# Assessing Contextual Suggestion

Adriel Dean-Hall
University of Waterloo

Charles L. A. Clarke
University of Waterloo

## ABSTRACT

Assessment for the TREC Contextual Suggestion Track is unusual in that it depends on the personal preferences of assessors. During the initial phase of the track, assessors rate points-of-interests in a source city (Philadelphia for TREC 2013) in terms of their own interests. These rankings are distributed to participating groups, who are given about a month to generate point-of-interest suggestions for fifty other target cities around the United States, with personalized suggestions generated for each assessor on the basis of their ratings for the source city. These suggestions are then returned for rating, with each assessor rating their personalized suggestions for one or two of the target cites. Effectiveness scores (e.g., precision at rank 5) are then computed from these ratings. Unlike traditional TREC tasks, such as adhoc retrieval, it is not possible to measure assessor agreement, since each assessor is rating each point-in-interest in terms of their own personal preferences. Instead, we measure assessor consistency, which we define as an assessor's tendency to rank systems in the same order as other assessors. While consistency can be quite high for some assessors, and appears reasonable for most assessors, we have been unable to identify predictors of assessor consistency, including past consistency.

## 1. INTRODUCTION

The contextual suggestion track [1] imagines a traveler in a new city. Given a set of the traveler's preferences for places and activities in their home city, participating systems suggested points of interest in the new city. For example, given that the traveler likes the "Underground Garage and Bar" in Toronto, a system might suggest the "Arrow Bar" in New York.

TREC participants were given profiles for 562 assessors acting as potential travelers some of whom were twenty-something university students (62) and some of whom were American Mechanical Turk workers (500). Each profile indicates the assessor's opinion, on a 5-point scale from strongly uninterested to strongly interested, regarding fifty places in and around Philadelphia, PA. Track participants were also given fifty target cities (contexts) in the United States. For each profile-context pair, participating systems produced a ranked list of fifty suggestions tailored to the assessor's preferences and a target city. Each suggestion included the name of the attraction, a brief description, and a URL referencing a webpage that provided more details about the attraction.

In total, 34 runs were submitted as part of the contextual suggestion track in 2013. Suggestions from these runs were then returned to the assessors for judging. Judgements were made on the same 5-point scale as before. These ratings were given immediately after reading the description and title of the attraction and assessors were given an opportunity to change their judgement after visiting the associated webpage.

## 2. BACKGROUND

The majority of our judgements come from crowdsourced assessors. One of the major concerns we had with recruiting assessors from Mechanical Turk was the quality of responses. Vuurens and de Vries [4] found substantial improvements to the overall results of relevance judgements if they detected and removed spammy assessors. However, unlike the contextual suggestion track, they were working with objective questions that a group of assessors could work together to judge. In the contextual suggestion track subjective responses are required and having multiple assessors come to a consensus in the same way or measuring inter-assessor agreement is not possible.

Difallah et. al. [2] also look at spammers on crowdsourcing systems noting that there may be various forms of attack, including group attacks where some subset of the assessors work together and agree with each other. They also describe several techniques to filter spam and note that designing tasks that are undesirable to spammers is important. However they don't have any statistics on the amount of spam that can be expected. The contextual suggestion track has taken certain steps to prevent spam, for example timing assessors and asking assessors respond to known objective questions. However we are here interested in how many assessors responded with spammy ratings. Hoßfeld et. al. [3] also have similar advice for crowdsourcing with subjective questions where asking a known or objective question to assessors is advised.

EVIA'14, December, 2014, Tokyo, Japan.

