QUALITY OF CROWDSOURCED RELEVANCE JUDGMENTS IN ASSOCIATION WITH LOGICAL REASONING ABILITY

***Sri Devi Ravana*[1]*, Parnia Samimi*[2] *and Prabha Rajagopal*[3]**

[1,3]Department of Information Systems, Faculty of Computer Science and Information Technology
University of Malaya, Kuala Lumpur, Malaysia
[2]Computer Eng. and IT Department, Shiraz University, Shiraz, Iran

Email: sdevi@um.edu.my[1], p.samimi@cse.shirazu.ac.ir[2], prabz13@yahoo.com[3]

DOI: https://doi.org/10.22452/mjcs.sp2018no1.6

ABSTRACT

Crowdsourcing has become a low-cost and scalable alternative to gather relevance assessments through crowdsourcing platforms. However, when gathering subjective human judgments, it can be challenging to enforce data quality due to the lack of vision about how judges make a decision. It is important to determine the attributes that could affect the effectiveness of crowsourced-judgments in an information retrieval systems evaluation. The purpose of the experiment that is discussed in this paper is to investigate if logical reasoning ability of the crowd workers is related to the quality of the relevant judgments produced through the crowdsource process. The study also evaluates the effect of cognitive characteristics on the quality of relevance judgment compared to the gold standard dataset. Through this experiment, a comparison study is done between the quality of the judgments obtained through the crowdsourcing process and the original baseline judgments generated by the hired experts by TREC. In the study, the systems performances were measured using both of these sets of relevance judgments to see its correlation. The experimentation reveals that quality of relevance judgments is highly correlated with the logical reasoning ability of individuals. The judgment difficulty level reported by the crowdsource workers and the confidence level claimed by the workers showed a significant correlation with the quality of the judgments. Unexpectedly though, self-reported knowledge about a given topic and demographics data have no correlation with the quality of judgments produced through crowdsourcing.

Keywords: information retrieval evaluation; human judgments; quality of relevance judgments; crowdsourcing; logical reasoning.

1.0 INTRODUCTION

The test collection is a widespread method for information retrieval systems evaluation. However, the variability and mass of test collections continue to expand, making it difficult and costly for engaging human assessors to generate relevance judgments [1]. To overcome the drawbacks of engaging expert assessors in creating relevance judgments, crowdsourcing is suggested. Based on previous literature, crowdsourcing was introduced by Howe[2]. Crowdsourcing found to be useful for tasks involving humans such as creating relevance judgments [3, 4]. Cost-effectiveness, fast results, and flexibility make the crowdsourcing approach appealing [5].

In a typical process of crowdsourcing, it is possible to have differences in the way workers judge the judgments which causes the inaccuracies in judgments. The discrepancies in turn can cause the unreliable evaluation of the retrieval systems. Hence, the reliability of creating relevance judgments through crowdsourcing in substitution of expert assessors poses a concern. The quality of relevance judgments could affect the crowdsourced relevance judgments. Therefore, various methods that could manage the value and reduce the inaccuracies caused by unreliable workers were developed. Nevertheless, the role of cognitive abilities in ensuring the high quality of the judgments is still unclear. Cognitive ability is related to the measure of general intelligence such as the ability to learn and problem solving skills [6]. This study evaluates the relationship between the ability of the workers in terms of logical reasoning ability and how that influences the judgments obtained through crowdsourcing. Logical reasoning can be defined as "the ability to reason from premise to conclusion or to evaluate the correctness of a conclusion" [6]. The study also investigates how reliable is the crowdsourced judgments in determining the ranking of the retrieval systems in comparison to the baseline ranking using the TREC generated relevance judgments. Respectively, the following research questions are addressed in this work (reliability here is referred to as the quality):

1. Does logical reasoning ability of a crowdsourced worker affect the reliability of crowdsourced judgments?
2. Does logical reasoning ability effect the information retrieval systems' rankings in information retrieval evaluation?
3. Do the self-reported competences of a crowdsourced worker influence the reliability of the crowdsourced judgments?
4. Do demographics data affect the reliability of crowdsourced judgments?

2.0 RELATED WORKS

The concept of relevance judgments and the factors that effect on its reliability has been broadly studied. Relevance judgment is subjective and can be varied among assessors [7]. It also can be different over time for the same assessor [8]. Producing efficient relevance judgments has raised questions for a long time [9]. However, system rankings are robust to some extent with this variation. Even with the high level of inconsistency in relevance judgments, a high level of agreement on system rankings appeared between TREC assessors and non-TREC assessors [10].

