On the Relation Between Assessor's Agreement and Accuracy in Gamified Relevance Assessment

Olga Megorskaya Vladimir Kukushkin Pavel Serdyukov

Yandex
Moscow, Russia
{omegorskaya, wowone, pavser}@yandex-team.ru

ABSTRACT

Expert judgments (labels) are widely used in Information Retrieval for the purposes of search quality evaluation and machine learning. Setting up the process of collecting such judgments is a challenge of its own, and the maintenance of judgments quality is an extremely important part of the process. One of the possible ways of controlling the quality is monitoring inter-assessor agreement level. But does the agreement level really reflect the quality of assessor's judgments? Indeed, if a group of assessors comes to a consensus, to what extent should we trust their collective opinion? In this paper, we investigate, whether the agreement level can be used as a metric for estimating the quality of assessor's judgments, and provide recommendations for the design of judgments collection workflow. Namely, we estimate the correlation between assessors' accuracy and agreement in the scope of several workflow designs and investigate which specific workflow features influence the accuracy of judgments the most.

Keywords

Relevance labels; agreement vs. accuracy; judgments collection workflow.

1. INTRODUCTION

Expert judgments is an important resource used by search engines for the purposes of machine learning and search quality evaluation. Creating an environment for collecting assessors' judgments is a nontrivial task that faces the challenges of managing the quality of judgments and providing scalability and flexibility of the labeling system. However, expert judgments are subjective by their nature, and hence it is not obvious how their quality should be estimated.

One might calculate an assessor's *accuracy*: the share of documents judged by the assessor correctly, i.e. by assigning the correct relevance label. A label might be considered correct if it matches the one from the *golden set* (a set of documents assessed by a reliable and knowledgeable expert). However, creating a high quality golden set is a costly process, because such a set needs to be comprehensive and representative (i.e. proportional to the number of assessors and the assessments they generate) to fairly and realistically evaluate assessors' accuracy. The reason is that, in the case of professional assessors (opposed to crowd workers), we need to constantly monitor if the assessors are keeping the quality of their judgments on a high level, and hence filtering plain cheaters with a small and trivial golden set is usually not enough. Apart for the high costs of maintenance of the golden set, such approach also implies a risk of systematically biasing judgments by accepting opinions of just one or several persons as the ground truth.

Another metric - *agreement level* - is also often used to estimate the quality of judgments. Measuring agreement level is practically convenient and cheap, as it does not require any additional efforts/costs. But does agreement level really correlate with the overall quality of judgments (i.e. with their accuracy)? If it is so, one would be able to create an effective infrastructure for collecting judgments where the quality control is maintained by only monitoring inter-assessor agreement level, which makes the whole system more scalable and agile.

In our study, we propose a new gamified workflow of collecting judgments with some varied parameters in order to study different aspects that might influence the relation between the agreement and accuracy. The idea of the workflow is that assessors are able to make their judgments collectively, so, for each assessor, the evaluation process splits in two parts: making a single judgment and arrival to the final decision during a discussion within a group. In the case of the absence of a consensus, an additional assessor - *a referee* - is assigned to be responsible for the final judgment. The varied parameters of the workflow are: assessors' motivation, communication ability, the size of discussion groups (*overlap*) and the strategy of choosing a referee. We investigate how certain parameters of the proposed workflow affect the quality of judgments in terms of accuracy, agreement level and correlation between them. We assume that the most optimal parameters should provide us with the most accurate judgments by ensuring a strong correlation between the agreement and accuracy. Also, we are interested in some additional aspects of assessors' behaviour which could affect the system, such as how often assessors call for a referee, how productive discussions are, how quickly the experience spreads through a group of assessors.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767727.

We begin our study with the estimation of the baseline correlation between agreement and accuracy keeping all the parameters unadjusted. Then we search for the answers to the following research questions that are induced by a number of intuitions that we wanted to verify:

- **RQ1**: Is monitoring of agreement level a sufficient measure of quality control? **Intuition**: Supposedly, if assessors are only motivated to maximize their agreement level regardless of the actual quality of their judgments then, in the ultimate case, all assessors can agree on always making the same judgment, and the workflow will collapse, if no additional mechanisms of quality control are introduced;

- **RQ2**: Does communication between assessors help to increase the quality of judgments? **Intuition**: We expect that if assessors do not discuss disagreement cases between each other, it is harder for them to gain experience and to learn, so the quality of these judgments is expected to be lower;

- **RQ3**: Does a larger overlap provide better correlation between agreement level and accuracy? **Intuition**: We suppose that increasing of the overlap leads to increasing of the speed of information propagation, so that the accuracy, the agreement and the correlation between them will also increase.

- **RQ4**: Are moderators better than "common" referees? We tested two strategies of choosing a referee: a referee could be chosen among the regular assessors or among highly experienced assessors (*moderators*). **Intuition**: We expect that the cases when all assessors make different judgments are more difficult than others, and to define the final judgment one needs to have a higher level of expertise in evaluation.

In this paper, we observed that agreement level, under certain circumstances, can be used as a substitute for accuracy for measuring the quality of assessor's judgments. Based on the lessons learned from our study, we provide recommendations towards a design of a workflow that sets agreement level as the target for assessors and provides high quality of collected judgments.

The next section describes the related work. The default workflow design and the dataset are presented in Section 3. The experimental setup is presented in Section 3.4. Section 4 describes the set of metrics used in this study. The experimental results are presented in Section 5. Section 6 concludes the paper and describes a few directions for the future work.

