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：

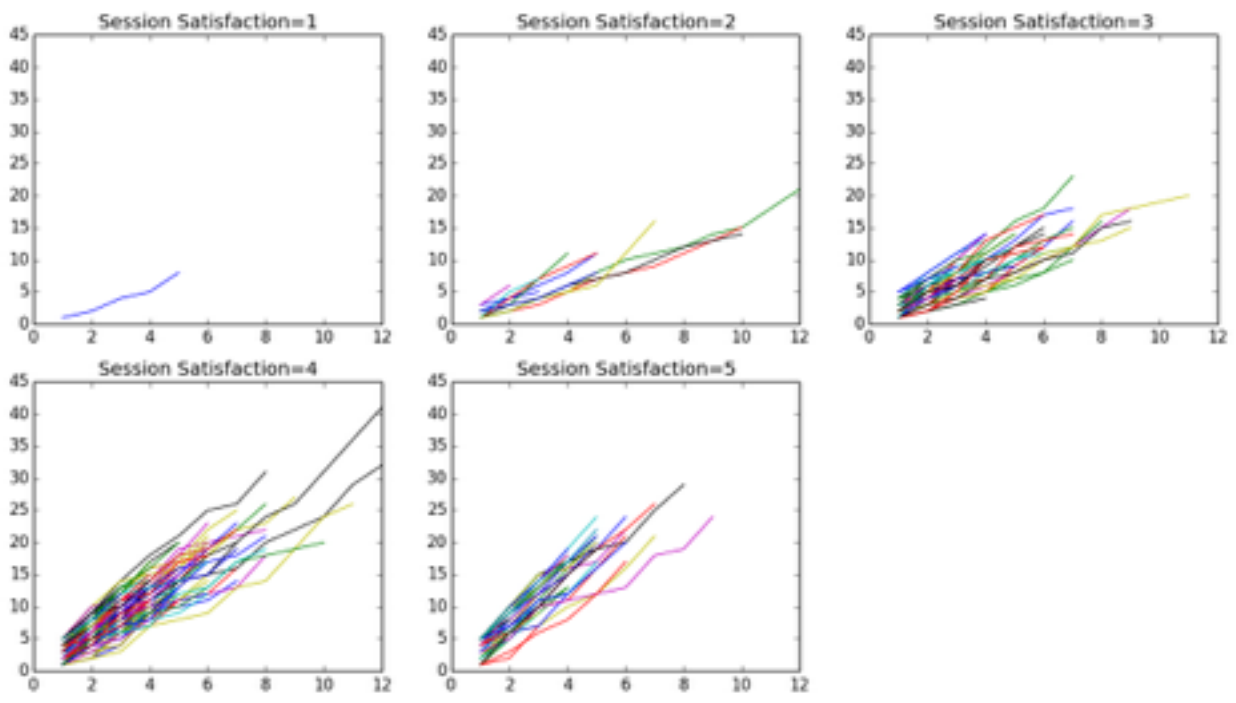
对于我们的实验和Jiepu Jiang的实验对比，在Search Outcome(sCG) 和Search Outcome/Effort (sCG/#queries) 上的趋势是一样的，都是Search Outcome要好一些。sDCG要稍微差一点，这个趋势也是正确的。

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。

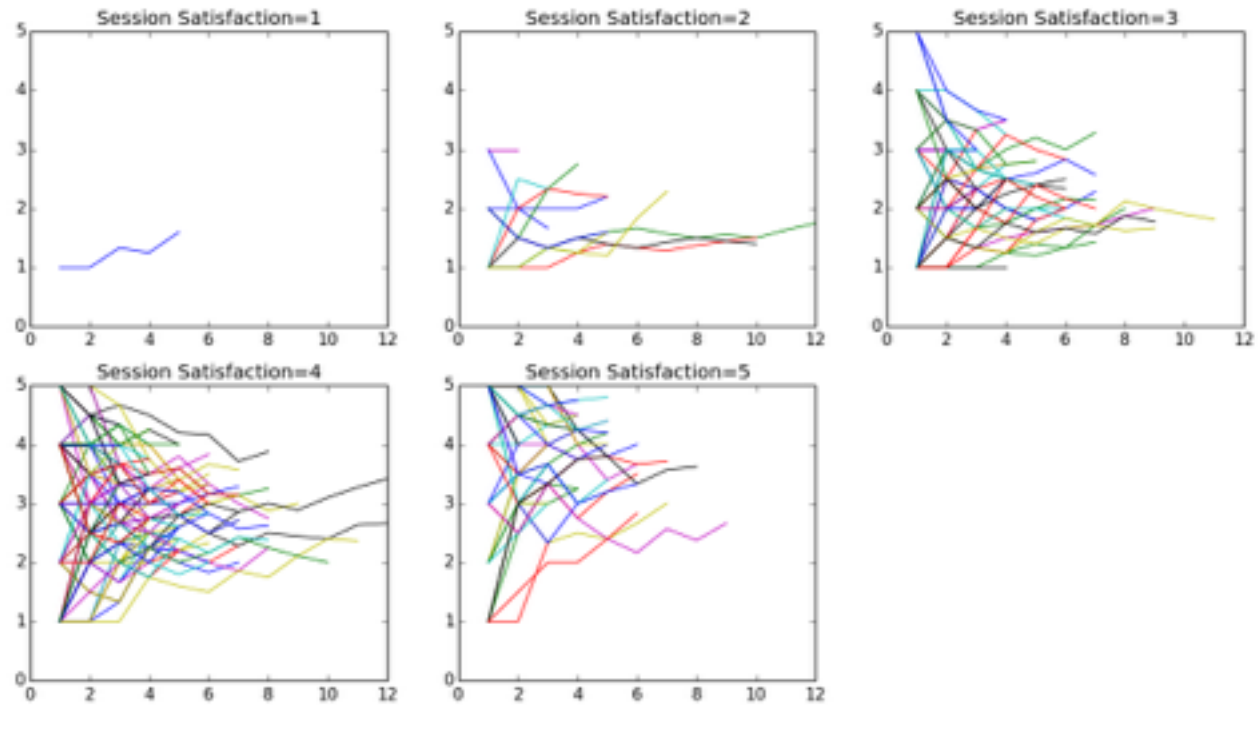
第二部分的实验，我们分析了sCG 和 sCG/#queries 两个指标随着query的数量增加的时候的变化，这里对query的满意度进行了一个预处理，在每一个人的维度上，将所有的query level/session level 的satisfaction归一化为z-score，