Recently, crowdsourcing is introduced in order to extend current test collections [11]. An important challenge in crowdsourcing is ensuring the quality of output [12] particularly in paid microtask platforms like Crowdflower or Amazon Mechanical Turk in which unknown and unskilled workers accomplish the tasks. Different scientific workshops have been allocated to use crowdsourcing in an efficient way [13]. The fist workshop including crowdsourcing track was offered by TREC 2011, focusing on using crowdsourcing for making relevance judgments [14]. Ever since, studies have focused on incorporating crowdsourcing in information retrieval.

Findings of former studies are a mixture about the quality of crowdsourced judgments. Some studies claim that crowdsourced relevance judgments can be used as a reliable alternative [1]. Kinney et al. [15] showed that judgments of non-expert for domain-specific queries led to notable errors influencing system rankings. Some other studies investigated the factors effect on quality of relevance judgments. For example, a study explored if the accuracy of relevance judgment through crowdsourcing is influenced by workers' personality and demographics [6]. In another case, it was highlighted that the interest and incentive of worker conducting the task highly influence the accuracy of relevance judgments [7].

Former approaches to ensure the quality of workers' output is to insert gold tasks (tasks with known answers) on which crowds' performance can be checked. However, using gold tasks are costly and have a narrow application. Recently, monitoring workers' behavior and interaction during a task are used to estimate their task quality [16]. In a study [16], the behavior of trained expert judges and crowdsourced workers was compared, and the trained judges' behavior used as a gold behavior in order to identify crowdsourced workers who perform poorly. The results of this study showed that the approach doubles the accuracy in some tasks. In another study, the effect of available time to make relevance judgments on its quality was assessed [8]. They showed that the cost of crowdsourced experiments can be reduced by decreasing the available time to make the judgments without effecting on quality. [8] contains further details on the factors influencing the quality of relevance judgment [8].

Besides the crowdsourced workers' personality, demographics, interest, incentives and behaviour, the cognitive abilities were explored by some studies including relevance judgments by crowdsourced workers. Personality variances in cognitive performance can be referred as cognitive abilities. The terms intelligence, aptitude and cognitive abilities, which are substitutable, are commonly defined as the learning ability, adapting new situations and solving problems. Internal cognitive abilities are also associated with problem solving performance [17]. In another word, a complex combination of cognitive abilities is intelligence. Variety of research studies investigated these cognitive abilities. For instance, cognitive abilities have been introduced as an important indicator of people performance in certain jobs in management research. Library research has shown that users with high level of cognitive abilities utilize IR systems more efficiently. Studies have shown the importance of cognitive abilities in the performance of technology-based tasks [18][19][20]. In a separate study, the author has investigated on how the cognitive ability affects the reliability of the judgments which are generated through crowdsourcing [10][11]. The results showed that verbal comprehension which is an element under the cognitive ability, does affect the reliability of these judgments.

In addition to cognitive abilities, the effect of logical reasoning on the performance in searching was investigated. It was found that logical reasoning affects information searching behaviour of users. Besides that, it was also found that the people with higher order of logical reasoning skills tend to be more skilful in searching activities [22]. In a previous study it was also found that people such as catalogers, curators and librarians tend to show a good level of reasoning skills (logical) [21]. Based on the previous findings, the authors hypothesized the possibilities of logical reasoning skills could affect the reliability of the relevance judgments by the crowdsource workers. There are possibilities that users with higher level of logical reasoning skill would possibly able to generate more reliable judgments. The authors predict that information retrieval practitioners could increase the effectiveness and the reliability of the evaluation of IR systems by selecting crowdsource workers that could produce high quality judgments. Hence, there is possibilities of using the logical reasoning attribute as the measure of worker's quality in judging the relevance. The logical reasoning can be measured using Nonsense Syllogisms Test whereby scores derived from the test lead to classification of an individual's logical reasoning ability.

3.0 METHOD

The experiment hypothesizes positive correlation or association between crowdsourced workers with varying levels of logical reasoning ability and the reliability of the judgments in giving an accurate evaluation results. Additionally, the association between the reliability of judgments and claims by the workers such as the topic knowledge, how difficult is the task assigned, and the level of confidence the workers have in the judgments task and their demographics data are investigated in this study. Based on the additional association investigated, this study further hypothesizes trustworthy judgments are made by those workers (i) with greater knowledge about the topic, (ii) workers who found the task to be easy, and (iii) workers who were very confident with their judgments.

3.1 Experimental Data

A total of ten topics from TREC 2011 Crowdsourcing Track[1] was used for this experimentation. For every topic, ten documents were randomly chosen from the ClueWeb09 dataset[2]. The documents may have been classified as highly-relevant-(HR), relevant-(R) or non-relevant-(NR) by hired expert assessors from TREC. A crowdsourcing platform, Crowdflower[3], was used to obtain 50 graded judgments for every 100 topics-documents in the experiment. In total the workers completed 5000 judgments.

