Agreement of Relevance Assessment between Human Assessors and Crowdsourced Workers in Information Retrieval Systems Evaluation Experimentation

Parnia Samimi and Sri Devi Ravana

***Abstract*—**Relevance judgment set is created by human assessors (e.g TREC assessors) in Information Retrieval (IR) evaluation which is a laborious task. Recently crowdsourcing was introduced as a low cost method to create relevance judgment set. One of the important issues in crowdsourcing is the quality of labels or judgments which were produced by workers. This study, investigate whether increasing number of judgments for each topic and document improve the quality of judgments through measuring the agreement of relevance judgments between crowdsourced workers and human assessors to validate the use of crowdsourcing for creating relevance judgments. The agreement is calculated for both individual and group agreement through percentage agreement and kappa statistics. The results show that there is a higher agreement between crowdsource and human assessors in group assessment while in the individual agreement the agreement was low. However, when the number of workers to judge the same topic and document increases, the agreement between TREC assessors and workers does not increase significantly. In addition, we investigate how the rank ordering of a set of retrieval systems change when replacing human assessors' judgments with crowdsourced judgments using different number of workers. The results show that the system ranking is approximately the same when number of workers increase for judging the same topic and document.

***Keywords*—** information retrieval, evaluation, crowdsourcing, TREC.

I. Introduction

TEST collections are the common Information Retrieval (IR) evaluation approach that referred to Cranfield experiments which is the beginnings of today's laboratory retrieval evaluation experiments [1]. In 1992, the Text REtrieval Conference (TREC) was also established in order to support IR researches to provide the infrastructure for large scale evaluation of retrieval methodologies. Human assessors who appointed by TREC are responsible for making relevance judgments set that called qrels which is a laborious task. Different methods of creating relevance judgments are proposed. Researchers validate their methods for creating relevance judgment in IR evaluation by measuring the inter-rater or inter-annotator agreement. The inter-annotator agreement is applied to measure the performance in order to analyse the agreement between the judgments generated through the proposed method and judgments generated via human assessors to see whether the proposed methods are a reliable replacement for human assessors. One of these proposed methods which used to create relevance judgments set is crowdsourcing which conquer the problems that current evaluation methods have through expert judges. The term crowdsourcing was devised by Howe [2]. Outsourcing tasks, which formerly accomplished inside an institution by personnel allocated externally to massive of potential workers through Internet is crowdsourcing. Running experiments within low cost and fast turnaround make this approach very remarkable [3]. But the important issue is that whether the crowdsourced judgments is reliable and how to improve reliability of judgments in crowdsourcing.

In this work: (i) we investigate the agreement between the judgments created through crowdsourced workers and judgments generated via TREC assessors while increasing number of workers for judging each topic and document to see whether more numbers of workers can improve judgments. (ii) we evaluate how the rank ordering of systems (in terms of effectiveness measures) change when replacing human assessors' judgments with crowdsourced judgments using varying number of workers to judge the same topic and document to see whether more numbers of workers can improve system ranking. One of the important concerns of crowdsourcing is quality control since labels from non-experts are often untrustworthy which causes low accuracy. Collecting high quality labels is a challenging task. Label quality depends both on the expertise of the labelers and on the number of labelers [4]. If we assume that one judgment per each example or task called single-labeling method, the time and cost may be saved. However, the quality of work is dependent on an individual's knowledge. In order to solve this issue of single labeling methods, integrating the labels from multiple workers

Parnia Samimi and Sri Devi Ravana are with Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.