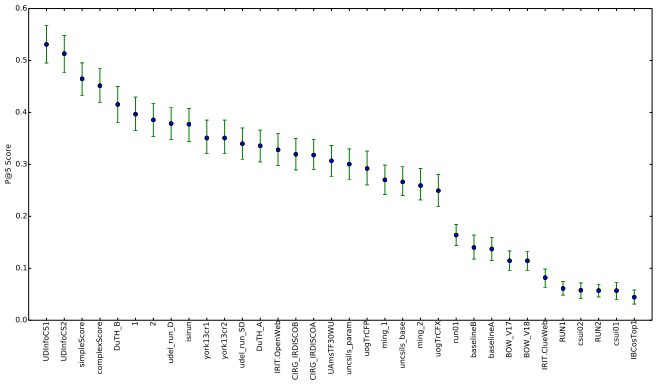



**Figure 1: P@5 score for each ranked system. 95% CI indicated.**

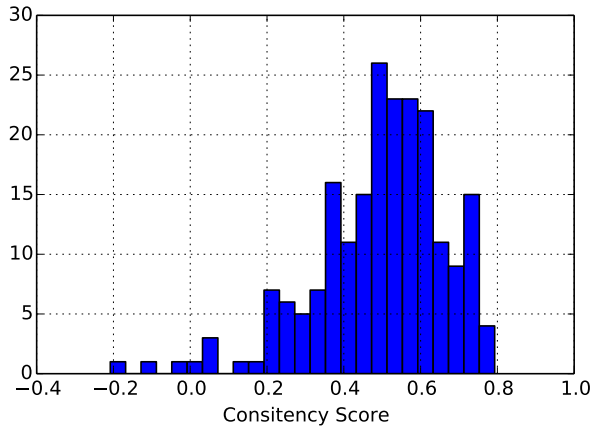



**Figure 2: Agreement of system rankings between profile-context pairs and all assessors.**

## 3. ASSESSORS

In 2013 the contextual track had 562 profiles for which track participants were expected to provide point of interest suggestions. In total, 136 assessors responded to an invitation to judge systems. The assessors were asked to either judge one profile-context pair or two. 59 assessors judged a single profile-context pair and 77 judged two. As can be seen in figure 1 the result of this judgement allowed us to rank systems where certain pairs of systems were significantly different from each other.

The task of judging is taxing on assessors, for each profile-context pair 170 suggestions (34 runs x 5 ranks) need to be rated on both how interesting the description of the attraction is and how interesting the website is. Additionally assessors were asked whether the suggestion was in the target city or not. This process takes time, assessors often took an hour to judge a single profile-context pair. Given the costs of judging it would be beneficial if we could identify careful assessors and only invite those assessors to judge systems.

### 3.1 Assessor Consistency

Ultimately the end goal of the whole judging process is to come up with a ranking of systems where the highest ranked system is the one that does the best overall job of providing attraction suggestions. The final ranking of systems is done by sorting the systems based on the mean P@5 score given to each system across all the judged profile-context pairs. Note that this technique could be used on any other metric besides P@5, however P@5 is the main metric used in the contextual suggestion track, so it is the metric that we use in this work.

Each assessor, on their own, can also order the systems. This is done by calculating the P@5 scores given to all the systems based on the judgements give for a single profile-context pair rather than all profile-context pairs. In general, we might expect careful assessors to order systems in roughly the same order as the aggregate of all assessors (assuming that some systems are, in fact, better than others). The closer an assessor is to matching the global ordering of systems the more *consistent* we say they are.

### 3.2 Consistency Score

In order to score assessors based on how consistent they are against the aggregate, we use Kendall's $\tau$ to compare scores given to systems based on judgements for a single profile-context pair to the scores given to systems based on judgements on all profile-context pairs. Note, however, that for a single profile only one of six P@5 values can be given to each system (0.0, 0.2, 0.4, 0.6, 0.8, or 1.0), producing many ties, which are handled in the standard way. We compute Kendall's $\tau$ as the number of concordant pairs minus the number of discordant pairs, divided by the maximum number of concordant pairs.

### 3.3 Assessor Comparisons

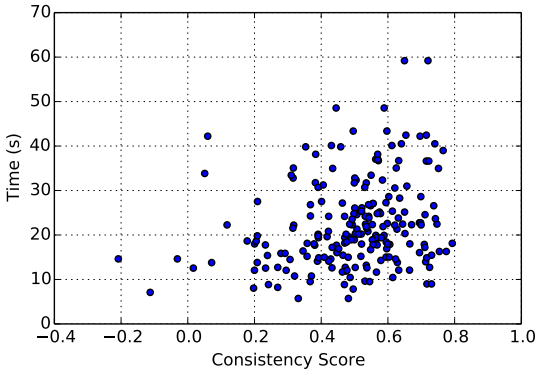
We can now take a look at the assessors to see how consistent they were. As we can see from figure 2 most assessors were fairly consistent, with the graph skewed to the right. The maximum $\tau$ score was 0.79 and the mean was 0.49. On the other hand, there were also a small number of assessors who were not consistent with the minimum being -0.2.

We can see here that there are some assessors who are dramatically more consistent with the group than the rest. Being able to identify which assessors are consistent early based upon the judgements of a single profile-context pair would be very helpful as we could then ask those assessors to judge more profile-context pairs rather than having the other assessors spent time judging. For the contextual suggestion track assessors were only asked to judge one or two profile-context pairs but if we can identify consistent assessors then they could be asked to judge even more attractions.

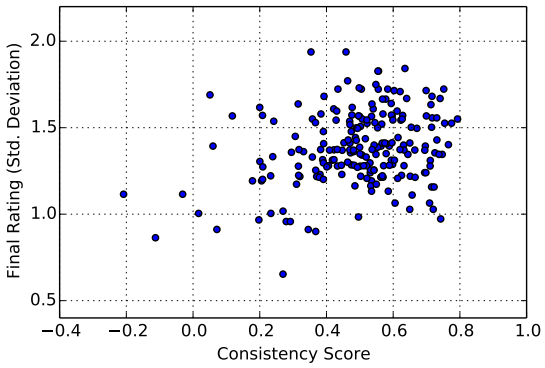
### 3.4 Assessor Attributes

In order to identify which assessors are consistent, we can look at a few different features and compare them with our consistency score. One such feature is the time it takes to make judgements. Some assessors might not be carefully making judgements, they might be very quickly making their judgments and hence be less consistent. Figure 3a shows that this doesn't seem to be the case. Although there is quite a range of times that assessors take to make judgements there doesn't seem to be any correlation to how long assessors take and their consistency.