3.2 Designing Tasks

This experimentation contains 20 HITs using Crowdflower. 50 workers completed each HIT which totals up to thousand HITs. Fig. 1 presents a schematic procedure of the designing task for this experiment.

As the first step, crowdsource workers are required to carry out the judgment tasks through the HITs provided. Each HIT is designed such as way that it contains 1-topic and 5-documents (for each topic). Once the judgment is completed by the worker, he/she is required to provide other related information which is in a form of 4-point scale. The questionnaire consists of three items which are: Q1-Knowledge on the topic, Q2 – Difficulty of the evaluation process, and Q3-Confidence on the evaluation completed below; that measures the judgment difficulty, their knowledge about the given topic and their confidence in judgments. The procedure produced valuable data on the relation (if exist) between the level of self-claimed competences of the worker and the reliability of the judgments.

Secondly, to evaluate the workers' cognitive abilities, the logical reasoning ability was measured. The Factor-Referenced Cognitive Tests (FRCT) [6] and Nonsense Syllogisms Test for logical reasoning was measured. The workers answered 10 questions that will quantify their ability to accurately draw conclusions from certain statements. The scores for logical-reasoning is computed by finding the difference between the frequency of correct answers and wrong answers. This

[1] sites.google.com/site/treccrowd/2011
[2] www.lemurproject.org/clueweb09.php
[3] www.crowdflower.com/

computation is done for the ten questions given to the workers. A sample of Nonsense Syllogisms Test is shown in the Fig.2.

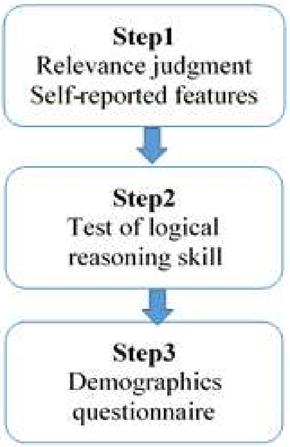

Fig. 1. Task design

In this experiment division of workers are done based on their scoring. Group 1 low-score workers, Group 2 is moderate-score workers, and Group 3 is high-score workers.

Every car has red wheels. Every van is a car. Therefore, every van has a red wheel.
(1) It is Correct (2) It is Not Correct

Answer: It is Correct .

Fig. 2. A sample of Nonsense Syllogisms Test

During the last step, the workers completed a set of five questions (including a trap question) about their demographic information. The demographic questions acquired some information about age, gender, educational, computer experience and the Net experience for each worker. This information was then used to find a statistical association (if any) with the quality of their relevance judgments. Table 1 shows the demographic attributes and the levels.

Table 1. Demographic attributes and the levels

Demographics	Level
Age	< 20
	>= 20 and < 30
	>= 30 and < 40
	>= 40 and < 50
	>= 50 and < 60
	> 60
Gender	Male
	Female
Level of Education	No education
	Basic schooling
	High school
	Bachelor degree
	Master degree
	PhD or higher
Level of Experience with Computer	1
	2
	3
	4
Level of Experience with Internet	1
	2
	3
	4

3.3 Quality of Relevance Judgments

The quality of crowdsourcing as an alternative to expert judges in creating relevance judgments needs to be evaluated. Inter-rater agreement between expert and crowdsourced judges were done to validate the findings. Cohen's Kappa and % agreement was used to compute the inter-rater agreement [23]. Since the Kappa method takes into account the random agreement of different judges/assessors, it is found to be more vigorous than % agreement [24]. Landis and Koch (1977) proposed a 5-level scale to understand the Kappa method. The scales are: (>=0.01 and <0.21: Slight-agreement), (>=0.21 and <0.41: Fair=agreement), (>=0.41 and <0.61: Moderate-agreement), (>=0.61 and <0.81: Substantial-agreement) and (>=0.81 and <1.00: Perfect-agreement).

Agreement between workers was measured either by ternary or binary agreement [26]. We say that there is agreement between workers when both the crowdsource worker and TREC judged the relevance similarly. This study has fifty different judgments created by fifty different crowdsourced workers for each topic and document (total 1000 HITs). An aggregating method known as Majority Voting (MV) is utilized to aggregate the number of assessment to a single assessment. The MV is commonly used for aggregation and is used in this study to aggregate the judgments. The group agreement between relevance judgments by crowdsourced workers and TREC assessors is conducted to detect any improvement in the group agreement compared to the individual agreement. The group agreement is also to determine if the aggregation of numerous high logical reasoning ability of crowdsourced workers is improved compared with low cognitive abilities ones.

3.4 Filtering Spam

Though crowdsourcing is convenient, it can attract unreliable workers to complete tasks quickly and inaccurately to gain payment. Therefore, quality control is necessary to filter out such workers [7]. Two assurance criteria and qualification settings were utilized for filtering in this experimentation. Workers who fail to meet the criteria were excluded. The two assurance criteria are trap questions indicating if workers perform tasks accurately. Another common filtering method is the task completion time which identifies malicious random answers [27] specifically in crowdsourcing [28]. Hence, tasks that are completed within 2 minutes were categorized as spammers.

