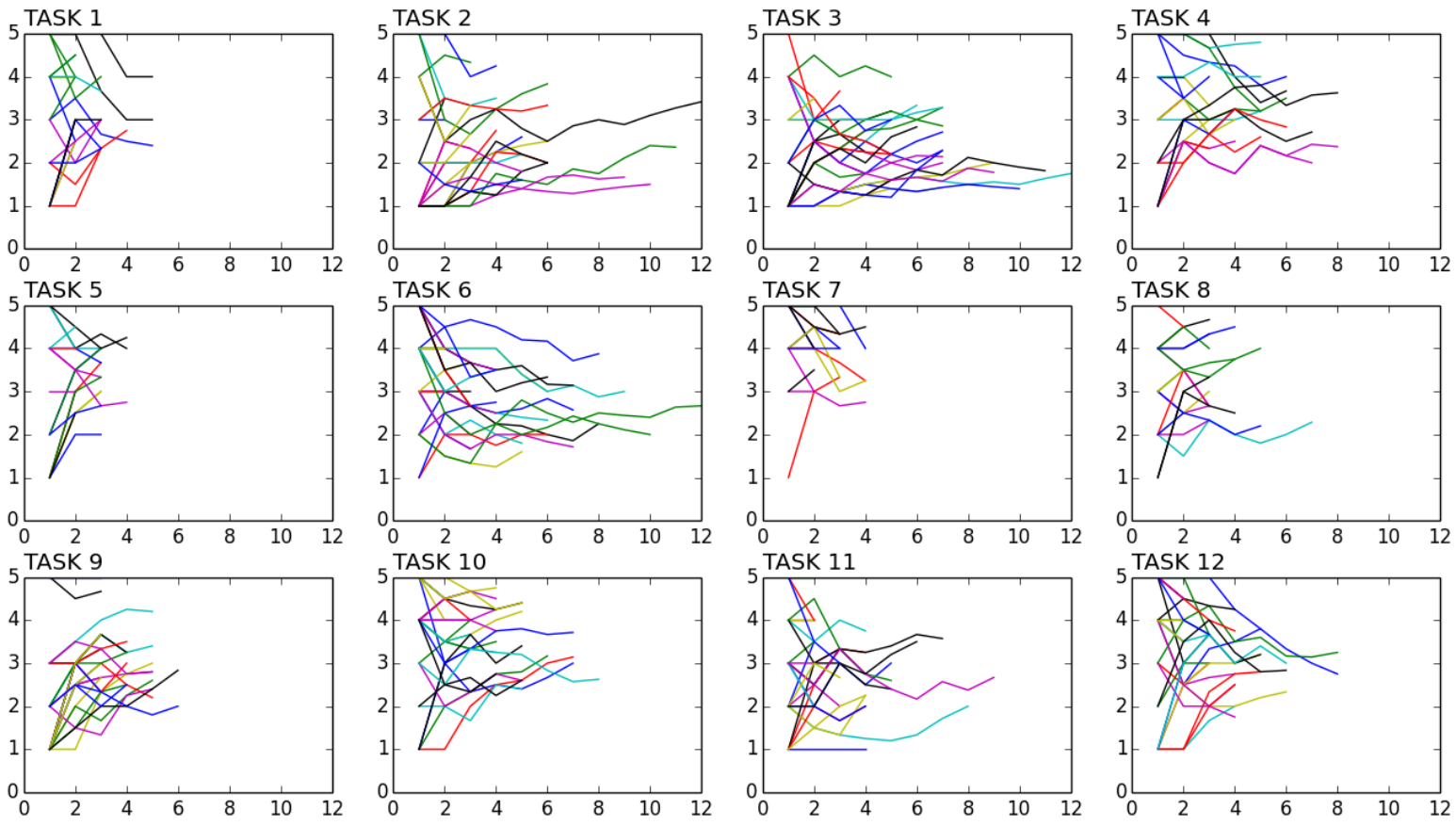


其次，按照1~12个任务区分：

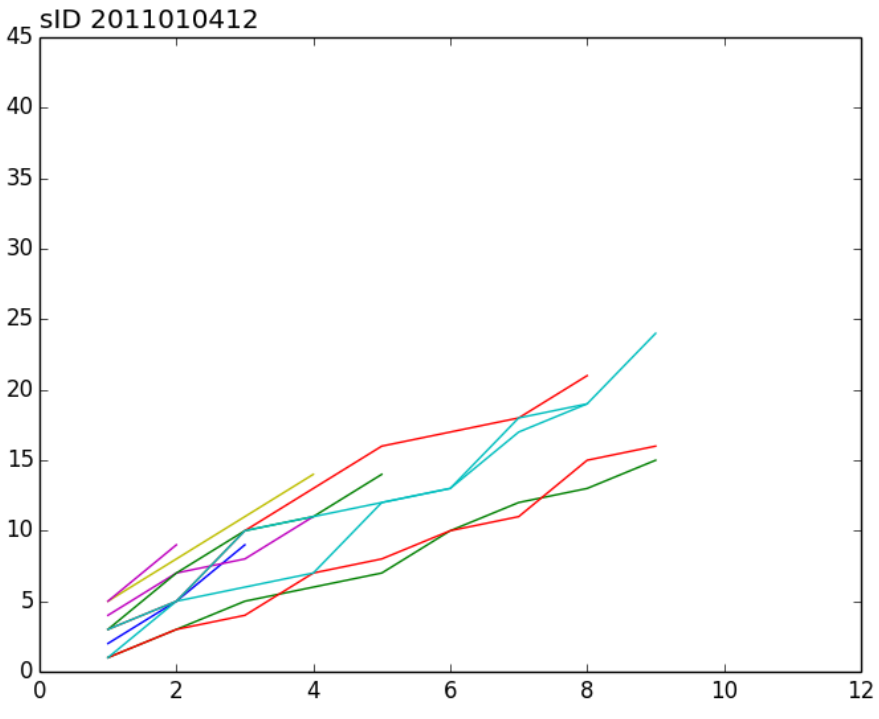
sCG

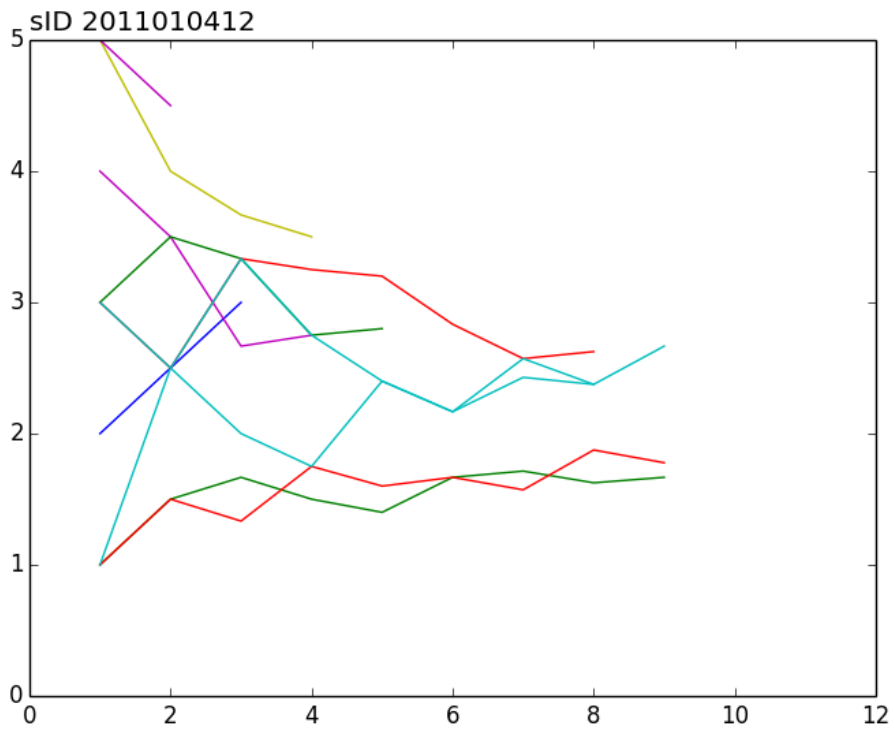


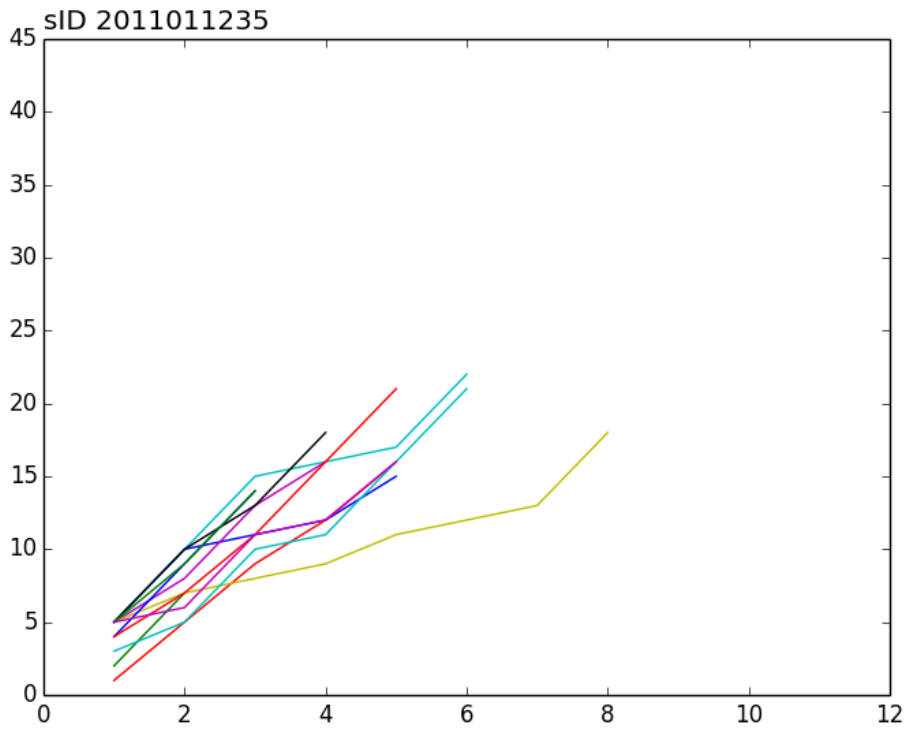