2. RELATED WORK

Many studies refer to metrics of assessors' agreement and accuracy, analyze them independently of each other and investigate how these metrics are affected by various parameters, such as HIT design, task difficulty, assessors' motivation and qualification, etc [5, 2, 1, 8]. Some of the studies investigate similar aspects of evaluation workflows, but address them with a different purpose. For example, Kazai et al. in [5] show that the quality control is necessary for obtaining consistent judgments. They compared two groups of assessors: the first group was under severe judgment control conditions and the assessors from the second group were exposed to a smaller number of quality control mechanisms. It was shown that the quality of judgments of the second group was substantially lower. Other studies [4, 7] compared professional trained assessors with crowd workers in traditional relevance assessment tasks and showed that professional assessors have higher level of both accuracy and agreement. The studies of Alonso and Baeza-Yates [1], as well as of Kosinski et al. [8] show that a collective judges opinion is better than a single vote, and the accuracy of a consolidated output can be increased by increasing the number of participants per task. Finally, Kazai et al. introduced a gamified HIT design for collecting judgments for book search purposes on a crowdsourcing platform, where crowd workers were allowed to communicate with each other before they make the final decision [6]. They showed that the initial judgments made by the assessors could change as the result of a dialogue, but neither measured how these dialogues influenced the quality of judgments, nor provided any other details on how communication affected the evaluation workflow and its participants in general.

While our study continues the research on the optimal evaluation workflow designs, it focuses on a number of novel aspects. First, we investigate the correlation between the assessors' accuracy and agreement, and how it depends on specific parameters of a workflow. Second, we focus on the impact of discussions on assessors' accuracy, agreement, experience and the overall effectiveness of a workflow. Besides, we study such parameters of communication as the size of a discussion group and analyze how it affects the quality of judgments when they are produced collectively.

3. WORKFLOW DESIGN AND DATA

In this section, we describe the design of our set of experiments: its participants, the evaluation set preparation and the default workflow design. In order to not only get quantitative data, but also to observe the real process of assessors communication, as well as its evolution in time, we have organized a series of experiments in a form of an offline game. During each of 6 games, participants had to evaluate a given set of documents in the scope of a respective workflow.

3.1 Participants

98 participants took part in the experiments described in this paper. They were randomly assigned to one of 6 groups: one *control* group and five *experimental* groups. The control group took part in a "pure" default experiment: participants evaluated documents according to the general rules (see Section 3.3). For all experimental groups, the evaluation rules and conditions were modified in different ways (see Section 4). None of the participants knew ahead to which group they would be assigned.

All the participants have passed a standard job entrance exam that all assessors have to pass to get hired (at the search engine under study). Out of them, 16 were randomly selected from professional assessors who had already been working on their position for several years, and all of them were assigned to the same experimental group. All the others did not have any reported experience in being an assessor before.

3.2 Evaluation sets

We prepared a special dataset for evaluation. Several volunteers among our colleagues who reported to have expertise in some topic were asked to generate keywords which related to the topic of their expertise. We did not limit our volunteers in choosing the topics of their expertise, so, finally, we obtained a very diverse list of keywords related to one of the following topics: *antiviruses, table tennis, guitar, athletics, hiking, management, football, medicine, linguistics, alpine skiing, Internet-memes, chess* and *repair*. Then we extracted real user queries that contained the above-mentioned keywords from the logs of a commercial search engine, hence obtaining a set of real user queries related to the topics of our volunteers' expertise. We randomly selected 3 documents out of the top 10 search results generated by the commercial search engine for every query and asked our volunteers to evaluate them according to a standard 6-grade relevance scale similar to the one used at TREC (Web search track)[1]. Thus, we have collected 64 queries corresponding to 192 documents evaluated by the experts on the topics of those queries, and used this set as the golden set for estimating the accuracy of the judgments produced by other assessors. As far as our experiment was conducted completely offline, all the documents were printed out on a color printer, so that all the participants evaluated the same instances of the web documents.

3.3 Workflow design

The default evaluation game that we organized has the following description and consists of 8 rounds. In every round each assessor receives one task. Each task is assigned to two assessors. Each task consists of a query and 3 related documents. None of the tasks is evaluated twice during the game. Assessors evaluate each document according to the same 6-graded relevance scale as the topic experts used when they created the golden set (see Section 3.2). Right before the start of a game, assessors were provided with a short 1-page description of the relevance grades with examples.

The workflow is based on the idea of communication between assessors. After receiving a task, assessors read the documents related to the task and make judgments on their own. Afterwards, they meet their partner, the assessor who were assigned with the same task, discuss their judgments and make the final collective decision. By design, two particular assessors cannot be paired in more than one round during the same game.

Depending on what judgments assessors made before and after the discussion, they can receive cards of different colors: green, blue, red, black or gold. Each of them stands for a special score which we calculate for each assessor. There are 3 possible outcomes of a discussion:

- Assessors make the same judgment initially, i.e. before the discussion. In this case, each of them receives one green card.

- Assessors make different judgments initially, but come to a consensus after the discussion. The assessor who initially made the judgment which is approved by the consensus gains a blue card. Another assessor gains a red card. If they decide that the correct judgment differs from both their initial judgments, they both receive red cards.

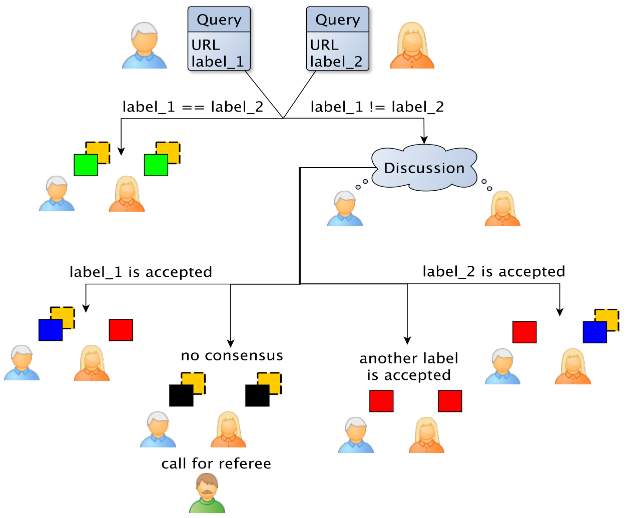

Figure 1: The workflow design. Dashed gold cards mean that the assessor can receive them if his/her label matches the corresponding golden set label.