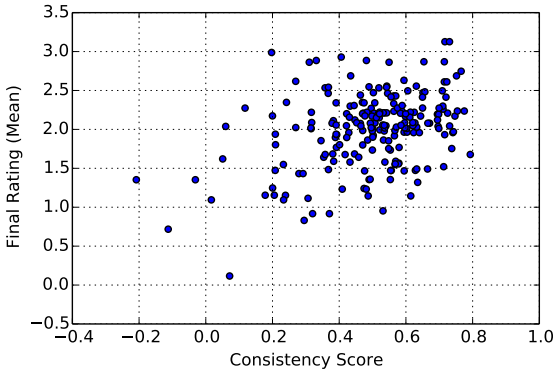
Instead of looking at time, we can look at the ratings the assessors provided. A assessor who always gives the same rating to every single attraction would not be very useful for ranking systems. We can look at the variance in the ratings



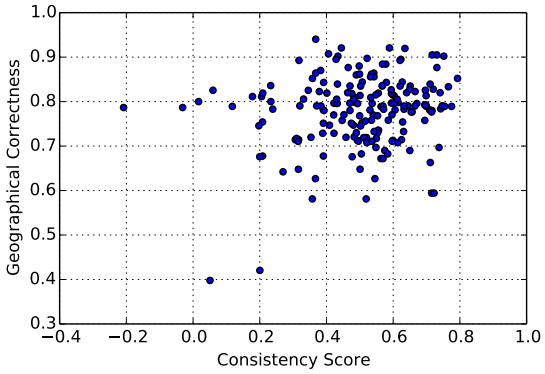

**(a) Time (s) to make judgements vs assessor consistency. Note that judgements that took more than 100s to make were removed.**




**(b) Standard deviation of final (website) rating vs assessor consistency.**




**(c) Mean of final (website) rating vs assessor consistency.**




**(d) Agreement of assessors with NIST assessments on geographical appropriateness vs agreement with all assessors.**

**Figure 3: Features against consistency score for every profile-context pair.**

provided by assessors and compare them to the consistency scores. We can see, in figure 3b the standard deviation of ratings vs. the consistency score and that again it does not seem to be a good indicator of assessor consistency.

Continuing to consider the ratings given by assessors we can look at the mean rating for each assessor. Some assessors might not like anything (so they would have a lower mean rating) and hence none of the systems are liked. On the other hand some assessors might have a high mean rating. Perhaps one group or the other is better suited to order systems? However, again this does not seem to be a useful metric, as seen in figure 3c.

In addition to rating attractions on interest level, assessors reported whether the attraction was in the target city or not. Whereas interest level is a subjective quality, geographic location is not: the attraction is either in the target city or it isn't. For the contextual suggestion track this objective question was answered by NIST assessors in addition to Mechanical Turk and student assessors. These NIST assessments can be used to compare against the other assessors to see whether they responded correctly. This provides another potential measure for assessors where assessors who were carefully rating attractions would have gotten more of the context judgements correct. Unfortunately, again, this is not reliable as an indicator of assessor consistency as can be seen in figure 3d.

There are other potential features that can be used to identify consistent assessors which yield similar results. If we look at population of the city being judged (where larger city might be easier to judge), the rating given based on the description (rather than the final rating given based on the website), the difference in consistency between Mechanical Turk or University students, or any other features we are equally unable to predict which assessors will be consistent.

## 3.5 Subgroups

As we have noted, some assessors are more consistent than others. For the assessors that are not consistent, two possibilities exist: 1) either the inconsistent assessors form one or more subgroups which are consistent with the rest of the subgroup or, 2) the inconsistent assessors are simply different from everybody else. In order to understand the behaviour of inconsistent assessors we compared every pair of profile-contexts pairs to each other against the average of the pair of profile-context pairs to the global ranking of systems. For example for a single pair of profile-context pairs: A and B we calculated the $\tau$ of the ranking of systems given by A vs. the ranking given by B, this value was plotted against