sCG/
#queries

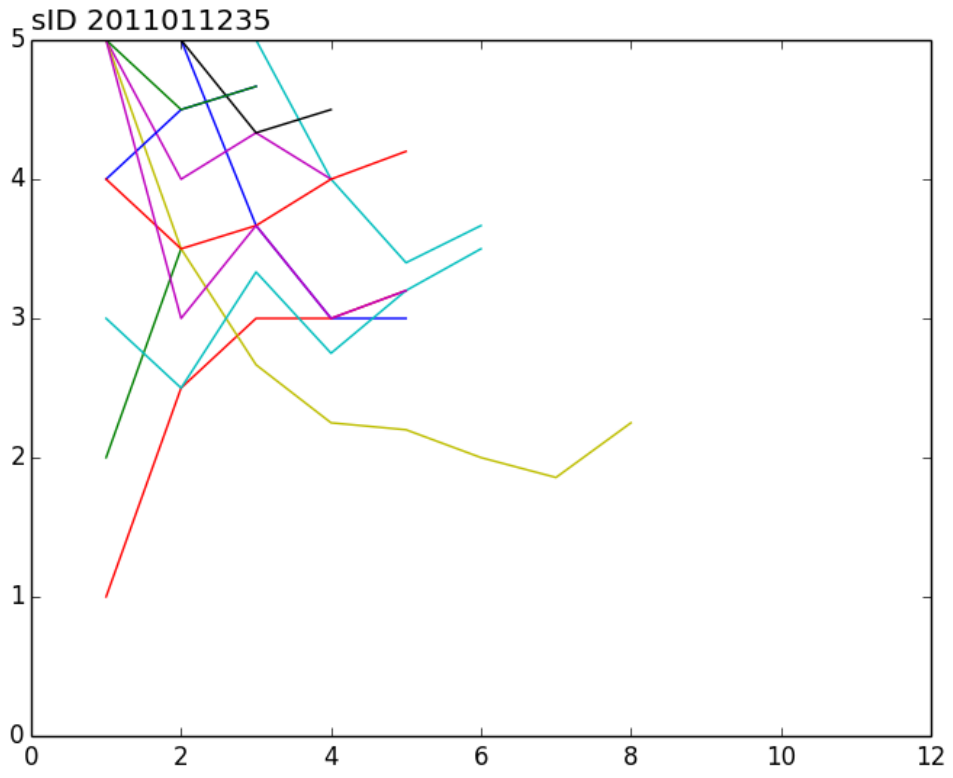


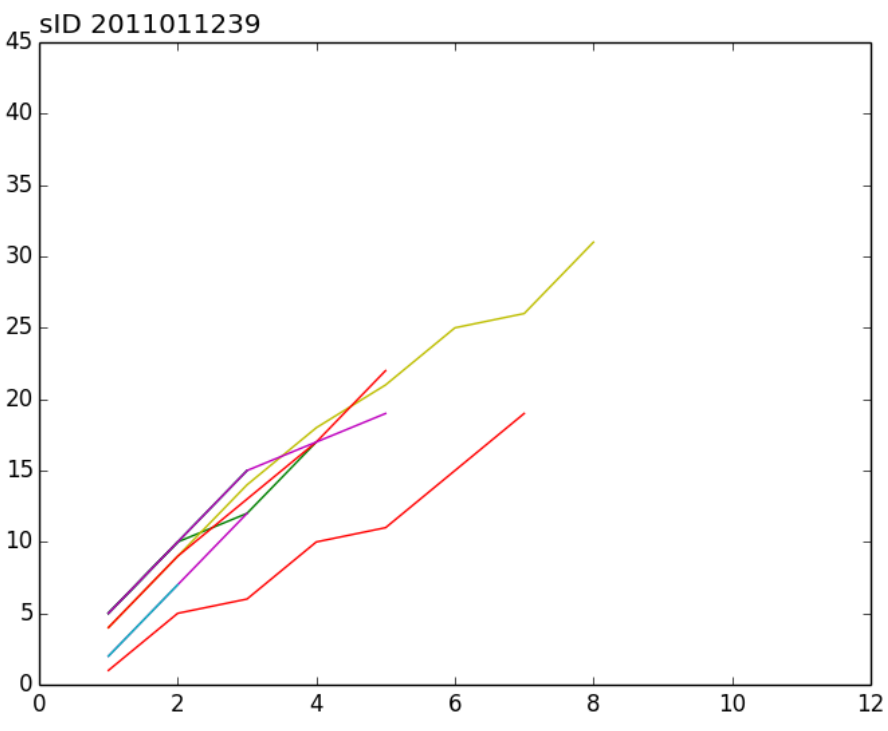
其三，按照被试区分，左边为sCG，右边为sCG/#queries

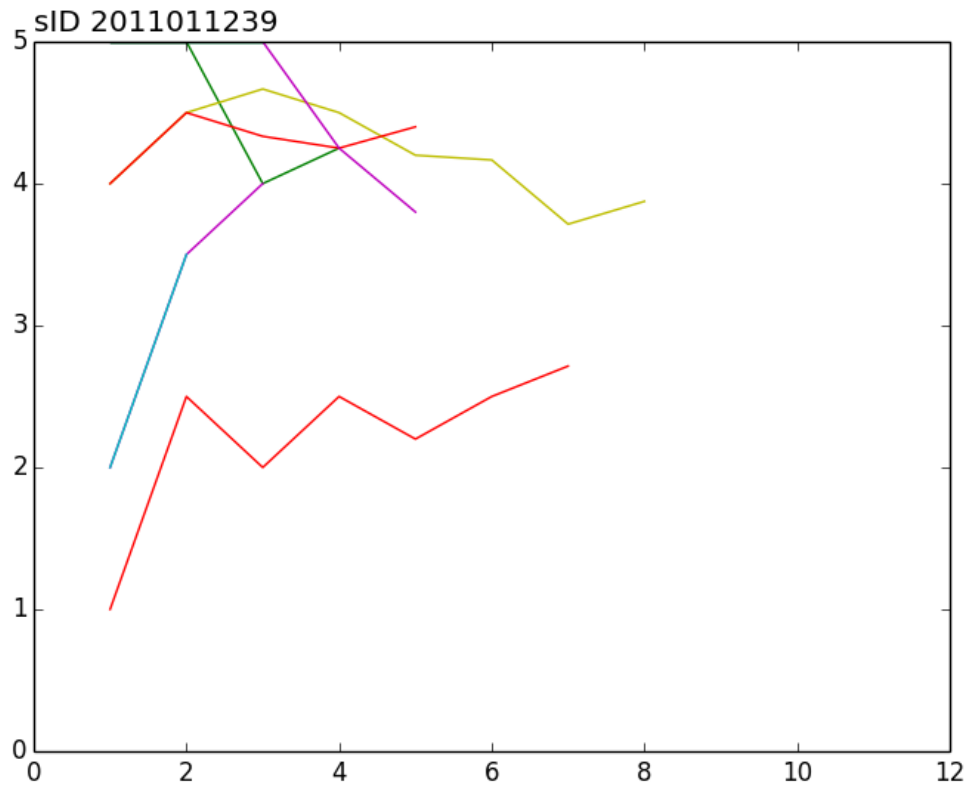