- Assessors make different judgments initially, but do not come to a consensus after discussion. The absence of the consensus might mean that at least two approaches are equally possible in evaluation of that document. For the purposes of machine learning and search quality evaluation, every document must have only one label, so, in order to choose the only judgment we need to make an additional decision. One of the possible strategies is to create and maintain an assessors' rating list, which could inform us about who of the two assessors is more reliable. This approach has a sufficient disadvantage: if both assessors have a low rating level, we might accept a wrong judgment. We propose another approach. In case of the absence of the consensus, the third assessor - a *referee* - is invited. At the first round, the referee is chosen randomly from all the assessors: at every following round, the referee is chosen among vacant assessors (i.e. who are not occupied with a dispute) with the largest number of green cards by that moment, thus, we are sure that a referee is a definitely reliable person. The referee evaluates the same document by taking the assessors' opinions and arguments into account. The referee's opinion is considered as final and cannot be appealed. Both assessors who called for the referee receive a black card, the referee receives nothing. If a pair of assessors has more than one unresolved dispute in the scope of the same task (each task consists of 3 documents to assess), still only one referee is called to make the final decisions in all unresolved cases[2].

[1] http://trec.nist.gov

[2] Of course, in these experiments we could choose a referee with the highest number of gold cards, but since the agreement metric is aimed to substitute the accuracy, we use it advisedly for this purpose.

As a result, every document receives one final judgment: either the judgment that both of the assigned assessors agreed on or the judgment made by a referee. After that, we compare the final judgment with the corresponding golden set judgment. If the final judgment matches the judgment from the golden set, the assessor who had initially made it additionally receives a gold card, even if she had to call for a referee to defend her judgment. If the assessor initially made the correct judgment, but was convinced by the opponent that the opponent's opinion is correct, she did not receive a gold card.

Once the cards are given to assessors, the next round begins. Figure 1 illustrates the described workflow. The game stops after the 8th round.

Green and gold cards might be considered as gamified metrics of the agreement and the accuracy correspondingly. Other cards: blue, red and black indicate an assessor's productivity during discussions. Blue cards show how successfully an assessor insists on her judgments during discussions. On the contrary, red cards show how often an assessor agrees with the opponent's judgment. Finally, black cards show how inflexible an assessor is in disputes.

After all, we are able to compose a portrait of each assessor in terms of her cards. From the point of view of judgments collection system we are interested in assessors who make precise judgments and capable of convincing their opponents of correctness of own opinion. Thus, a successful assessor should have many gold, some blue, and as few as possible red and black cards. As for green cards, we do not know how many of them a good assessor should have. Ideally, if we have perfect judges, they should have as many green cards as gold, but, practically, we do not know if it is possible for one person to have many green but only a few gold cards, and vice versa.

There are two possible winners in a game: the assessor who gains the most number of green cards (i.e. the most predictable assessor), and the assessor who gains the most number of gold cards (the most accurate assessor), however, there could be one person who get both prizes. The other types of cards did not determine the winners, they were used only for our post-analysis and the participants were informed about that.

The described workflow was the default one. Depending on a specific experiment, this workflow was changed in the ways which are described below.

3.4 Experiment setup

Experiment 1: Control. The control experiment had the default workflow. We are highlighting the most important points to compare them with the other experiments. 16 inexperienced assessors took part in our control experiment. They did not have any reported experience of working as an assessor before and were not familiar with the detailed evaluation guidelines (see Section 3.1). Each task was evaluated by 2 assessors, the referees were selected among the participants. The game lasted 8 rounds and each assessor evaluated 24 documents. Two targets were set for the participants: to gain as many green and/or gold cards as possible. The purpose of this experiment was in the baseline estimation of the assessors agreement level and the accuracy, calculating the correlation coefficient between them and comparing these metrics across all experiments.

Experiment 2: With Instigators. The difference of this experiment comparing with the control experiment was in the assessors' target. The participants were only told that they need to gain as many green cards as possible, and they were not told that their judgments are also compared with the golden set. We expected that if assessors knew that their only target is to maximize their agreement with other assessors, sooner or later they would start to agree with each other more often, and, in the ultimate case, they would start making one and the same judgment for every document, thus obtaining a very high agreement level, but a very low quality of judgments. In order to model such extreme situation during our experiment, we invited two *instigators* to take part in this experiment. One of the instigators was selected from our professional assessors, another was selected from the participants and had no assessor experience. They were asked to indirectly and informally (during coffee breaks or discussions of disagreement cases) hint to other participants to follow that "optimal" strategy. Thus, we assumed that the average agreement level should be high opposed to the average accuracy which should be low. Besides, we expected to see no correlation between these variables.

Experiment 3: With No Communication. We expected that discussing disagreement cases helps assessors share and gain useful experience, as well as synchronize their approaches towards evaluation. In Experiment 3 we did not let assessors discuss disagreement cases. At every round in this experiment, two assessors compared their judgments for the same document; if the judgments matched, assessors received green cards, if the judgments did not match, they received nothing and proceeded to the next round of the game. Consequently, as there were no discussions, there were no referees to arbitrate cases when assessors could not come to a conclusion, and there were no blue, red and black cards. In the case when two assessors made different judgments, the final judgment for the document was chosen randomly among these two. All other parameters of the workflow were the same as in the control experiment. Since the conditions of this experiment are considered as the degradation in respect to the control, we expected both the average agreement level and the average accuracy to be lower in comparison with the control group. As for the correlation, we did not have any expectation how strong it should be in this case.