4.0 Results and discussion

A total of 519 workers contributed to this experimentation. Out of the 1000 HITs, 25 HITs failed the trap question and were removed. From the remaining 975 HITs, 186 HITs were classified as unreliable HITs because the task completion time was less than 2 minutes. Finally, 789 HITs (from the initial 1000 HITs) or 3945 judgments were identified as "reliable HITs". Fig.3 shows the number of HITs assessed by each crowdsourced worker in the logical reasoning experiment. A number of workers judged all the HITs, with the most conscientious workers accomplished all 20 HITs. Most workers performed only a task. The workers were divided based on their logical reasoning scores and percentile-split. In summary, the Grp 1: Workers with low-scores, Grp 2: Workers with moderate-scores and Grp 3: Workers with high-scores. The purpose for this division is to test the hypothesis if the group of workers with higher logical reasoning skills could produce better quality judgments similar to the ones generated by the expert assessors of TREC.

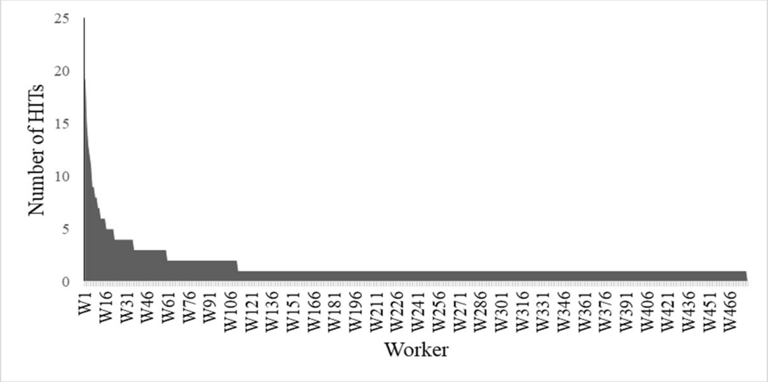

Fig. 3. Number of HITs judged by crowdsource-workers

4.1 Effect of Logical Reasoning Ability on Accuracy

Based on the experimental results, logical reasoning grade correlates with the quality of relevance judgments. There is a moderate correlation coefficient (r=0.30) between binary accuracy and logical reasoning. As for the ternary accuracy, the correlation coefficient (r=0.24) is small but significant with the logical reasoning ability.

Table 2 shows individual percentage agreement and Cohen's Kappa agreement between crowdsourced workers and TREC assessors of relevance judgments. From the table, mostly the ternary agreement is lower than a binary agreement for both percentage agreement and Cohen's kappa. The ternary agreement is lower due to lack of exact level of relevancy agreement between both assessors. The ternary and binary agreement between high grades workers and TREC assessors are 51.99% and 78.4% respectively. The binary agreement for relevance judgments between high grades workers and TREC assessors shows a moderate agreement (kappa=0.47). There is a slight agreement between low grades workers and TREC assessors with a kappa value of 0.10 whilst the binary agreement between moderate grades workers and TREC assessors is fair (kappa=0.37) and lower than the binary agreement between TREC assessors and Group 3 (kappa=0.47).

Table 2. Individual agreement

Workers	Ternary agreement		Binary agreement	
	Percentage (%)	**Kappa**	**Percentage (%)**	**Kappa**
All	47.81	0.21	71.59	0.30
Group 1	40.73	0.09	63.53	0.10
Group 2	51.21	0.26	74.04	0.37
Group 3	51.99	0.27	78.4	0.47

In a previous study [29], 75% overall agreement was found between TREC and non-TREC assessors for relevance judgments, close to the binary agreement found between Group 2 and the TREC assessors (74.04%) in this study. However, a strong individual agreement was not found between the groups and TREC assessors. As the previous study [30] also showed 70-80% overall agreement between two TREC assessors in a relevance judgment, this could be due to the subjective matters in judgment.

Table 3 summarizes the group agreement between TREC assessors and crowdsourced workers' relevance judgments. There is a slight kappa agreement (0.11) for relevance judgments between low grades workers and TREC assessors. The kappa agreement between Group 2 and TREC assessors is considered fair at 0.37. The highest kappa (0.60) value is between high grades individuals and TREC assessors in relevance judgments. With a higher agreement for both the individual and group agreement with the TREC assessors in relevance judgments suggests that Group 3 is more trustworthy compared to the other two groups. Besides, Group 2 is more reliable than Group 1 for creating relevance judgments. In addition, group agreements have higher values than individual agreements between workers and TREC assessors, indicating that group agreement is more reliable. In addition, utilizing the MV method to produce one judgment out of multiple judgments appears filtering out untrustworthy judgments.