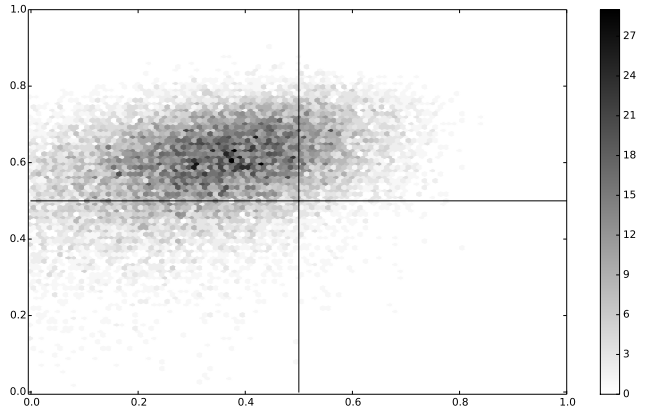


**Figure 4: Comparing pairs of assessors vs all assessors. Top-left: assessors disagree with each other but agree with all assessors (13 952); top-right: assessors agree with each other and all assessors (3324); bottom-left: assessors disagree with both each other and all assessors (4238); bottom-left: assessors agree with each other but disagree with all other assessors (222).**

the mean ranking of systems given by both A and B vs. the global ranking of systems. Figure 4 presents the result.

The figure divides the pairs into four groups. In the top-left group are pairs that agree with the group but not each other, this is the largest group and not surprising, because there are a few inconsistent assessors and assessors with varying degrees of consistency we expect to see many pairs here. The bottom-left group is the second largest, and consists of assessors that disagree with each other and the group, again it is not surprising to find that the inconsistent assessors don't agree with each other. The third-largest group is the top-right group which contains assessors that agree with both each other and the group. Again this is expected: this is simply pairs of higly consistent assessors. Finally the smallest group, in the bottom-right are assessors that agree with each other but not with the group. If the number of pairs in this group were large we could conclude that some subgroups had formed with groups of assessors with an inconsistent ranking from the whole group but consistent within the smaller subgroup. However this is a very small group of pairs so this does not appear to be the case.

For this analysis we have chosen the value 0.5 for the division into groups, which is a somewhat arbitrary choice. However the figure gives an indicate of what would happen if this threshold was altered to be higher and our conclusions would remain the same. Having a lower threshold for consistency is not really justifiable for a Kendall's tau based comparison.

## 3.6 Consistency

Finally we compare assessors with themselves. About half of the assessors who judged profile-context pairs judged two profile-context pairs, we now turn to these assessors and see if the assessors remain at about the same consistency level. In figure 5 we compare assessor's consistency score for the first vs the second profile-context pair that they judged. This consistency score was not calculated based on a profile-context pair ranking vs the global ranking but

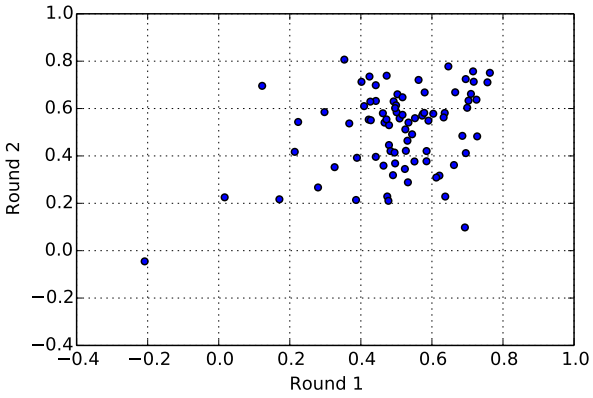



**Figure 5: Consistency score for the assessor's first judged profile-context pair vs their second profile-context pair.**

rather a profile-context pairs rankings vs the aggregate ranking of either all the first set of judgements or the second set of judgements.

We can see that even assessors themselves do not remain consistent from one profile-context pair to the next. Judging for the track took place over a 2-3 week period so most assessors should not have drastically changed their opinions on attractions. Even if somehow we had been able to identify the assessors who were consistent inviting them back for further rounds would not have guaranteed that the next set of judgements from that assessor was consistent.

# 4. CONCLUSIONS

In this paper we have explored consistency among assessors judging the contextual suggestion track. Our goal was the identification of careful and consistent assessors, allowing us to minimize assessment costs and improve assessment quality. Unfortunately, we were unable to find a method of reliably detecting consistent assessors. Moreover, assessors themselves don't remain consistent from context to context. However, despite this lack of consistency on the part of individual assessors, the group as a whole is able to identify significant differences between systems (figures 1 and 4).

# 5. REFERENCES

[1] A. Dean-Hall, C. L. A. Clarke, J. Kamps, P. Thomas, N. Simone, and E. Voorhees. Overview of the TREC 2013 contextual suggestion track. In *22nd Text REtrieval Conference*, Gaithersburg, Maryland, 2013.

[2] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, pages 26–30, 2012.

[3] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *Multimedia, IEEE*, 2013.

[4] J. B. Vuurens and A. P. de Vries. Obtaining high-quality relevance judgments using crowdsourcing. *Internet Computing, IEEE*, 16(5):20–27, 2012.