Experiment 4: With Experienced Assessors. We expected that trained assessors have higher consistency of judgments and wanted to make sure that both our metrics - agreement level and accuracy reflect it. Unlike in the other experiments whose participants never worked as assessors before, all participants of Experiment 4 were experienced assessors who had already been working as assessors for a long time. All other parameters of the workflow were the same as in the control experiment. Obviously, since the participants were more experienced in relevance assessment than the other assessors, we expected to see higher levels of both accuracy and agreement than in the control group.

Experiment 5: With Overlap 3. In this experiment, each task was evaluated by 3 assessors at every round. Since the total number of participants should have been divisible by three, 18 assessors (instead of 16 as in the previous experiments) took part in this game. As far as three judgments per document are compared with each other, the agreement notion should be adjusted here. An assessor received a green card if her initial judgment matched the majority vote (e.g.

Experiment	Description	# of participants	Experience	Accuracy Control	Communication	Overlap
1	Control	16 assessors	no	yes	yes	2
2	With instigators	14 assessors 2 instigators	no	**no**	yes	2
3	With no communication	16 assessors	no	yes	**no**	2
4	With experienced assessors	16 assessors	**yes**	yes	yes	2
5	With overlap 3	18 assessors	no	yes	yes	**3**
6	With moderators	16 assessors **4 moderators**	no	yes	yes	2

Table 1: Experimental design. Modified parameters are marked in bold.

at least two assessors out of three independently made the same choice). A referee was called only if all three assessors made different judgments. All other parameters of the workflow were the same as in the control experiment. Apparently, increasing the overlap parameter should increase the intensity of information propagation between the participants and, consequently, assist assessors in learning rules-of-thumb of relevance assessment. Thus, we expected to see all three statistics: the agreement level, the accuracy and the correlation to be higher than in the control experiment.

Experiment 6: With Moderators. We assumed that in the cases when all assessors, first, independently make different judgments, and then cannot come to a common decision even after a discussion are specifically difficult. We hypothesized that such extra-difficult cases cannot be fairly resolved by regular assessors, and we should ask for an opinion of some specific highly-trusted referees. In order to test this idea, we have introduced an additional role to our game: *moderator*. In all other experiments, a referee was chosen among the assessors with the highest agreement level, and any assessor could become a referee during a game. In Experiment 6, in addition to 16 inexperienced assessors, we invited 4 *moderators*: highly-trusted experienced assessors. In those cases when two assessors had initially made different judgments and could not come to a conclusion during the discussion, they called for a moderator to arbitrate their dispute. All other parameters of the workflow were the same as in the control experiment. As far as moderators are much more experienced than "common" referees, the final judgments in the cases of disputes were expected to be more accurate than in the other experiments. Moreover, the moderators could share their knowledge during the discussions, so this could affect not only the final judgments, but both the accuracy and agreement levels, which we expected to be higher than in the control group.

All the specifics of each game performed are summarized in Table 1. The modified parameter in every experiment is highlighted.

4. METHODOLOGY

To search for an answer for our Research Questions (see Section 1), we calculated a number of metrics for every experiment. These metrics are based on the numbers of cards received by each assessor during a game. The following 4 metrics describe the agreement level:

- Assessor's agreement level I (Agr. I): the percentage of the cases in which the assessor's initial judgment matched the judgment made by another assessor responsible for the same task, i.e. the number of green cards gained by the assessor divided by the number of documents she evaluated. We could not use just the number of green cards as an agreement level metric, since in Experiment 5, assessors evaluated less documents than in the others, and hence they could gain less green cards than assessors from the other groups. In order to eliminate this problem, we applied this normalization (division by the number of evaluated documents).

- Assessor's agreement level II (Agr. II): Cohen's kappa calculated for the assessor's judgments with respect to the judgments which were made by all her partners assigned to the same documents. Agr. I metric was convenient to discuss with assessors, because it is transparent and clear for understanding. However, in our analysis we used both Agr. I and Agr. II metrics.

- Average agreement levels for all assessors participating in the experiment (avgAgr I, avgAgr II).

In the further analysis we will show that Agr I and Agr II strongly correlate with each other, and, consequently, avgAgr I and avgAgr II have a high correlation level too.

The following 3 metrics relate to the notion of accuracy:

- Assessor's accuracy (Acc.): the percentage of the cases in which the assessor's initial judgment matched the corresponding golden set judgment (and was defended during the discussion if that occurred), i.e. the number of gold cards gained by the assessor divided by the number of documents she evaluated. We compared each assessor's judgments with the corresponding golden set judgments and calculated the fraction of correct answers;

- Average accuracy of assessors per experiment (avgAcc).

- Accuracy of the final set of judgments generated during an experiment (FA). A final judgment is a judgment either made by the majority of assessors, or as a result decision after a discussion, or as a referee's judgment. These judgments were compared with the golden set

Exp. No.	Description	avgAcc	avgAgr I	Corr. I	avgAgr II	Corr. II	FA	Corr. III
1	Control	0.378	0.479	0.417	0.315	0.381	0.534	0.979
2	With instigators	0.388	0.531	0.232	0.342	0.367	0.492	0.846
3	With no communication	0.318($\star$)	0.391($\star$)	0.638	0.193($\star$)	0.71	0.433($\star$)	0.969
4	With experienced assessors	0.448	0.615($\star$)	0.465	0.462($\star$)	0.521	0.543	0.987
5	With overlap 3	0.369	0.459	0.72	0.227	0.705	0.53	0.888
6	With moderators	0.344	0.422($\star$)	0.358	0.232	0.324	0.515	0.96

Table 2: The Results of the experiments. Average values of agreement levels I and II, average accuracy, final accuracy for every experiment; correlation between agreement level I and accuracy, correlation between agreement level II and accuracy, correlation between agreement I and agreement II. Values that differ significantly from the results of the control experiment are marked with $\star$ ($\alpha = 0.05$).

judgments and the fraction of correct answers was calculated.