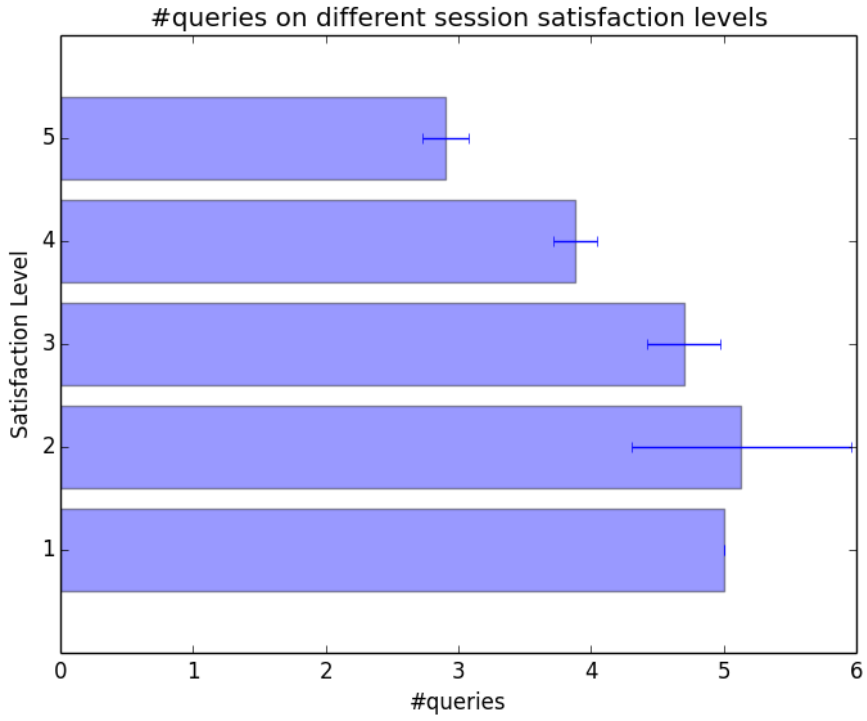




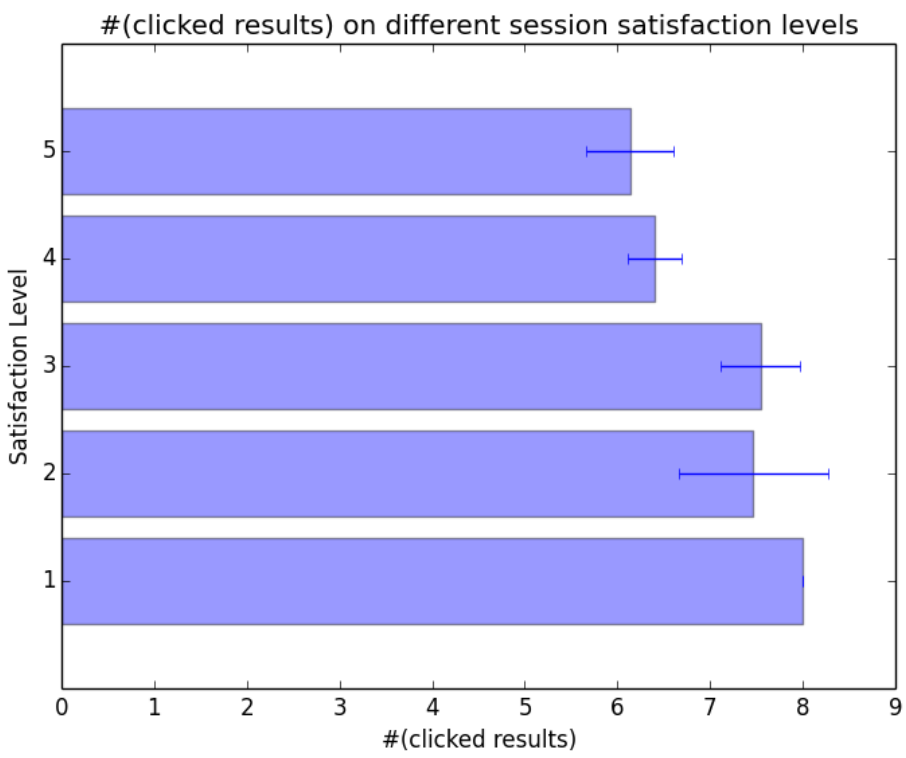


接下来，我们统计了一下在不同的Satisfaction Level上的一些指标

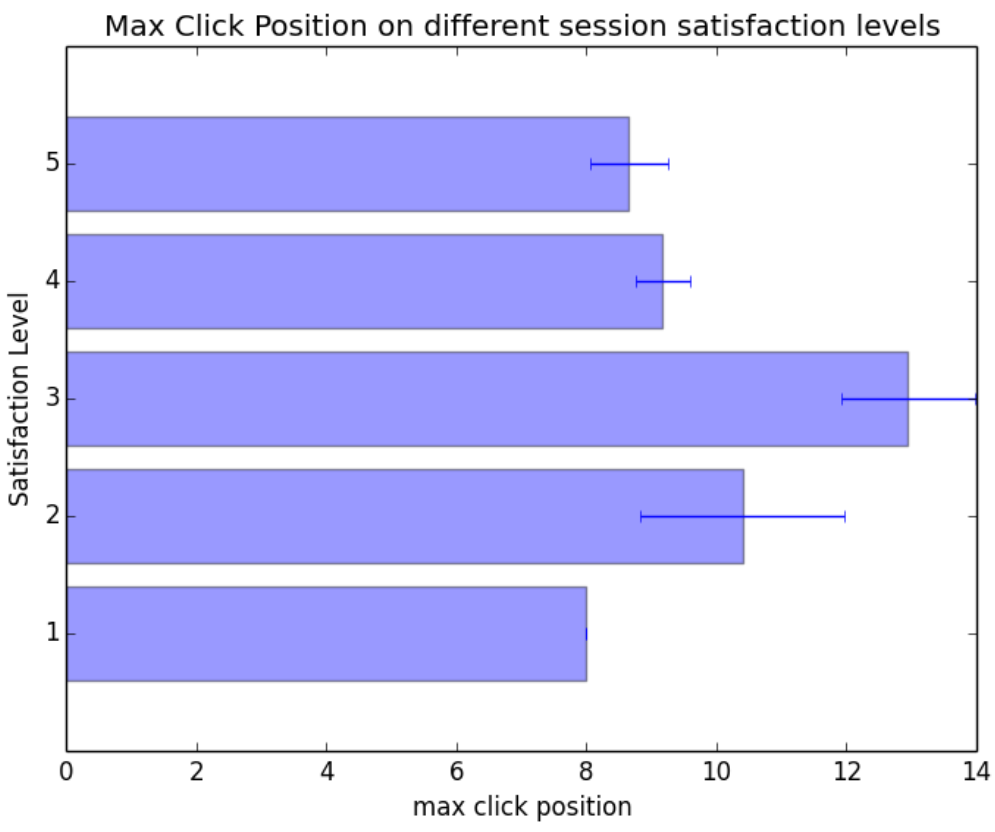
我们现在的Satisfaction Level是完全根据用户最后的标注统计的，这样的一个问题是，对于不同的Satisfaction Level上的session的数量实质上差距比较大。比如Satisfaction Level是1的session数量只有1个。
Jiepu Jiang的论文中，对于每一个session都有多个标注人员的打分，这样根据打分可以把Satisfaction分到4个水平上，并且保证每一个水平上有足够多的session，可以对结论进行显著性的检验。所以现在比较着急的是引入客观评价的部分，用打分的方式，总能把session分开。目前正在标注，这个标注比较费时间，差不多标注1个task要1个小时左右。



从左边来看，基本是满意度越高，session中对应的平均查询数量越少。Satisfaction 为1的时候，只有1个session，为2的时候方差比较大，也是因为session的数量比较少。



左边的图表示的是不同的满意度水平上，用户在整个session中点击的结果的数量。明显地，在满意度较高的session中，用户提交的查询更少。
这说明满意度和用户的effort是有关系的。