Table 3. Group agreement between crowdsourced workers and TREC assessors' relevance judgments

Workers	Group agreement	Kappa
All	75%	0.35
Group 1	67%	0.11
Group 2	75%	0.37
Group 3	84%	0.60

Significance tests tell us whether there is a significant difference among the groups. One-way statistical significance test (ANOVA test) was conducted to identify statistically significant differences among the groups (if any) for ternary and binary accuracies**.** Statistically significant difference for both ternary and binary accuracy between the three groups were observed for $p < 0.01$. The results show that individual differences in crowdsourced workers' cognitive abilities are related to the quality of their relevance judgments. Based on the previous study [31] workers' characteristics can be deemed to assess the quality of their outputs if specific characteristics of individuals are related to their quality.

4.2 Effect of Logical Reasoning Ability on Rank Correlation

In this experimentation, five different sets of relevance judgments were used. The relevance judgments set were generated by (i) TREC assessors (a part of *qrels*), (ii) all of the workers, (iii) workers with low logical reasoning ability, (iv) workers who have moderate logical reasoning ability, and (v) individuals with high logical reasoning ability.

A total of 25 systems were ranked and graded using relevance judgments set created by low, moderate and high-grade individuals. Fig. 4 and Fig. 5 shows the rankings using Mean Average Precision (MAP) (k=1000) and MAP (k=10) based on relevance judgments set created by different grades workers and TREC assessors. The systems are arranged in ascending MAP scores from the TREC assessors' judgments. The system rankings based on relevance judgments by high-grade workers is similar to the system rankings based on TREC assessors compared with the other two system rankings. The moderate grades workers' system rankings are better than low grades workers in terms of the system ranking closeness to that of TREC assessors.

The Kendall's tau between the system rankings from the different groups is presented in Table 4. Kendall's tau compares the system ranks between distinct sets of relevance judgments by crowdsourced workers and TREC assessors. The system rankings based on TREC assessors, and all of the workers, low grades, moderate grades, and high grades workers have high correlation coefficients (MAP (10) and MAP (1000)). Furthermore, the correlation coefficients between high grades workers and the TREC assessors with is the highest with 0.81 for MAP (1000) and 0.75 for MAP (10). The correlation coefficients between moderate grades workers and TREC assessors are 0.61 for MAP (1000) and 0.66 for MAP (k=10), which are higher compared to that of low grades workers (0.54 for MAP (1000) and 0.52 for MAP (k=10)). Overall, the system rankings generated from workers with high logical reasoning ability were more reliable compared to low logical reasoning ability due to the higher correlation coefficient with system rankings based on TREC assessors.

Results of system rankings indicate that workers' logical reasoning ability has a little impact on system rankings. Meanwhile, system rankings produced by high grades workers (Group 3) is more trustworthy due to the highest correlation coefficient with system rankings by TREC assessors. Previous studies show that other factors such as different HIT design [26], assessor errors [10] and domain expertise [32] influence the system rankings. In the study that the effects of HIT design on system rankings were investigated, better system rankings can be achieved by a complete set of quality control methods [3].

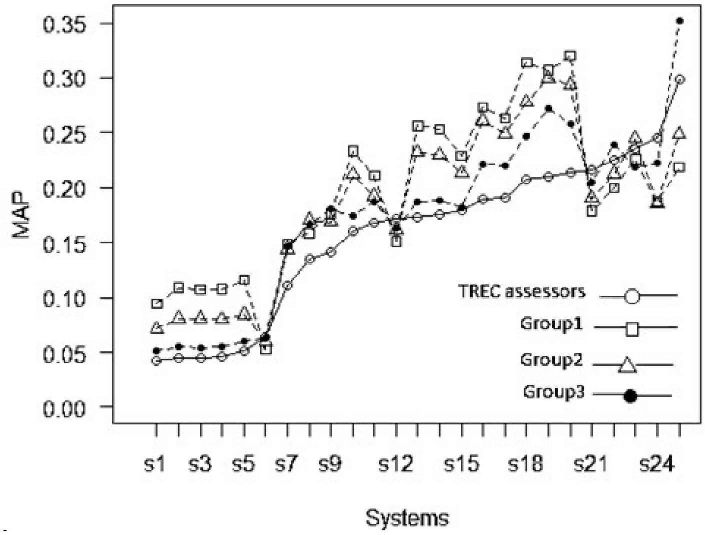

Fig. 4. Rankings of systems for different groups; MAP (k=1000).