We normalize the number of gold cards in the same way as we normalized green cards. The final judgments accuracy reflects the overall quality of judgments that can be collected within a given workflow.

The following 2 metrics show the correlation between the accuracy and the agreement.

- Pearson correlation coefficient between Agr. I and Acc. (Corr. I);
- Pearson correlation coefficient between Agr. II and Acc. (Corr. II).

It is the target group of statistics. If a workflow provides a strong correlation between the accuracy and the agreement, and keeps the accuracy on a high level, this allows us to consider it as a successful workflow, because such a workflow allows us to monitor the quality using the agreement level.

Finally, the following 2 metrics relate to the black cards:

- Normalized number of calls for a referee (CR): the number of the assessor's black cards divided by the number of documents she evaluated (i.e. maximum possible number of disputes);
- Average number of calls for a referee per experiment (avgCR):

For each experiment we also estimated the statistical significance of the differences between the experimental and the control group for each of the described metrics except Corr. I and Corr. II. The significance of the differences was measured by using the bootstrap method with resampling and estimating 95% confidence interval. The cases where the difference is significant are marked with an asterisk sign in Table 2. Pearson correlation coefficients were interpreted according to the widely accepted rule-of-thumb. Namely, the correlation between 0.2 and 0.4 was considered as weak, between 0.4 and 0.6 – moderate and between 0.6 and 1 – strong.

As far as there were no discussions in Experiment 3, its final judgments were sampled randomly from the two judgments of the assessors who disagreed on them. In order to reduce random effects in our analysis we perform this selection 1000 times and then average the accuracy and both agreement levels I and II.

Additionally, in order to estimate the relation between two agreement metrics, we calculated Pearson correlation coefficient between Agr. I and Agr. II (Corr. III).

We also collected some subjective reports. After each experiment, we had a discussion with the participants in which we asked them to tell us about their impressions, discovered winning strategies and other thoughts on the pros and cons of the game.

5. RESULTS

In this section we present the results of our experiments in terms of accuracy, agreement level, correlation between them and effectiveness of disputes between assessors. The results are provided in Tables 2 and 3, and illustrated in Figures 2 and 3.

5.1 Accuracy

In this section we analyze the average accuracy per experimental group (avgAcc) and the accuracy of the final set of judgments generated (FA). Both these values differ significantly from the values in Experiment 1 only in Experiment 3. We see that the communication capability has a significant influence on the quality of judgments: if assessors do not discuss disagreement cases with each other, they do not get any feedback on the quality of their judgments and cannot improve it, and, as the result, the accuracy of the judgments produced within this experiment was significantly lower than in the others. In all other experiments we do not observe statistically significant differences in accuracy levels, but some trends are interesting to analyze.

First of all, despite our expectations, we do not observe significantly higher accuracy of judgments provided by the experienced assessors in Experiment 4, though the accuracy of its judgments was the highest among others. In fact, we estimated that p-value of that difference is bounded between 0.05 and 0.06. The absence of significance might be related to the nature of the particular set of documents assessors had to evaluate.

In the experiment with instigators, while the average accuracy is the second best among all experiments (which means that assessors were generally making reasonable judgments at the stage of individual evaluation), the final accuracy is the second worst among all. Since the final accuracy depends on the judgments assessors agreed on after the discussions, this fact shows that assessors were not motivated to search for the correct judgment during the discussions. In subjective reports some of the participants admitted that they were really influenced by the instigators and tried to predict their opponent's thoughts and make a judgment that would rather match their opponent's judgment than the judgment that they believed it had to be, and, particularly, they had no motivation to come to the correct answer during discussions. However, not all assessors were affected by this temptation, some of them blamed the cheaters for this dishonest strategy and ignored such behavior.

In fact, the average accuracy in Experiment 2 is slightly higher than in the control group because of the influence of one of the instigators. As it was mentioned in Section 3.4, one of the instigators was selected from the professional assessors. He played his role of an instigator perfectly, but, at the same time, he made very accurate judgments. As a result, he received the most number of gold cards and became one of the winners in this experiment. If we do not take him into account while calculating the average accuracy (avgAcc), it decreases from 0.388 to 0.372.

Experiment 6 demonstrates a mirrored situation: the rank of this experiment according to the final accuracy is relatively higher than according to the average assessors' accuracy. This difference shows the impact of moderators who joined the discussions and generated accurate judgments, while the quality of judgments produced by the assessors themselves was not high. This effect will be considered later in detail in Section 5.4.

5.2 Agreement level

We observe that Experiment 4 shows the highest level of agreement both in terms of green cards gained by assessors (avgAgr I) and in terms of Cohen's kappa (avgAgr II). This observation matches with our expectations: experienced assessors who were trained to follow the same guideline are expected to have a high agreement level. At the same time, Experiment 3 (with no communication) has the lowest agreement level according to both metrics. Indeed, if assessors do not communicate, do not get any feedback and do not share their experience with each other, it is difficult for them to synchronize their approach towards evaluation. Although, the agreement level in Experiment 6 is significantly lower only according to avgAgr I, but not according to avgAgr II, it is one of the lowest among all other experiments. We think that it reflects the fact that assessors in this experiment were not motivated to learn from each other: instead they were waiting for the moderator to come and present the correct answer. The difference in the agreement level between other experiments is insignificant.