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

sCG

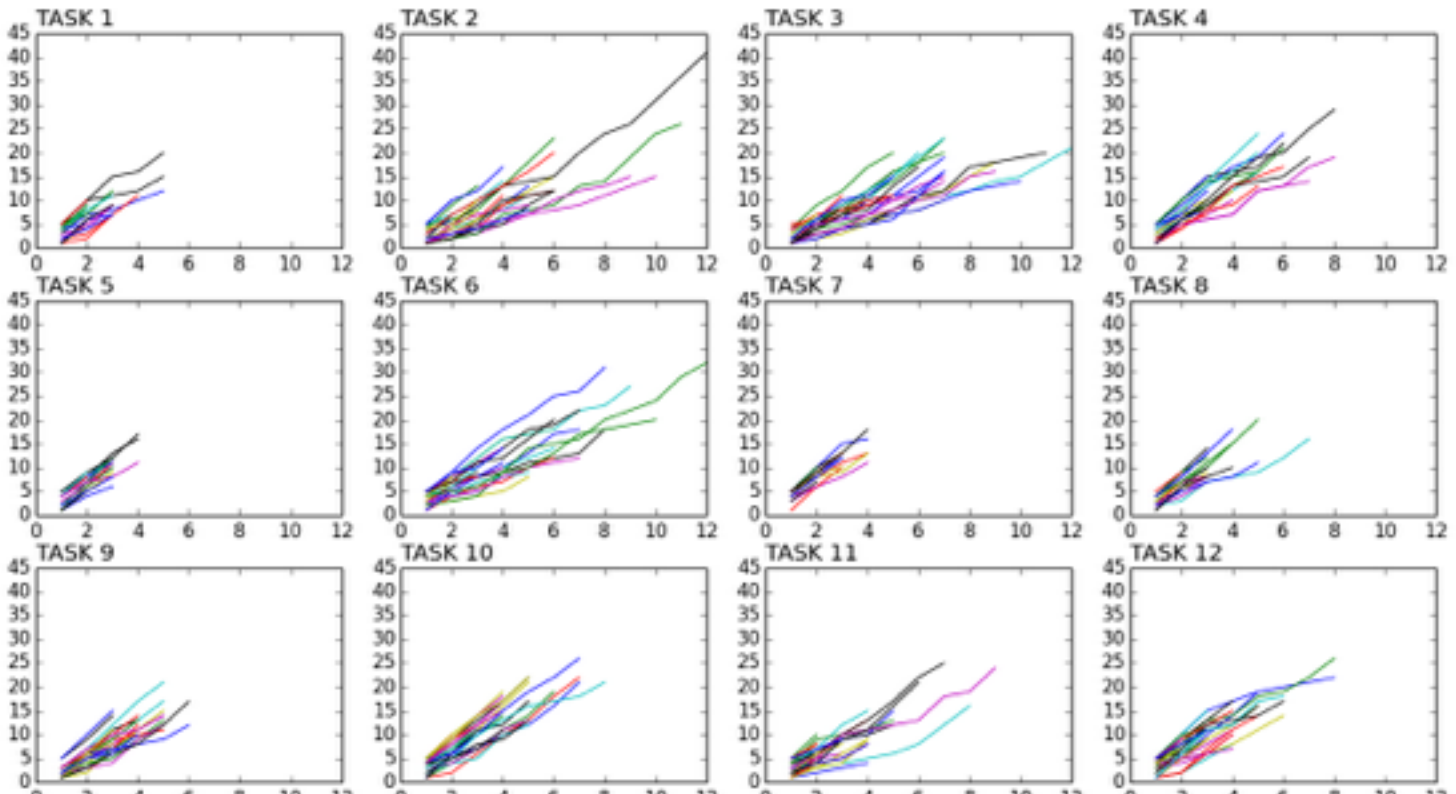


sCG/#queries

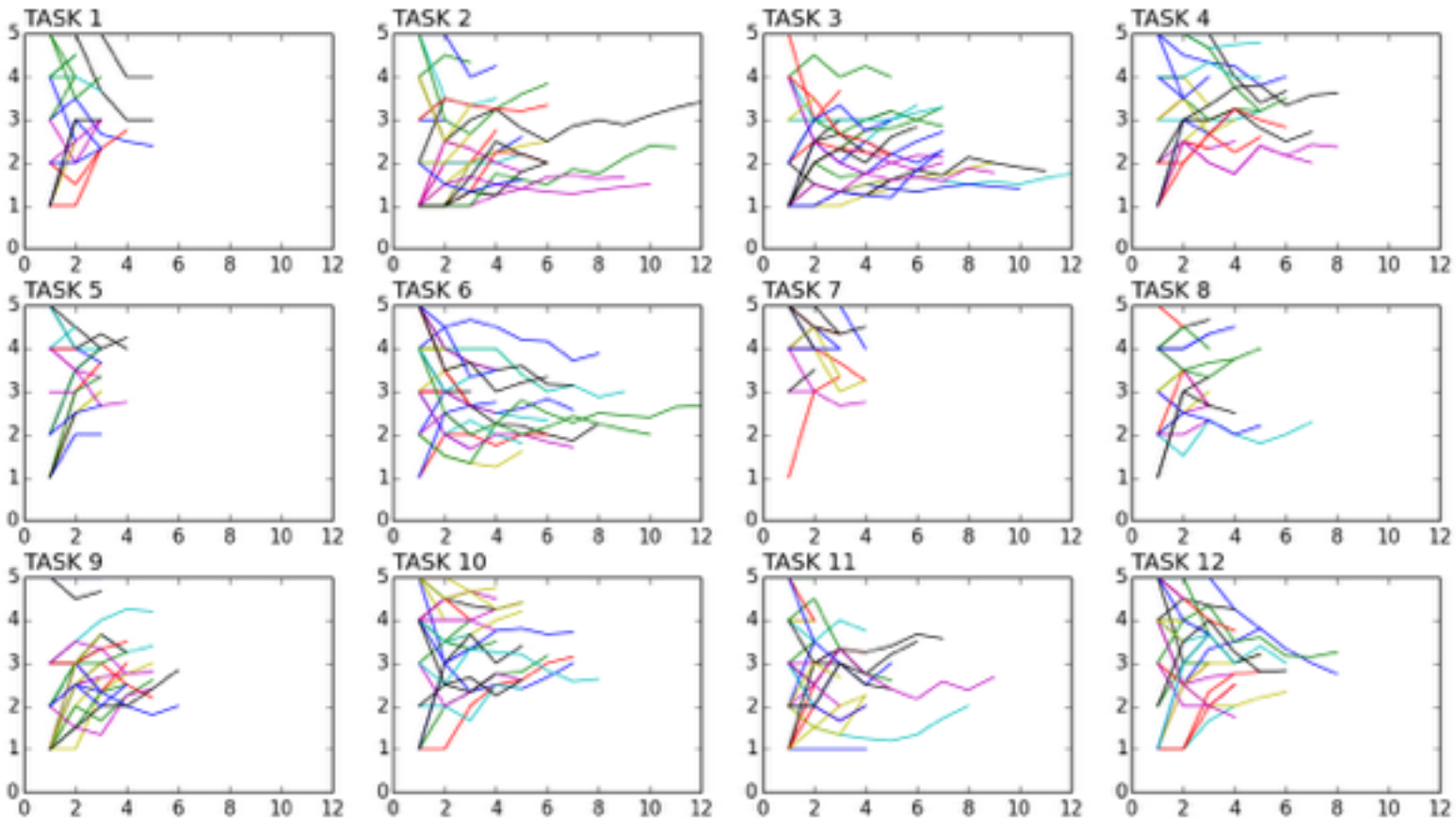


其次，按照1~12个任务区分：

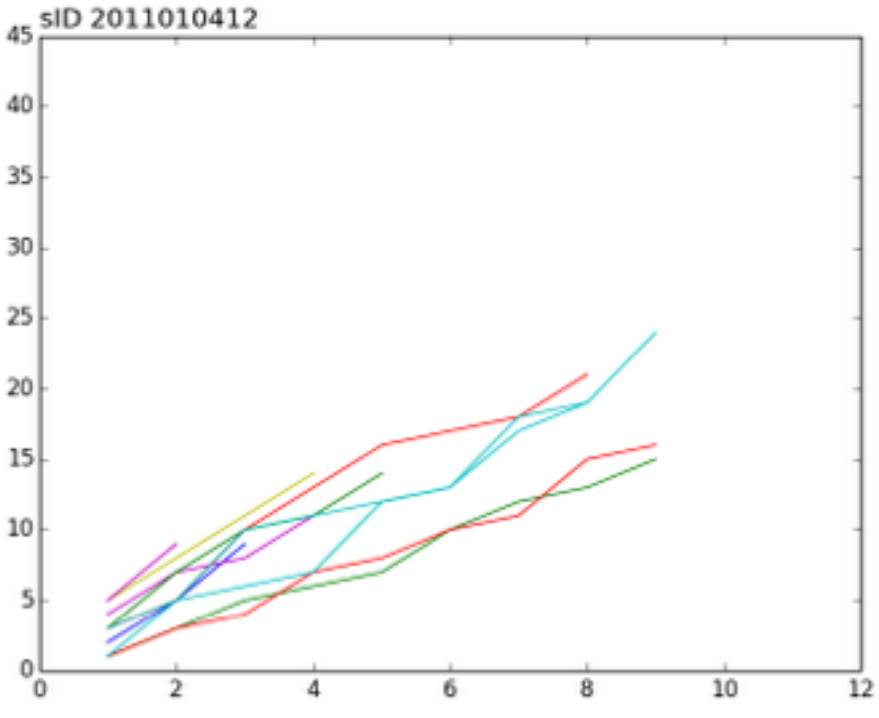
sCG

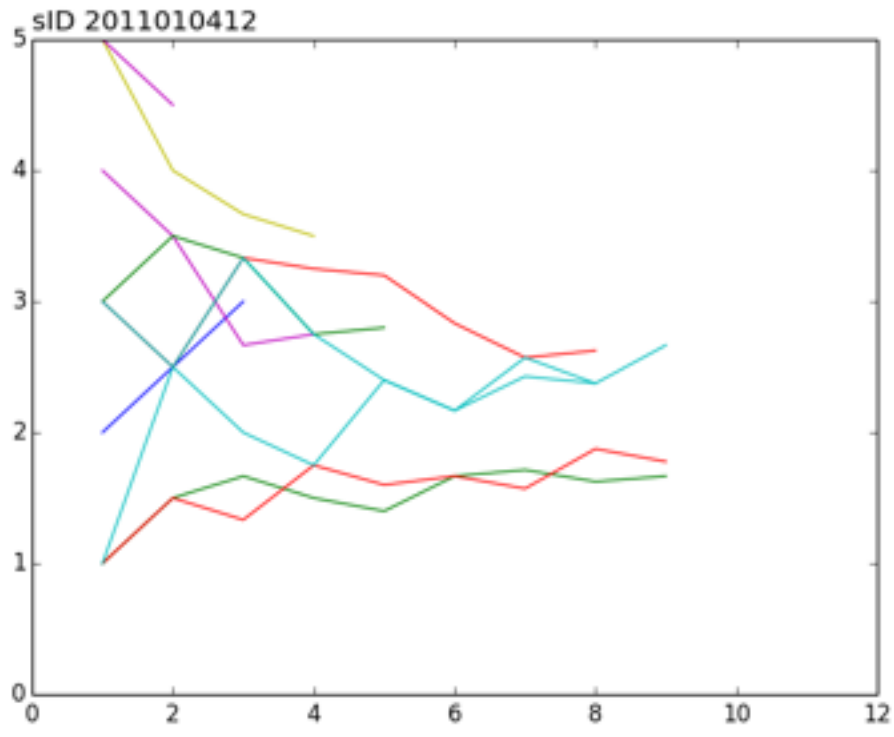