Table 4. Kendall's tau correlation coefficient

Workers	Kendall's tau	
	MAP (*k*=1000)	MAP (*k*=10)
All	0.58	0.64
Group 1	0.54	0.52
Group 2	0.61	0.66
Group 3	0.81	0.75

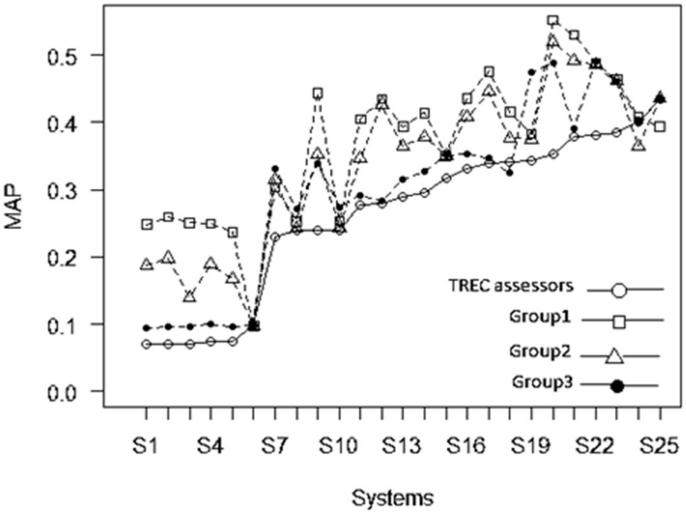

Fig. 5. Rankings of systems for different groups; MAP (k=10).

4.3 Effect of Self-Reported Competence on Accuracy of Judgments

This section details the results on the effect of various self-reported competence about the workers. Chi-square test for independence was applied to examine the effect of self-reported competence on crowdsourced judgment reliability. The Chi-square test for independence is used to explore the relationship between categorical variables. The test compares the observed proportion of cases for each category, and test the null hypothesis that the population proportions are identical [33].

Table 5. Self-claimed competence and judgment-accuracies

	Level	Relevance-Judgments	Ternary-judgments(Correct)	Binary-judgments(Correct)	Ternary-accuracy	Binary-accuracy
Confidence in judgment	1	159	63	85	0.39	0.53
	2	821	349	505	0.42	0.61
	3	1349	622	949	0.46	0.70
	4	1616	852	1285	0.53	0.79
Difficulty of the judgment	1	1682	882	1331	0.52	0.79
	2	1183	555	818	0.47	0.69
	3	889	368	562	0.41	0.63
	4	191	81	113	0.42	0.59
Knowledge on the topic	1	351	700	1268	0.40	0.72
	2	171	318	616	0.37	0.72
	3	211	374	742	0.35	0.70
	4	56	97	198	0.35	0.71

Note: Authors used the four-point Likert-scale for each self-claimed attributes.

Table 5 presents the ternary and binary accuracy for each level of confidence across the 3945 relevance judgments. The ternary and binary accuracy is increasing as the level of confidence are increasing. The trend shows that workers who were more confident with their judgments obtained a higher accuracy while less confident workers achieved lower accuracy. A

Chi-square test shows significant relationship between confidence and ternary accuracy ($\chi2 = 117.65$, $p < 0.001$). Similarly, the Chi-square test shows significant relationship between confidence and binary accuracy ($\chi2 = 25.27$, $p < 0.001$).

Moreover, the binary and ternary accuracy are calculated across all the 3945 relevance judgments for each level of difficulty to see the effect of difficulty of judgment on the accuracy of relevance judgments (Table 5). The results show that while the difficulty of the judgment is increasing, both ternary and binary accuracy is decreasing. A Chi-square test shows the relationship between difficulty and binary accuracy ($\chi2 = 95.65$, $p < 0.001$) to be significant, as well as between difficulty and ternary accuracy ($\chi2 = 31.71$, $p < 0.001$). For knowledge on the given topic, the results of binary and ternary accuracy are also shown in Table 5, but surprisingly the relationship between knowledge on the topic and accuracy of relevance judgments is not significant under a Chi-square test for independence.

Our results showed an association between confidence in judgment and quality of relevance judgments. The more confident workers were more trustworthy in their relevance judgments. These results were consistent with a former study [15] that showed a deficiency of confidence increases the likelihood of inaccurate judgments. In a separate study [34], the confidence score was introduced as a beneficial information to assess the accuracy of crowdsourced workers. In fact, the confidence score of workers was used for integrating crowdsourced judgments and the results showed improvement in the accuracy of crowdsourced results. Further, our results showed that there is a relationship between the difficulty of judgment and quality of relevance judgments. The workers who feel the judgments are easy; made more accurate judgments. The results are in line with a former study [7], which showed that the difficulty of the task is a representative of workers' performance. When the workers feel that the task is challenging, a clear descent appeared in the accuracy of workers' output. No association was found between knowledge on the topic and quality of relevance judgments, which is disaccord with what would be expected. Our results are in line with a former study [7] which showed there is no relationship between the familiarity with the topic and accuracy of relevance judgments. But, our results are in conflict with another study which found unfamiliarity with the topic and task influence the accuracy of relevance judgment [32].