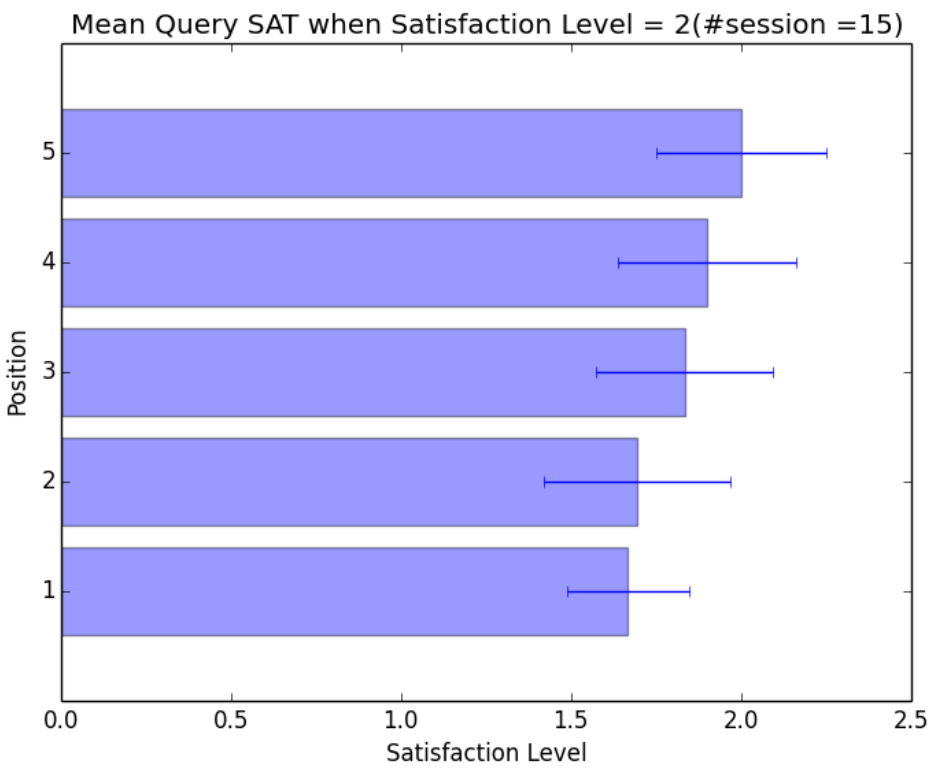


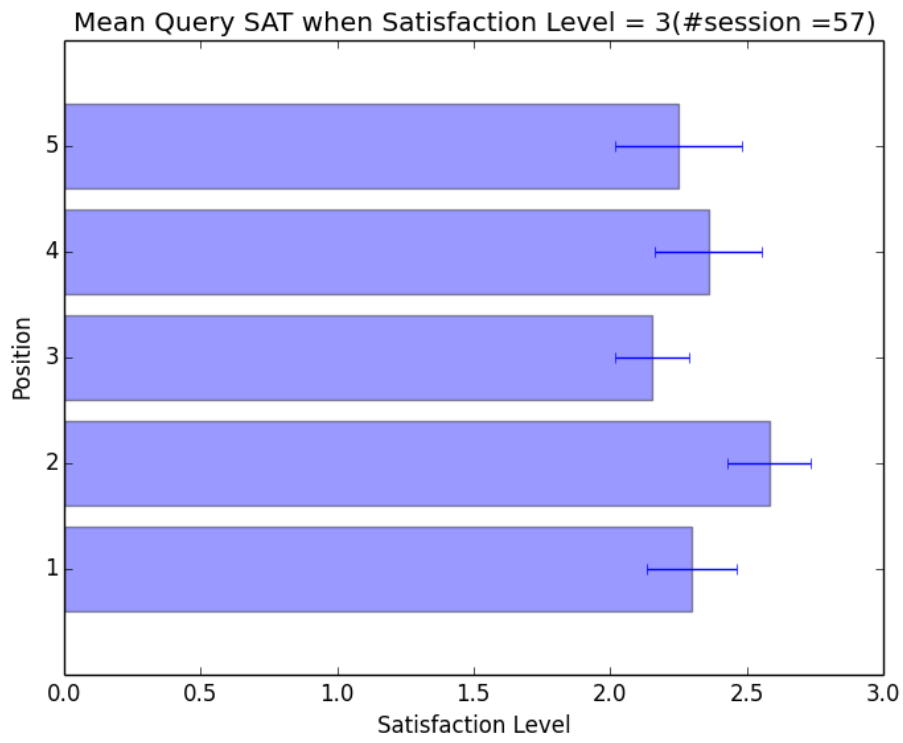
左边的图表示在不同的满意度水平上，整个session中某个query最深点击位置的关系，可以看到在session足够多的情况下，越满意的session点击地越浅。