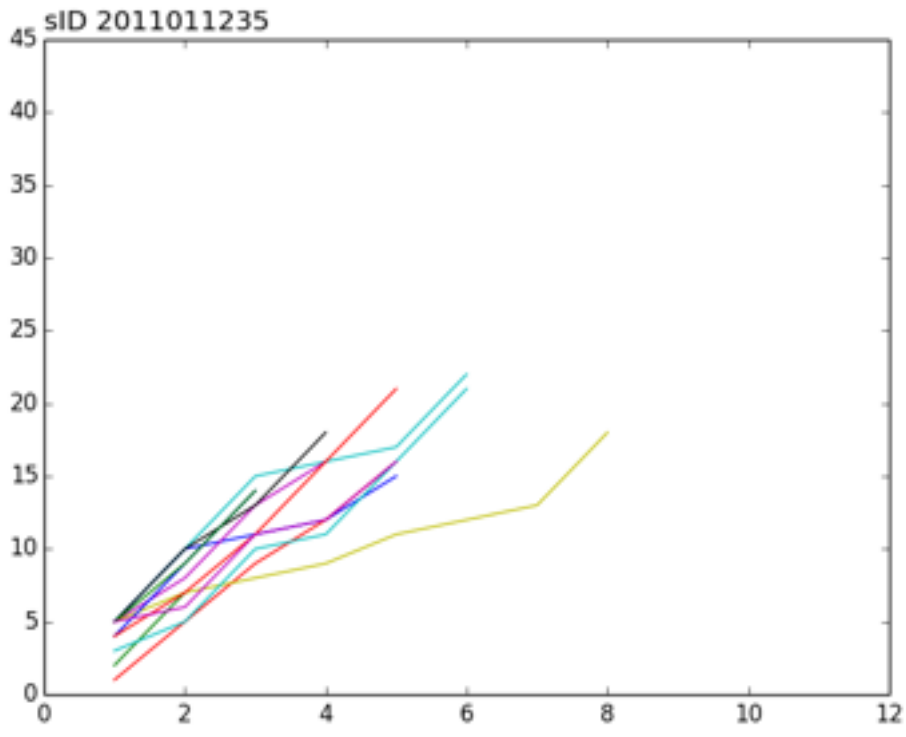
sCG/
#queries

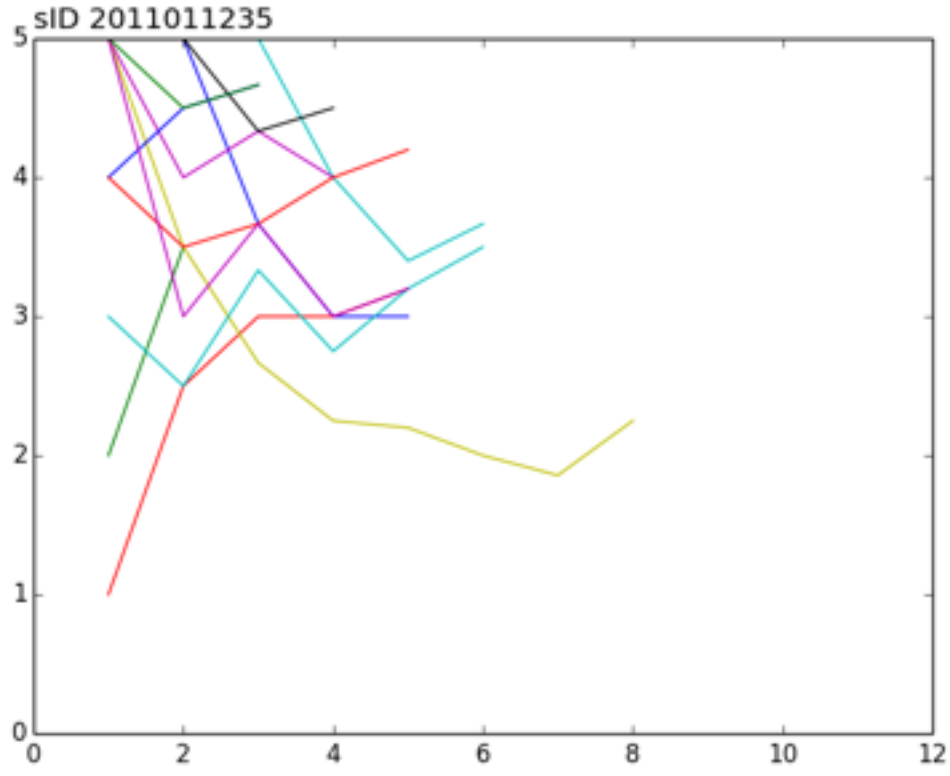


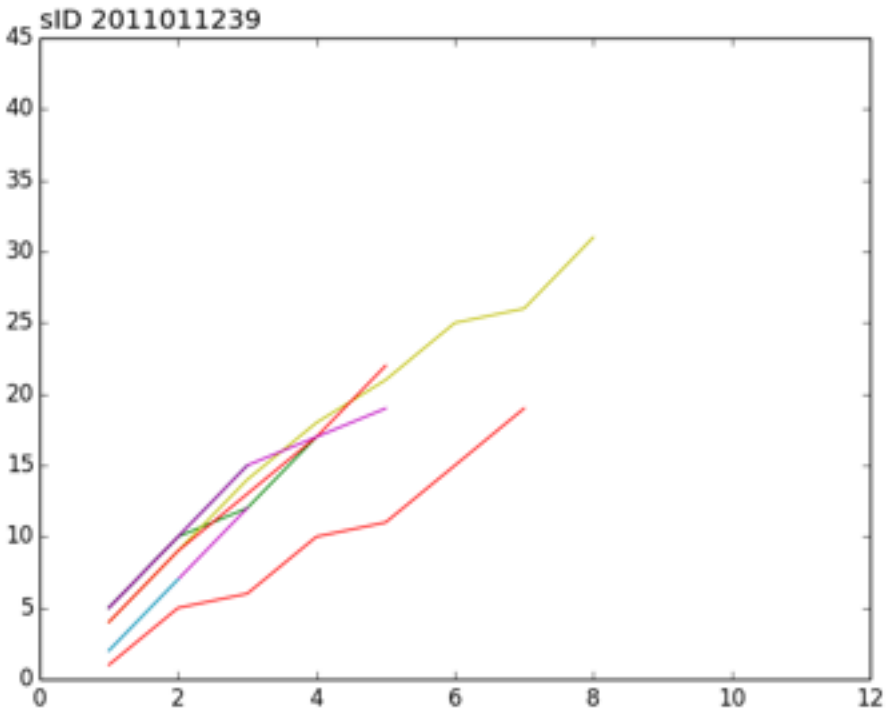
其三，按照被试区分，左边为sCG，右边为sCG/#queries

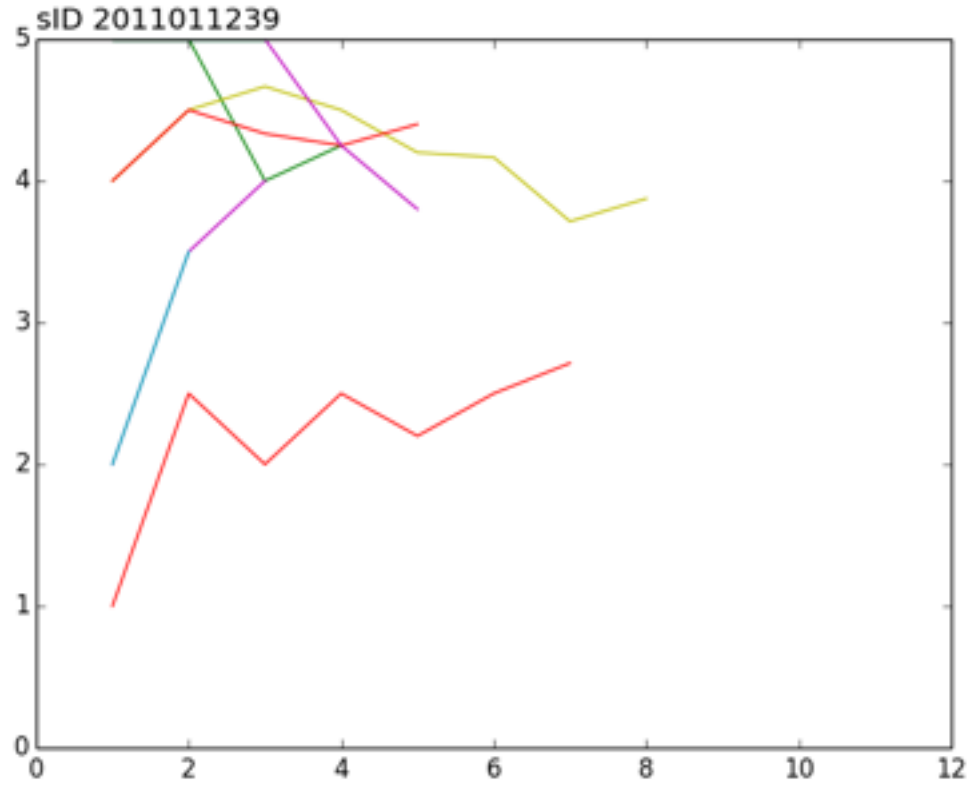