4.4 Effect of Demographics on Accuracy of Judgments

In this study, some demographic information was acquired about the workers, which consists of age, gender, level of education, level of computer experience and level of Internet experience and their country as provided by Crowdflower. The demographics are assessed to find out how various demographics information about the workers is related to the quality of their relevance judgments. Table 6 shows the ternary and binary accuracy for each demographic information, across the 789 HITs. Chi-square tests show there is no relationship between demographic information and judgment quality. Looking at the demographics, our results show no connection between demographics and judgment quality of workers.

The findings for age are relatively in accord with a previous work [35] which found a small correlation between age and accuracy overall data, and no significant correlation coefficient between age and accuracy in a simple design HIT. This finding for gender supports the previously published work [35] reporting no significant relationship between gender and the accuracy of the results overall data. In terms of education, the expectation was that more educated workers would be better in creating relevance judgments, however, the finding is in accord with a previous work [35] which found no correlation between accuracy and education. Geographical location of the workers also showed no correlation with the judgment quality. A previous study [35] found that location has a very strong correlation with accuracy of judgments and the Asian workers had significantly lesser performance than American and European workers. However, as our HITs were limited to the English language countries mostly American and European workers, it is reasonable that no significant difference was found among different countries in their accuracy of relevance judgments in our study.

Table 6. Demographics and judgment-accuracies

	Level	Relevance-Judgments	Ternary-judgments(Correct)	Binary-judgments(Correct)	Ternary-accuracy	Binary-accuracy
Age	not yet 20	110	53	79	0.48	0.72
	in my 20's	1215	568	875	0.47	0.72
	in my 30's	1130	517	784	0.46	0.69
	in my 40's	750	368	557	0.49	0.74
	in my 50's	500	257	357	0.51	0.71
	60+ years old	240	123	172	0.51	0.72
Gender	Male	2045	934	1437	0.46	0.70
	Female	1900	952	1387	0.50	0.73
Education	no education	0	0	0	0.00	0.00
	primary school	55	31	40	0.56	0.73
	high school	1785	829	1256	0.46	0.70
	Bachelor degree	1590	775	1147	0.49	0.72
	master degree	465	221	338	0.48	0.73
	PhD or higher	50	30	43	0.60	0.86
Computer Experience	1	10	3	6	0.30	0.60
	2	320	128	213	0.40	0.67
	3	1335	633	928	0.47	0.70
	4	2280	1122	1677	0.49	0.74
Internet Experience	1	15	3	5	0.20	0.33
	2	225	109	158	0.48	0.70
	3	1510	687	1044	0.45	0.69
	4	2195	1087	1617	0.50	0.74
Country	AUS	95	41	62	0.43	0.65
	BHS	0	0	0	0.00	0.00
	CAN	725	349	511	0.48	0.70
	GBR	1200	592	864	0.49	0.72
	IRL	130	56	96	0.43	0.74
	NZL	50	26	33	0.52	0.66
	USA	1745	822	1258	0.47	0.72

5.0 CONCLUSION

One of the disadvantages of producing relevance judgments through hiring human experts is the high cost in the experimentation. A reasonable alternative method to generate relevance judgments set via crowdsourcing requires a precise assessment to ensure the quality of the outcomes. Through this study it was observed that the workers who scored high in the logical reasoning skills tend to be more reliable in producing judgements that are more correlated with the judgments produced by the human experts (baseline). Besides that, the system rankings generated using the baseline relevance judgments and the crowdsourced relevance judgments, showed that workers with high logical reasoning skills could produce highly correlated rankings with the rankings of the baseline method. In line with previous studies, the findings have in fact supported the claims that there is a relationship between cognitive abilities and the reliability of the crowdsourced judgments. This study emphasizes the significance of considering cognitive skills as a crucial factor in the relevance judgment task to achieve outcomes that are more trustworthy. IR practitioners are recommended to deem these characteristics while designing their experiments in future using crowdsource tools. In conclusion, there are many other cognitive abilities as well as a cognitive style that can be suggested to be included in future experimental designs. It would be motivating to investigate the impacts of various other psychological factors such as emotion and other personality traits on the quality of judgments. Besides that, the scalability of crowdsourcing for large-scale IR evaluation is a thrilling area of research that is highly recommended for future assessments.

ACKNOWLEDGMENT

This research was supported by UMRG Programme RP059A-17SBS from University of Malaya.

REFERENCES

[1] O. Alonso, "Implementing crowdsourcing-based relevance experimentation: an industrial perspective," *Information Retrieval. Boston.*, Vol. 16, no. 2, Apr. 2013, pp. 101–120.