As it was expected, assessors in Experiment 2 have the highest agreement level among the inexperienced assessors, which confirms that the instigators' target has been hit. At the same time, the absence of the significance in this case could mean that their efforts were not enough. Anyway, we estimated the corresponding p-value and it is bounded between 0.07 and 0.08. On the other hand, we see from the subjective reports that some assessors were not affected by the instigators even after their attempts to incline them to cheating. Thus, we can optimistically suppose that there should be always fundamentally honest persons who never follow a cheating strategy, and such assessors make the difference insignificant. However, we think that the time period of the experiment was too short, and we would see the significant difference between the agreement levels if we extended the number of rounds. Of course, in this case, participants from other experiments could also realize "benefits" of cheating, but we suppose that the speed of information propagation with instigators is much higher than without them, so we still would see the significant difference.

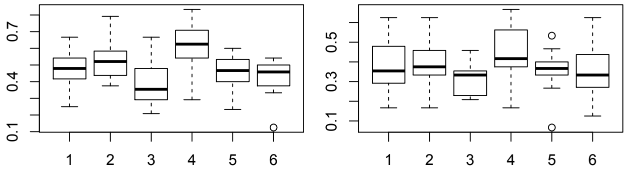

Figure 2: Accuracy (left) and agreement level I (right) per experiment.

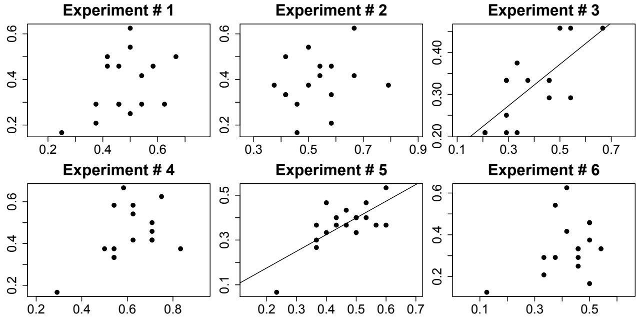

Figure 3: Correlation between accuracy and agreement per experiment. We plot the regression lines on the scatterplots if the corresponding correlation coefficients are greater than 0.6.

Finally, there is a very strong correlation Corr. III in all experiments, which confirms that both agreement metrics Agr. I and Agr. II are applicable. We may also note that in Experiments 2 and 5, Corr. III is not so large than in the others. In Experiment 5, such differences are really possible here, because in the case when overlap is 3 (opposed to the other experiments where overlap is 2), Agr I is calculated as the number of cases when an assessor's opinion matches the majority's opinion, but Agr II is still calculated as Cohen's kappa, i.e. as the number of cases when an assessor's opinion matches all partners' opinions she met. Anyway, since both avgAgr I and avgAgr II values in Experiment 5 are large (0.72 and 0.705 correspondingly), there are no contradictions with overlap 3 here. As for Experiment 2, the reason of a comparably low correlation is one of the assessors who adopted cheating strategy and made a lot of identical judgments. According to a property of Cohen's kappa, if one of the assessors always makes the same judgment, the kappa equals to zero [3]. Thus, the kappa of this cheater was very low, while the number of green cards was normal since she made one of the most frequent judgments, and hence, this slightly decreased the correlation. If we remove her, the correlation increases up to 0.97.

5.3 Correlation between accuracy and agreement

Next, we analyze the correlation between two metrics discussed above: the accuracy and the agreement level of assessors in each experiment. All the conclusions within this subsection are based on Corr I metric. Conclusions based on Corr II are the same. Two of six experiments - 2 and 6 - have a weak correlation (0.232 and 0.358), two experiments - 1 and 4 - have moderate correlation (0.417 and 0.465) and the other two - 3 and 5 - have a strong correlation (0.638 and 0.72).

Changing assessors' target to maximization of green cards in Experiment 2 led to the expected results: the agreement level in this experiment is one of the highest, while the accuracy is not so high and is near to the control group. In other words, in Experiment 2, assessors made approximately the same number of mistakes, but there were more "agreed" mistakes. As a result, this led to a strong decrease of the correlation between these statistics in comparison with the control group.

In Experiment 6, we observe a low correlation level too. This result could be associated with the specific behavior of the participants which will be described later in Section 5.4.

As we have already mentioned in Section 5.1, the accuracy of judgments produced by the experienced assessors (Experiment 4) was not significantly higher comparing to the control experiment, while the agreement level of experienced assessors was very high. Consequently, we observe a moderate correlation between accuracy and agreement level in this experiment.

Interestingly, in Experiment 3, we observe a strong correlation, but, as discussed above, both the accuracy and the agreement level are the lowest among all experiments. So, when assessors work completely on their own, do not communicate with each other and do not get any feedback on their work, their agreement level may well predict the accuracy of their judgments. However, there appears to be no benefit of such correlation, as though the accuracy can be predicted by the agreement in this experiment, its value stays way below the level that we expect from a high-quality evaluation workflow design.

Our control Experiment 1 shows a moderate correlation between accuracy and agreement, but we see how increasing the parameter of overlap in Experiment 5 may lead to an increase of this correlation. At the same time, the average accuracy in Experiment 5 is by 0.009 less than in the control experiment, which is not too much, and the difference is not statistically significant. So, strong correlation allows us to use the agreement as a quality metric, but only in the scope of this workflow out of all six we experimented with.