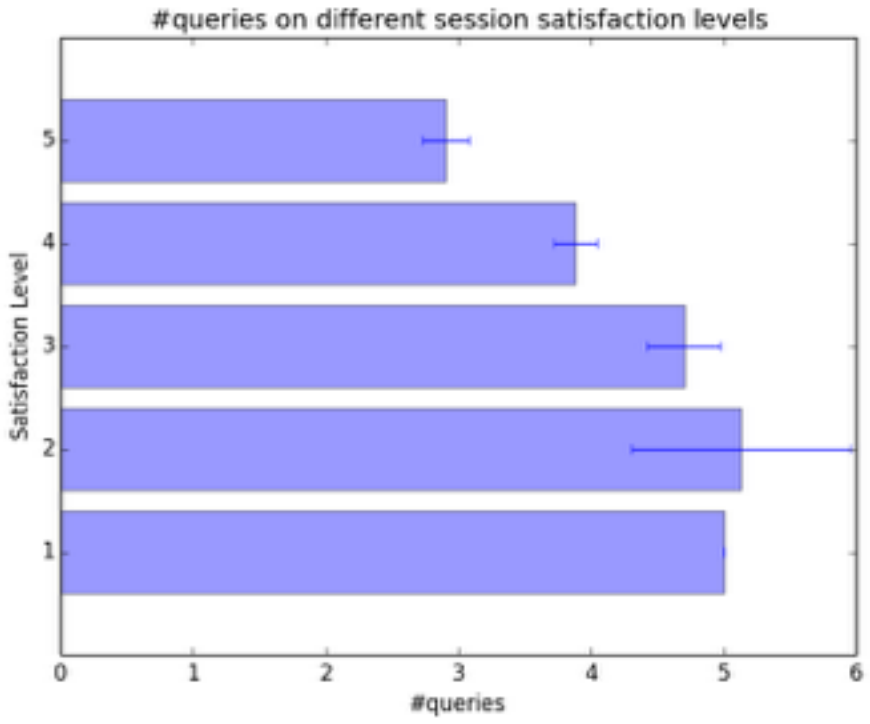




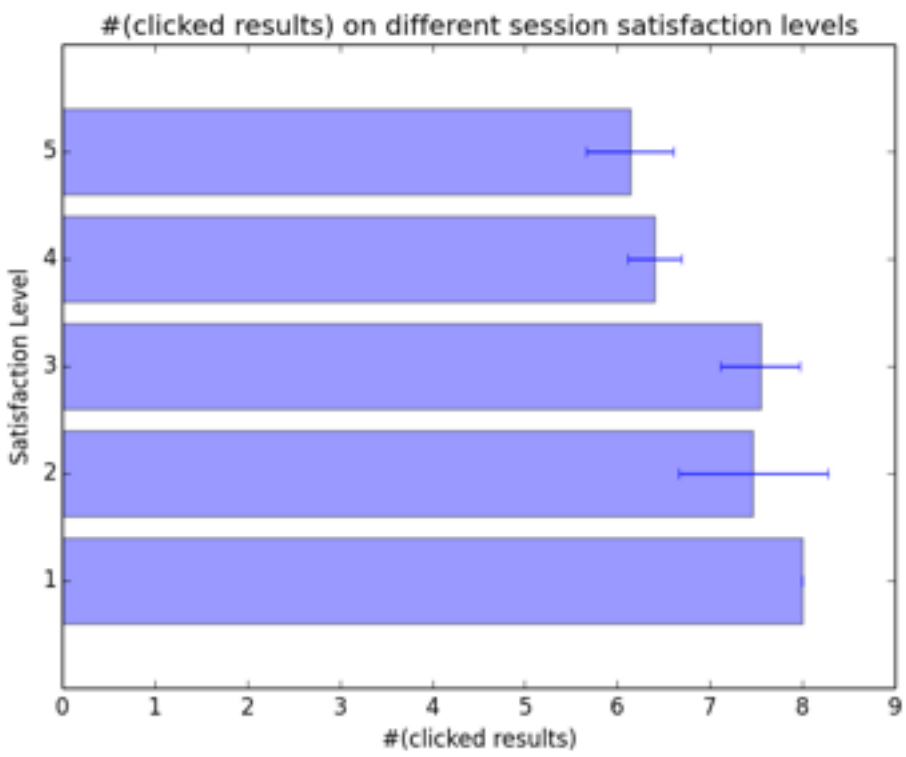


接下来，我们统计了一下在不同的Satisfaction Level上的一些指标

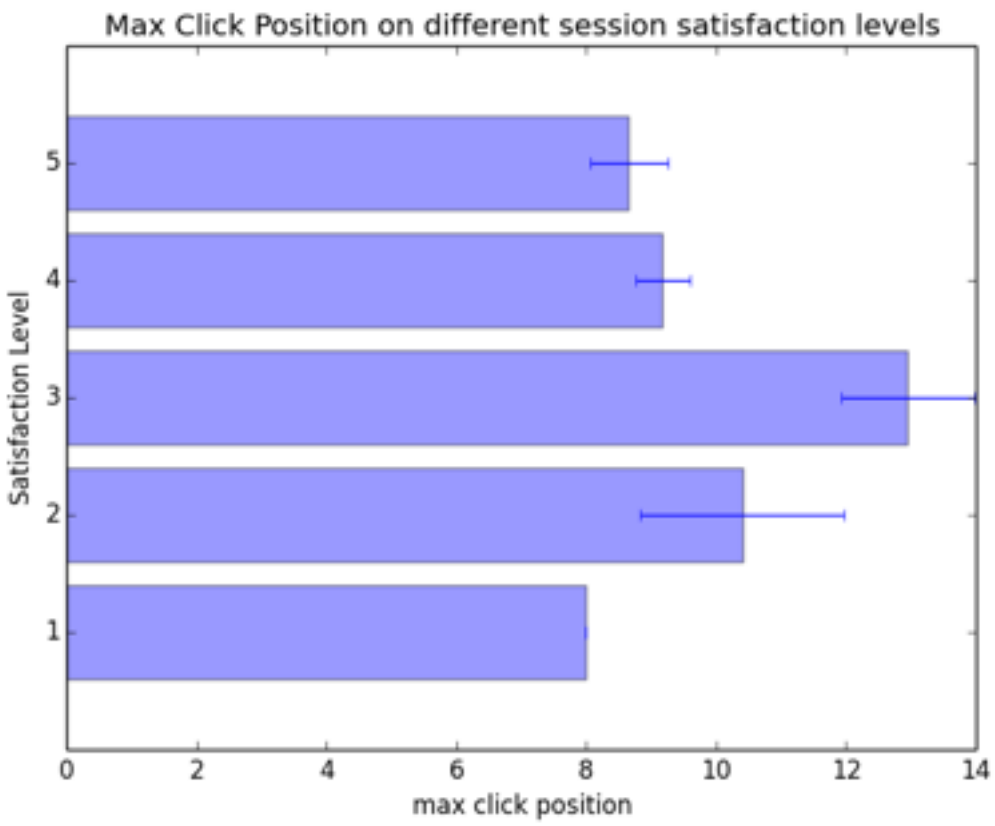
我们现在的Satisfaction Level是完全根据用户最后的标注统计的，这样的一个问题是，对于不同的Satisfaction Level上的session的数量实质上差距比较大。比如Satisfaction Level是1的session数量只有1个。
Jiepu Jiang的论文中，对于每一个session都有多个标注人员的打分，这样根据打分可以把Satisfaction分到4个水平上，并且保证每一个水平上有足够多的session，可以对结论进行显著性的检验。所以现在比较着急的是引入客观评价的部分，用打分的方式，总能把session分开。目前正在标注，这个标注比较费时间，差不多标注1个task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