[2] J. Howe, "The Rise of Crowdsourcing," Vol. 14, no. 6, 2006, pp. 1-4.

[3] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, 2011, pp. 1941.

[4] C. Grady and M. Lease, "Crowdsourcing Document Relevance Assessment with Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 172–179.

[5] O. Alonso and S. Mizzaro, "Using crowdsourcing for TREC relevance assessment," *Information Processing & Management*, vol. 48, no. 6, Nov. 2012, pp. 1053–1066.

[6] R.B. Ekstrom et al., *Manual for kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service, 1976.

[7] G. Kazai, J. Kamps, and N. Milic-Frayling, "An analysis of human factors and label accuracy in crowdsourcing relevance judgments", *Information Retrieval*, Vol. 16, No. 2, 2013, pp. 138-178.

[8] E. Maddalena et al., "Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge", in *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[9] M. Halvey, R. Villa, and P.D. Clough, "SIGIR 2014: Workshop on Gathering Efficient Assessments of Relevance (GEAR)", in *ACM SIGIR Forum*, ACM, 2015.

[10] B. Carterette and I. Soboroff, "The effect of assessor error on IR system evaluation" in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2010.

[11] A. Tonon, G. Demartini, and P. Cudré-Mauroux, "Pooling-based continuous evaluation of information retrieval systems", *Information Retrieval Journal,* Vol. 18, No. 5, 2015, pp. 445-472.

[12] A. Kittur et al., "The future of crowd work", in *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM, 2013.

[13] M. Lease and E. Yilmaz, "Crowdsourcing for information retrieval", in *ACM SIGIR Forum*, ACM, 2012.

[14] M. Lease and G. Kazai, "Overview of the trec 2011 crowdsourcing track", in *Proceedings of the text retrieval conference (TREC)*, 2011.

[15] K.A. Kinney, S.B. Huffman, and J. Zhai. "How evaluator domain expertise affects search result relevance judgments. in *Proceedings of the 17th ACM conference on Information and knowledge management*. 2008. ACM.

[16] G. Kazai and I. Zitouni, "Quality management in crowdsourcing using gold judges behavior", in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ACM, 2016.

[17] Mayer, R.E., Thinking, problem solving, cognition. 1992: WH Freeman/Times Books/Henry Holt & Co.

[18] Charness, N., et al., "Word-processing training and retraining: effects of adult age, experience, and interface," *Psychology and aging*, 2001, 16(1), pp. 110.

[19] Czaja, S.J., et al., "Examining age differences in performance of a complex information search and retrieval task," *Psychology and aging*, 2001, 16(4), pp. 564.

[20] Sharit, J., et al., "Effects of age, speech rate, and environmental support in using telephone voice menu systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2003, 45(2), pp. 234-251.

[21] Allen, B. and G. Allen, "Cognitive Abilities of Academic Librarians and their Patrons," *College & Research Libraries*, 1993, 54(1), pp. 67-73.

[22] Allen, B., "Cognitive Differences in End User Searching of a CD-ROM Index", in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992, ACM.

[23] Alonso, O. and S. Mizzaro, "Using Crowdsourcing for TREC Relevence Asessment," I*nformation Processing & Management*, 2012, 48(6). pp. 1053-1066.

[24] Cohen, J., "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, 1960, 20(1), pp. 37-46.

[25] Landis, J.R. and G.G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, 1977, 33(1), pp. 159-174.

[26] G. Kazai et al., "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,* ACM, 2011.

[27] D.E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms" in *CrowdSearch 2012 workshop at WWW 2012, Lyon, France*. 2012.

[28] T. Xia et al. "Real-time quality control for crowdsourcing relevance evaluation" in *2012 3rd IEEE International Conference on Network Infrastructure and Digital Content*, 2012.

[29] A. Al-Maskari, M. Sanderson, and P. Clough, "Relevance Judgments Between Trec and Non-Trec Assessors" in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2008.

[30] E.M. Voorhees and D.K. Harman, "TREC: Experiment and evaluation in information retrieval", Vol. 63, 2005: MIT press Cambridge.

[31] H. Li, B. Zhao, and A. Fuxman, "The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing" in *Proceedings of the 23rd international conference on World wide web*, ACM, 2014.

[32] P. Bailey et al., "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*, ACM, 2008.

[33] Soboroff, I., C. Nicholas, and P. Cahan., "Ranking retrieval systems without relevance judgments", in P*roceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, ACM.

[34] S. Oyama et al., "Accurate integration of crowdsourced labels using workers' self-reported confidence scores" in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 2013: AAAI Press.

[35] G. Kazai, J. Kamps, and N. Milic-Frayling, "The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy" in *Proceedings of the 21st ACM international conference on Information and knowledge management*, ACM, 2012.