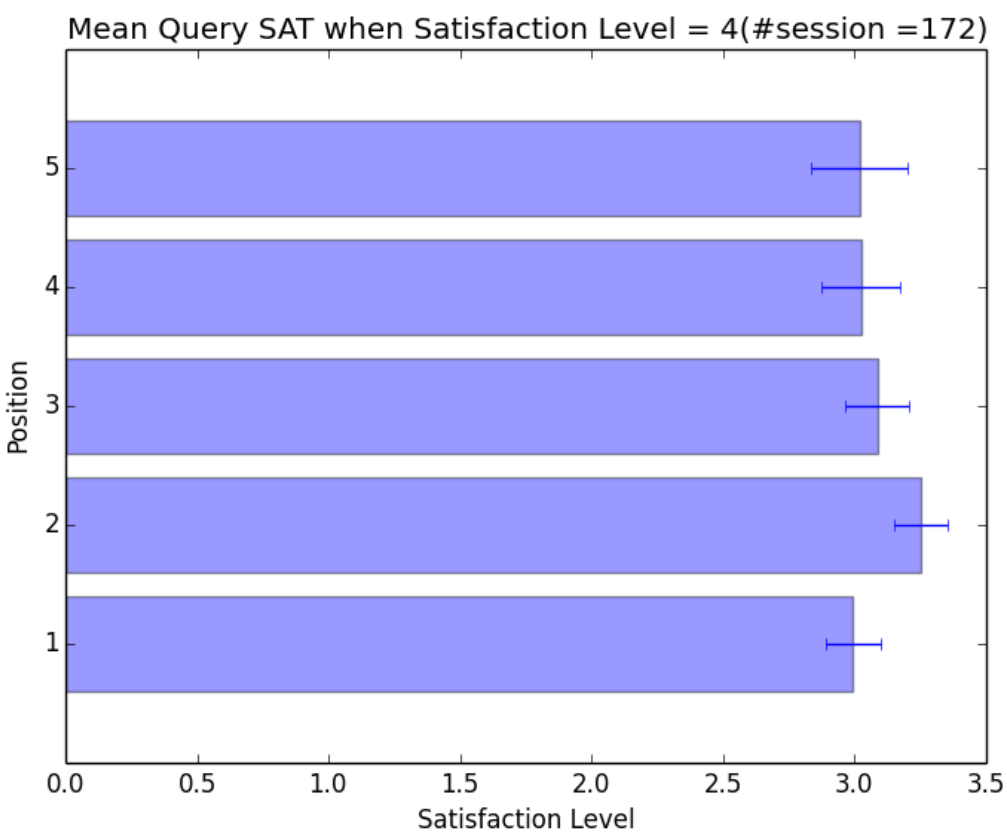
（刘老师提到这个不太有说服力）

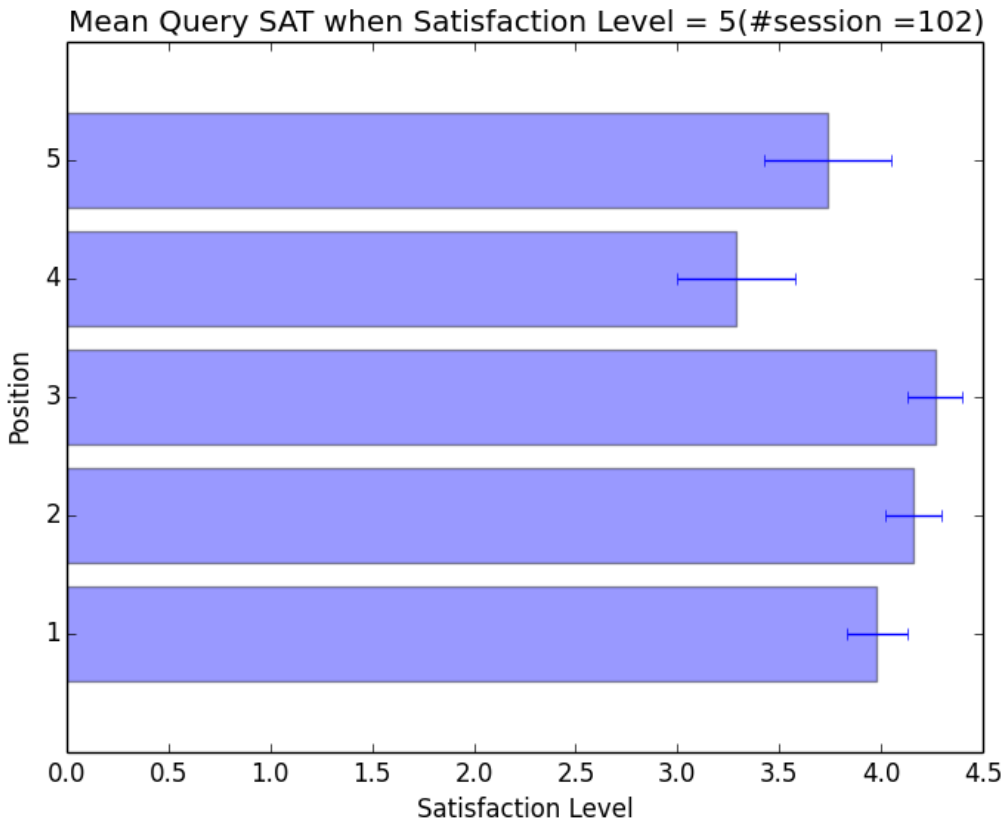
在不同的满意度水平上，随着query的增多，每一个query的Gain的变化，

下面四个图分别是Session Satisfaction = 5、4、3、2 的时候，在不同位置上的满意度的情况









从上面的这个图，可以看到：

1. 在Session Satisfaction = 2时，第一个query的满意度是比较差的；在满意度比较高的情况下Satisfaction Level = 4、5的时候，First Query的满意度是高一些。
2. 观察Query中满意度最高的那个查询，在Session Satisfaction = 2时，是最后一个查询，这样的情景大概是用户找了很久，终于找到了一个自己觉得不错的结果页，然后完成任务。

   在Session满意度比较高的时候，第二、第三个查询的满意度是整个session中平均最高的，这可能是我们给定的第一个查询用户觉得无法了解到全部信息，然后修改了查询，获得了更满意的结果。

3. 从四个满意度之间比较的话，趋势是比较一致的，综体上就是Session Level的满意度越高，Query Level的也越高。