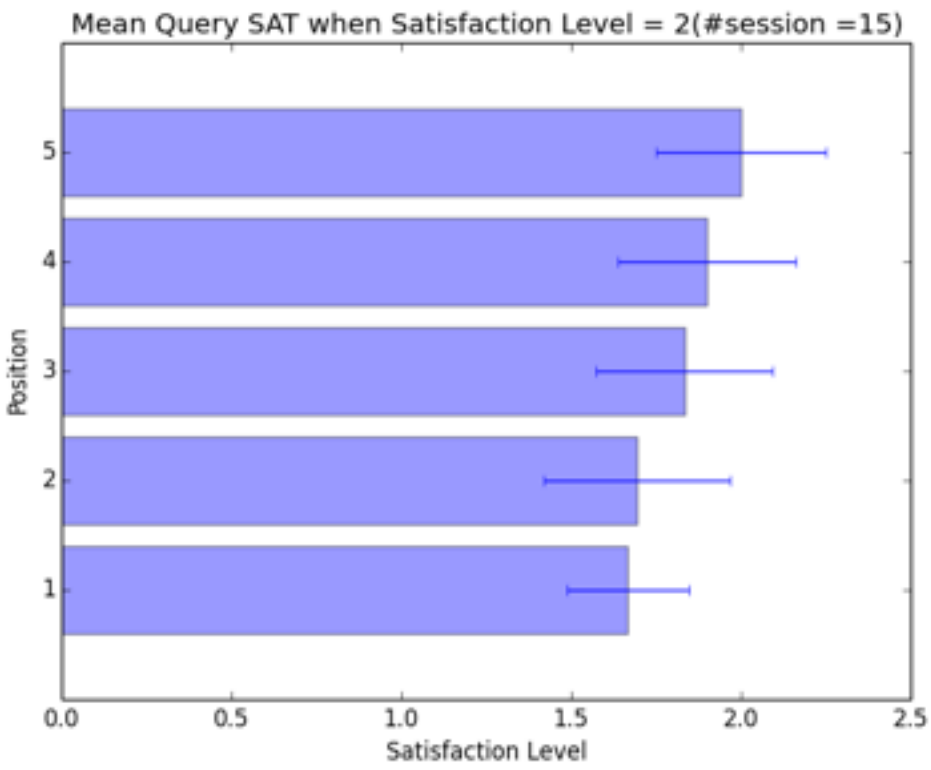


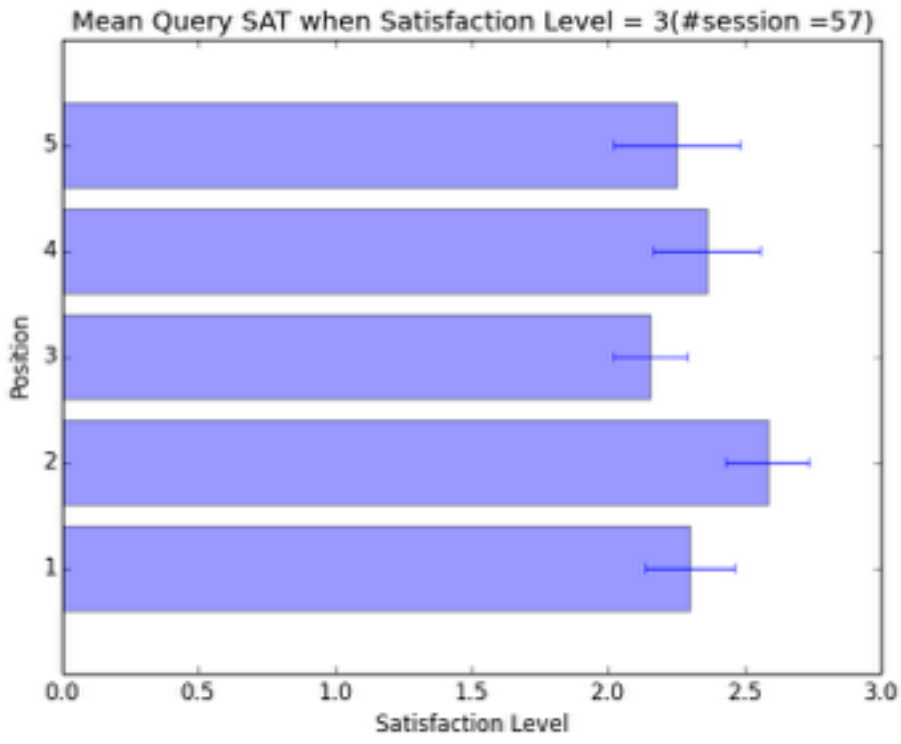
左边的图表示在不同的满意度水平上，整个session中某个query最深点击位置的关系，可以看到在session足够多的情况下，越满意的session点击地越浅。