This means that assessors' quality monitoring should be based not only on the agreement metric, but on the accuracy metric too. So, we cannot abandon golden sets completely, but we can reduce the size of the golden set. Indeed, if we use only the accuracy as an assessor's quality metric, the size of the golden set for one assessor could be calculated according to a well-known formula related to finding confidence interval for a single proportion [9]:

$$n \simeq p(1-p)\left(\frac{z_{\alpha/2}}{\delta}\right)^2,$$

where n is the sample size (i.e. the size of the golden set), p is the expected assessor's accuracy (e.g. the accuracy for the previous period of time), δ is a margin error, α is a significance level, $z_{\alpha/2}$ is a $\alpha/2$ quantile of the normal distribution. Obviously, we use the same golden set for all assessors, but still there are too many documents. As a matter of fact, commercial web search engines need to collect a huge amount of judgments, and the number of the documents to be checked may be equal to several thousands of documents per month. Now, in the scope of the proposed workflow, we are allowed to check the same volume of judgments using the agreement level (i.e. without the golden set), and the golden set judgments will be used just to make assessors feel that they are watched. We do not have a specific recommendation for the size of a golden set to be used for such sanity checks only, but, obviously, it should be essentially less than in the case when accuracy is the only means of control.

5.4 Efficiency of disputes

This section is devoted to the efficiency of disputes between assessors. The efficiency might be considered from two points of view. The first one relates to the accuracy. Indeed, if several assessors independently make the same judgment of the given document intuitively, this judgment is likely to be correct. However, if assessors independently make different judgments, may one expect that their discussion will help them choose the correct judgment? In general, are the judgments made after a discussion more accurate than the judgments that were initially made by the assessors? Moreover, if one of the assessors initially made the correct judgment, which did not match the judgment of her partner, would the correct judgment be chosen as the result of the discussion with that partner? In order to obtain an answer we need to calculate the average accuracy before and after the disputes.

So far, we measured the assessor's accuracy as the number of gold cards divided by the number of total documents the assessor evaluated, but, according to our workflow, an assessor does not gain a gold card if she cannot defend her judgment in the dispute. Now, we calculate "pure" assessors' accuracy as the proportion of cases where their initial judgment matches the corresponding judgment from the golden set, this is what we call *assessor's accuracy before a dispute*. *Assessor's accuracy after disputes* is the proportion of correct final judgments, no matter whether they are obtained with or without a dispute, with or without a referee. Obviously, referees affect the accuracy after disputes, but still we want to take them into account, because we are interested in the cases when one of the assessors knows the correct judgment and finally this judgment is accepted, even if a referee is called for that. Afterwards, we calculate delta of the difference between the accuracy before and after disputes and their statistical significance using paired t-test for two population means. The results are provided in Table 3.

The efficiency might be also related to the frequency of calls for a referee. Obviously, in our workflow, calling for a referee (or a moderator as in Experiment 6) means additional costs, so given that all major metrics (accuracy, agreement and correlation between the two) are comparable, the workflow that produces less calls for referees is more effective. The frequency of calls for a referee (avgCR) is also presented in Table 3.

Experiments 2 and 5 have a significantly lower number of calls for a referee comparing to other experiments. In Experiment 2 assessors were only motivated to gain as many green cards as possible, i.e. to match with a counterpart at the stage of independent evaluation. So, after they did not match, they were not motivated to fight for the correct judgment and preferred to agree on something rather than call for a referee. We also do not see any significant increase of average accuracy after the discussions in this experiment. When assessors are only motivated to maximize their default (pre-dispute) agreement level, disputes mean nothing to them and do not help increase accuracy of final judg-

Exp.	avgAcc before	avgAcc after	p-value	delta	avgCR
1	0.435	0.536	0.0003	0.102	0.208
2	0.453	0.49	0.199	0.036	0.062($\star$)
3	0.424	0.424	NA	NA	NA
4	0.516	0.542	0.283	0.026	0.146
5	0.454	0.533	0.002	0.08	0.033($\star$)
6	0.419	0.516	0.002	0.096	0.359($\star$)

Table 3: Average accuracy in groups before and after discussion, their delta, t-test p-value and the average number of calls for referee. avgCR values that differ significantly from the control experiment are marked with $\star$ ($\alpha = 0.05$). NA values relate to Experiment 3 with no communication.

ments. This conclusion fits the previous conclusion about the motivation of this group of assessors made in Section 5.1.

However, we observe a different case in Experiment 5: even though the number of calls for referees was also low, the efficiency of disputes here was very high. Evidently, if three people participate in the discussion, they are more likely to come to a consistent decision which will be supported by at least two of them, that is why they did not need to call for a referee often within this workflow.

Experiment 6 demonstrates a significantly higher number of calls for a referee (a moderator) comparing to other experiments. It means that the number of cases when assessors could not come to an agreement was very high. We tend to explain it by a psychological effect: in a game where any of assessors may become a referee, participants of the discussion try to come to a conclusion and learn by themselves, because they believe that there is no any person with a principally higher level of expertise to help them. But, in the experiment with moderators, assessors' motivation to learn by themselves was shifted to the motivation to be taught by a more knowledgeable person: instead of thinking "We both do not know the answer, so let us think together and find a solution", assessors' way of thinking changed into "We both do not know the answer, so let us go and ask someone who will provide us with an explanation". On the other hand, the deltas of accuracy before and after disputes in both experiments 1 and 6 are approximately the same, almost 0.1, so experienced moderators gave almost the same profit as regular referees who were chosen from among the participants. So, since the final accuracy in these experiments is comparable (0.534 and 0.515 correspondingly) and the difference between them is not significant, there is no need to assign a special role of moderator, because it is more expensive and because it makes assessors lazier.

Despite that both Experiments 1 and 4 had a moderate frequency of calls for a referee, in Experiment 1, disputes helped to significantly increase the accuracy of judgments, while in Experiment 4 disputes did not affect the final accuracy significantly. It may be explained by the fact that the experienced assessors in Experiment 4 initially had a very high accuracy, and it could not go much higher than their common level of expertise in given topics. At the same time, for untrained assessors in Experiment 1 these discussions were very useful, which emphasizes the value of communication specifically for new assessors at the stage of learning.