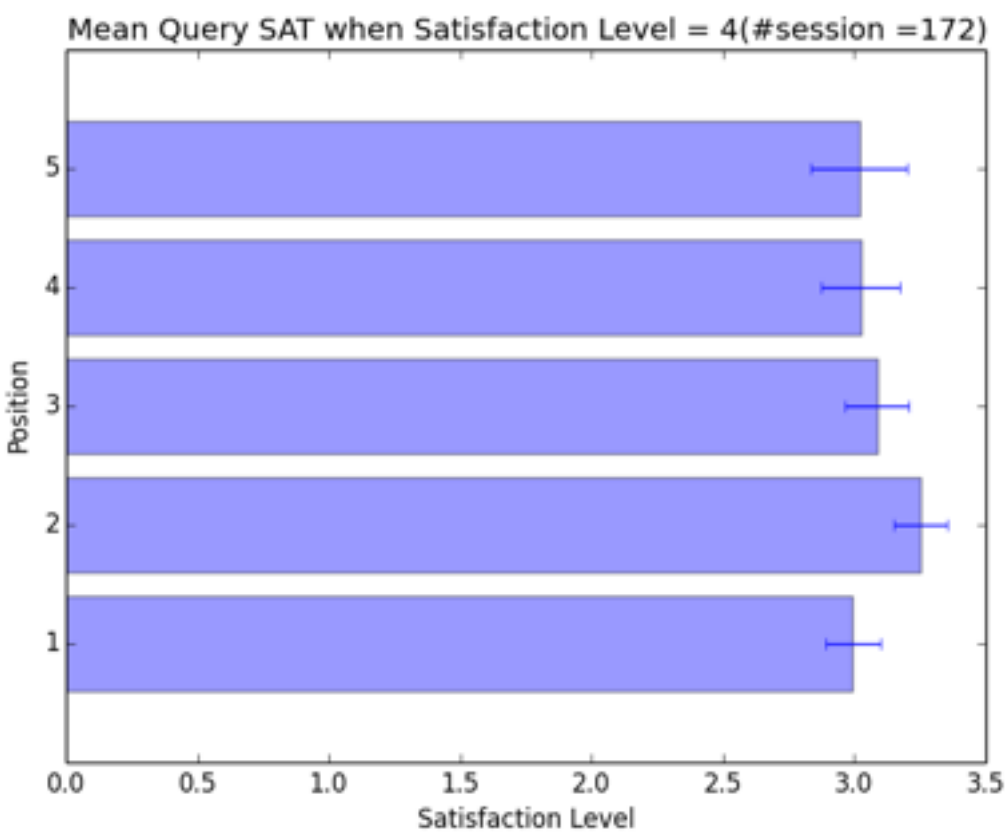
（刘老师提到这个不太有说服力）

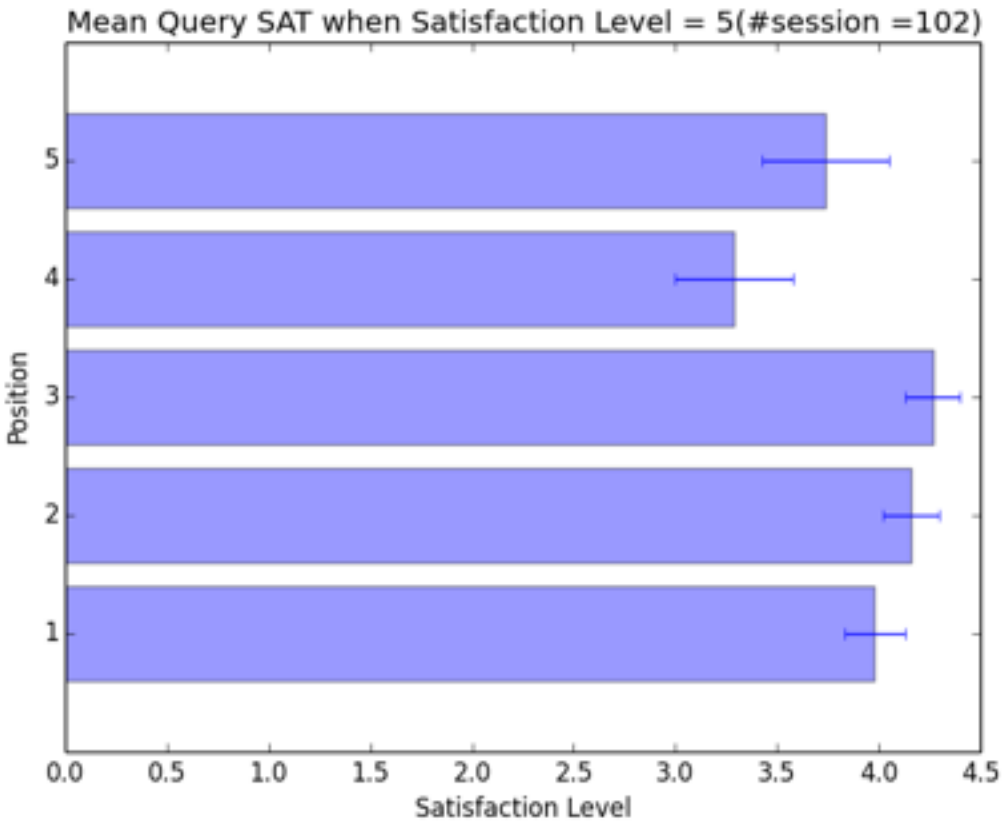
在不同的满意度水平上，随着query的增多，每一个query的Gain的变化，

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







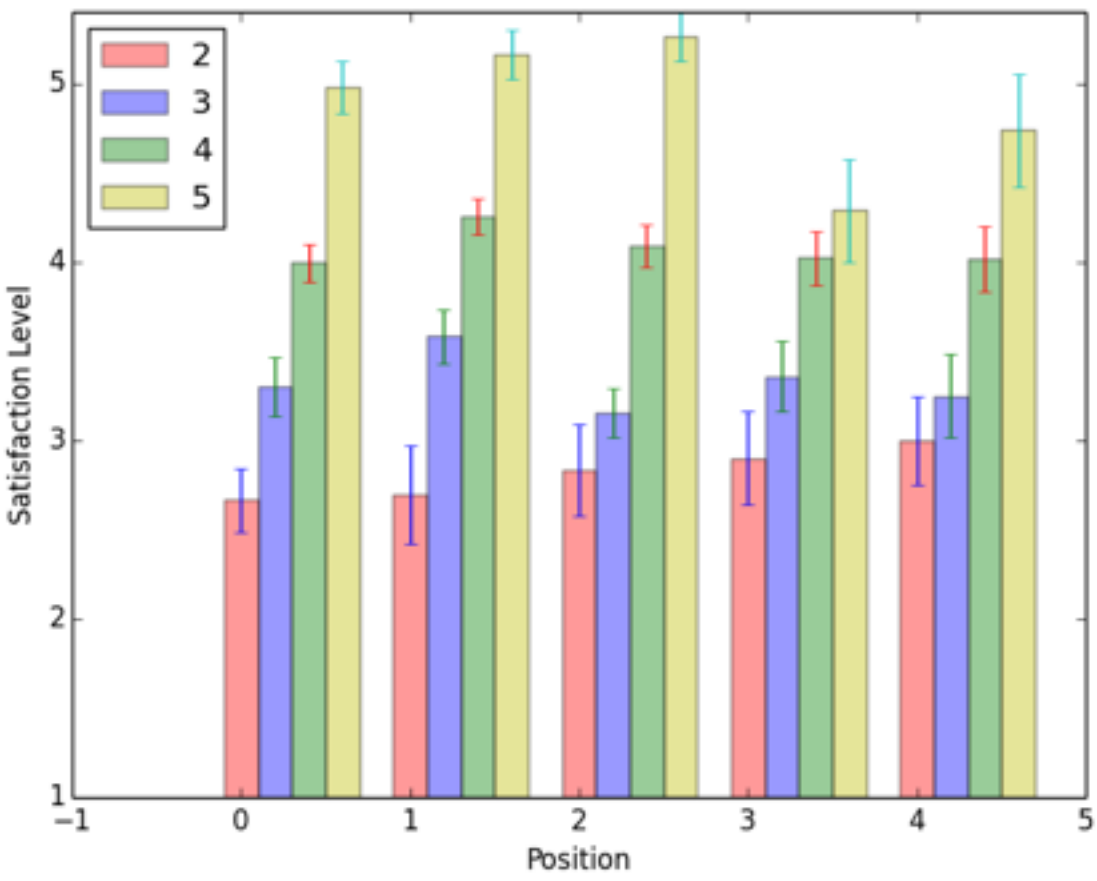


从上面的这个图，可以看到：

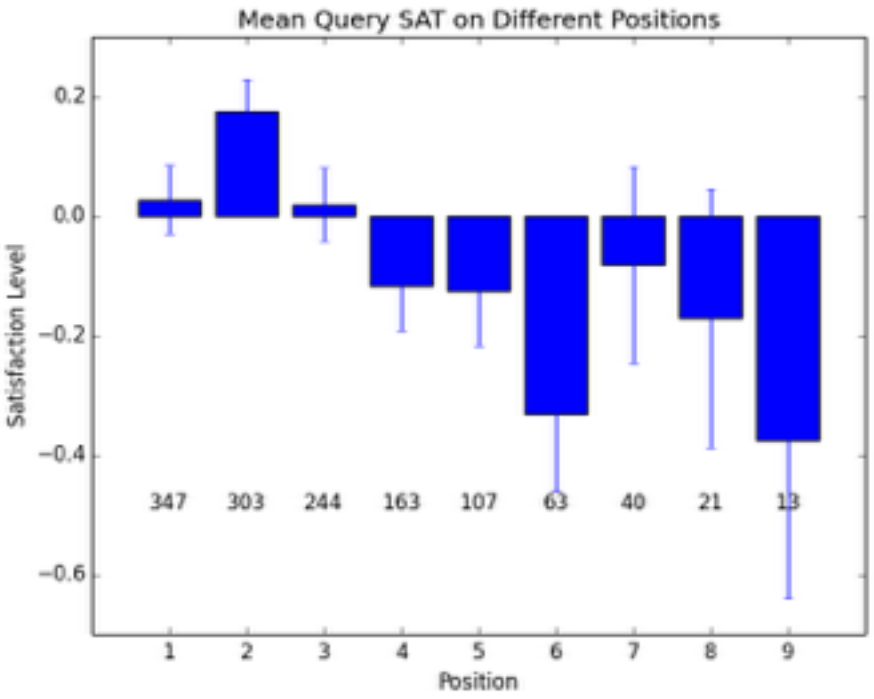
1. 在Session Satisfaction = 2时，第一个query的满意度是比较差的；在满意度比较高的情况下Satisfaction Level = 4、5的时候，First Query的满意度是高一些。
2. 观察Query中满意度最高的那个查询，在Session Satisfaction = 2时，是最后一个查询，这样的情景大概是用户找了很久，终于找到了一个自己觉得不错的结果页，然后完成任务。

   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。



不同位置上的满意度情况



不同位置上的满意度的情况，可以看到，用户最满意的是第二个查询，这可能因为第一个query是我们给定的，通过对第一个查询的结果浏览和检验，用户进一步明确了需求，进行查询改写，获得了更满意的结果。

（这里是否需要补充Dwell time/ Fixation 等指标？）

客观标注

对3个标注人员的标注进行了统计：

两两kappa, Kendall’s tau 值为：

Annotator 1	Annotator 2	Kappa	Kendall’s tau
2011011208	2011013986	0.526	0.346
2011011208	2222222222	0.610	0.462
2011013986	2222222222	0.573	0.425

4.2 Normalisation of Scores

As the ME process allows people to assign values in an unrestricted way, it is usual to normalise the scores using geometric averaging, to obtain responses on a comparable scale [5]. We adopt the approach recommended by McGee [8] as follows. If ℓ_{jd} is the logarithm of the score assigned by judge j to document d, and J is the total number of unique judges, then we compute the normalised score

$$s^*_{jd} = 10^{\ell_{jd}+\mu-\mu_j}$$

where

$$\mu_j = \frac{1}{4}\sum_{d=1}^{4} \ell_{jd} \quad \text{is the mean score used by judge } j, \text{ and}$$

$$\mu = \frac{1}{4J}\sum_{j=1}^{J}\sum_{d=1}^{4} \ell_{jd} \quad \text{is the over all mean score.}$$

All logarithms are base ten in this paper. Note that for clarity of presentation we use “Log Normalised” scores in the figures and tables, which is simply $\ell_{jd} + \mu - \mu_j$. This does not alter most of the analysis as it is done on ranks (non-parametric analysis). Where appropriate, we use s^*_{jd} for comparisons.

三个人的标注序列，两两计算Kappa还是比较高的，计算的方式采用了之前EVIA中的方法，在实际的计算中，我发现会出现这样的情况，三个人的打分的范围不太一样，这样的话，最后给出的得分会有1个人，得分总在100~1000之间，有一个人的得分在40~70之间，有一个人的得分在10左右。

暂时我还没用这部分数据，想设计一下怎么更好的结合三个人的标注。

Effort 和Session Satisfaction的相关性

	Kappa	Significance	Jiepu Jiang's	Kendall's tau	Significance	Jiepu Jiang's
Session Dwell time	**0.225295**	0.000022		0.151210	0.000026	
#Clicks	**-0.312504**	0.000000	-0.020000	-0.244483	0.000000	0.000000
#Queries	**-0.446187**	0.000000	-0.240000	-0.311255	0.000000	-0.230000
Average Query Length	**-0.244875**	0.000004	-0.140000	-0.177867	0.000001	-0.120000
Max click position	**-0.182740**	0.000613	-0.140000	-0.205219	0.000000	-0.120000
Avg click position	**-0.158544**	0.003019	-0.160000	-0.123420	0.000590	-0.120000
Min click position	**0.134081**	0.012295	-0.160000	0.133883	0.000193	-0.100000
Max Fixation Position	**-0.227937**	0.000018		-0.226942	0.000000	
Average Fixation Position	**-0.202821**	0.000139		-0.171125	0.000002	
Min Fixation Position	**nan**	1.000000		nan	nan	
Sum of Fixation Duration	**-0.377156**	0.000000		-0.300452	0.000000	
#Fixations	**-0.397376**	0.000000		-0.321542	0.000000	
Avg Click per Query	0.066550	0.215578	0.050000	0.031691	0.377596	0.040000

以上是各种Effort和Session Level的Satisfaction 的correlation 可以看到正相关最明显的是Session Dwell time，其次是最浅的点击位置；
负相关最明显的是#queries 这个和Jiepu Jiang的结论是一致的。

Dwell Time和#queries都是effort的一种刻画，不同的是Dwell time本身还代表了用户可能的效用；#queries则主要effort。

#Fixations 也是一个和满意度负相关的统计，这里的Fixation 统计了在SERP上，在各个搜索结果上的Fixation总数。