6. CONCLUSIONS

In our study we proposed a new gamified workflow of collecting judgments. In order to investigate how certain parameters of the proposed workflow affect the quality of judgments in terms of accuracy, agreement level and the correlation between these two metrics, we have conducted a set of offline experiments in a form of a game. Based on the results of these experiments we obtain the following answers to our research questions.

RQ1. Agreement level may be used as one of the targets for assessors in our workflow. However, it is not sufficient: once assessors are told that their only target is maximizing agreement level, they often lose motivation to care about the accuracy of their judgments; a high agreement level of an assessor in such circumstances does not predict high accuracy of her judgments and cannot serve as the only measure of quality control. The workflow of collecting judgments should necessarily contain some additional quality control mechanism that would not allow assessors to always make one and the same judgment, thus artificially maximizing their agreement level. Assessors should always know that some of their judgments may be checked in order to estimate their accuracy. This conclusion allows practitioners to essentially reduce the size of the golden set, and use it only for maintaining assessors' feeling that they are watched.

RQ2. Communication between assessors plays a great role in the process of learning and sharing experience between assessors: they are likely to choose the correct judgment during the disputes. Also, we see that the absence of communication in our workflow affects the accuracy of judgments negatively, and discussing disputable cases by assessors helps increase both the accuracy of judgments and the inter-assessor agreement level, and its effect is even more significant for inexperienced assessors.

RQ3. Increasing the number of assessors performing the same task indeed leads to a stronger correlation between the accuracy and the agreement.

RQ4. Difficult evaluation cases in which assessors cannot come to a common conclusion even after a discussion should not be solved by any specially chosen highly-trusted experts. Despite our expectations, the institute of moderators introduced into our workflow led neither to an increase in accuracy, nor to an increase in agreement level, demotivated the assessors to learn by themselves and increased the costs of the whole workflow, because the assessors tended to call for a referee to arbitrate their disputes much more often.

Based on these results, we formulate our recommendations towards the collecting judgments workflow. Agreement level may serve as the main metric of assessor's quality, but assessors should know that the accuracy of their judgments is also observed. Assessors should have the opportunity to discuss disagreement cases with each other, and the process of training of a new assessor may be reduced to simply performing a set of tasks and discussing them with other assessors. Difficult disagreement cases can be solved by adding one more assessor to the discussion, and no specific roles are needed. Finally, we assume that the overlap in a real production workflow should be from 3 to 5 assessors performing the same task (participation of more than five assessors may make discussions less transparent). In our experiments, Experiment 5 adopted all the recommendations given above, and it provided the best correlation between the accuracy and agreement, keeping all other measures high, such as accuracy, agreement, final accuracy and efficiency of discussions.

To sum up, the workflow we propose does not require significant managerial efforts towards the preparation of comprehensive detailed guidelines (assessors mainly learn from each other), spending time on training the assessors and checking samples of assessors' judgments (first of all one monitors the agreement level, which is calculated automatically, and the accuracy checking may be reduced to a minimum needed to maintain assessors' motivation to make reasonable judgments), which makes the whole process of collecting judgments more agile.

7. FURTHER WORK

In the future, we plan to create and test an online interface which would allow assessors to evaluate real web documents online and discuss disagreement cases with their colleagues in an anonymous chat. Once we get a more convenient instrument for producing further experiments, we aim to experiment with larger overlaps and provide a deeper investigation of the role of overlap in the effectiveness of an evaluation workflow.

Another interesting question is what size of the golden set in the proposed workflow is enough for stimulating assessors to do their work properly and do not follow a cheating strategy.

Also, for the workflows with high overlaps we plan to explore in which cases an assessor should receive a green card. In the current version of Experiment 5, an assessor received a green card when she matched with the decision of the majority. However, another possible approach is to give a green card to an assessor when her judgment matches with the median of the set of all assessors' opinions. On the one hand, if all assessors make different judgments, median judgment can always be obtained, while the majority vote decision will not exist and we will have to call for a referee. On the other hand, encouraging assessors to match with the median judgment may discourage them from making extreme judgments: when they are in doubt, assessors may tend to choose judgments from the middle of the evaluation scale because they will more likely match with a median one.

Finally, it is interesting to take a closer look at the psychological aspects of communication in the evaluation process. Does the communication strategy which assessor decides to follow depend on the features of the characteristics of assessor's personality? Does it depend on the difficulty of certain tasks? Is there a relation with the design of evaluation workflows in general? It seems that there is a relation between the communication strategy and assessor's personal characteristics, because judgment discussions are similar to conflict situations, which are investigated by psychologists in conflict theories. Next, we suppose that assessors' can behave in different ways depending on the difficulty of the task. For example, they could lose their motivation to come up with the correct judgment if the task is hard to assess. And if it is true, how can we motivate them in such cases?

8. REFERENCES

[1] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval. 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 153–164. Springer, 2011.

[2] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: Are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 667–674, New York, NY, USA, 2008. ACM.

[3] K. Gwet. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, 2:1–9, 2002.

[4] G. Kazai, N. Craswell, E. Yilmaz, and S. Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 105–114, New York, NY, USA, 2012. ACM.

[5] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 205–214, New York, NY, USA, 2011. ACM.

[6] G. Kazai, N. Milic-Frayling, and J. Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 452–459, New York, NY, USA, 2009. ACM.

[7] G. Kazai, E. Yilmaz, N. Craswell, and S. Tahaghoghi. User intent and assessor disagreement in web search evaluation. In *Proceedings of the 22Nd ACM International Conference on Conference on Information &#38; Knowledge Management*, CIKM '13, pages 699–708, New York, NY, USA, 2013. ACM.

[8] M. Kosinski, Y. Bachrach, G. Kasneci, J. Van-Gael, and T. Graepel. Crowd iq: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 151–160, New York, NY, USA, 2012. ACM.

[9] P. Mathews. *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Mathews Malnar and Bailey, Incorporated, 2010.