was introduced which called repeated-labeling [5] [6]. If labels are noisy, multiple labels can be desirable to single labeling even in the former setting when labels are not particularly low-cost [7]. Repeated-labeling leads to have more accuracy for relevance judgments [8]. Snow et al. suggested using more workers to improve quality of the results and encounter worker errors while doing tasks [9]. An important issue related to multiple labels per example is how to aggregate labels accurately and efficiently from various workers into a single consensus label. Majority Voting (MV) is one of the aggregating methods which applied in this study. MV is a straightforward and common method which eliminates the wrong results by using the majority decision [5] [9] [10]. The MV is a proper choice for routine tasks which are paid a lower payment since it is easy to implement and achieve reasonable results depending on truthfulness of the workers [11]. In 2009, Alonso et al. [12] investigate the use of crowdsourcing for IR evaluation by measuring the agreement between crowdsource workers and TREC assessors. Five different workers judged the same topic and document in the experiment. The results proved that crowdsourcing is a low cost, reliable and quick solution and an alternative to create relevance judgment by expert assessors but it is not a replacement for current methods as still there are several gaps and questions that left for future research. As an example, the scalability of this approach has not been investigated yet.

This study investigates the agreement between crowdsourced workers and TREC assessors while using different numbers of workers to judge the same topic and document. The main goal of this study is to examine whether increasing number of workers can improve quality of relevance judgments set for a test collection campaign. Firstly, the agreement between crowdsourced workers and TREC assessors is examined while testing different number of workers. Secondly, the rank ordering of systems is investigated when replacing human assessors' judgment set with crowdsourcing with different number of workers. The following section elaborates the experimental design. Section 3 presents the results of the experiment. Finally, the discussion and conclusion display in Section 4.

II. Experimental Design

One of the prevalent platform for implementing crowdsourcing is Crowdflower [13]. This experiment was conducted in Crowdflower. Topics were selected from TREC-9 Web Track and documents were chosen from WT10g collection which is a 1.69 million page corpus [14]. Each worker should answer a relevance question and Fig. 1 shows the task as seen by the workers.

The experiment consists of eight topics and 20 documents were chosen randomly for each topic which contains ten relevant and ten non-relevant documents to have a rational mix. Nine binary judgments were collected from different workers for each <topic, document> through Crowd flower. In total there are 1440 judgment.

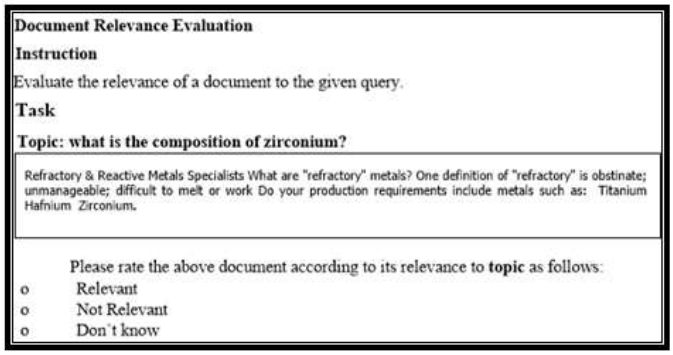


Fig. 1: A screenshot of the task

III. Results

A. Individual agreement

Individual agreement is measured for each worker and TREC assessor. If the worker and TREC assessor judgment is the same for the pair <topic, document>, they are considered to be in agreement. Individual agreement means that each worker is considered individually. There are different methods to measure agreement between each worker and TREC assessor. In this study, the individual agreement is calculated through two different methods: (i) percentage agreement and (ii) free-marginal kappa which is explained more in the following.

This measure sums the judgments which have the same judgments by two assessors (Crowdsourced workers and TREC assessors) and divide by the total number of judgments judged by two assessors. Table 1 displays the results and graphical representation of the percentage agreement. There is a 65.68% agreement between crowdsourced workers and TREC assessors (37.5% on relevant and 28.18% on not relevant). In order to evaluate the reliability of the agreement among assessors, kappa statistics was used. Formerly it was proposed by Cohen [15] that utilized to compare the agreement between two assessors. In this study, Free-marginal kappa was used which measures the degree of agreement while removing the effect of random agreement. Free-marginal kappa can be used for the case of multiple judges and when the assessors are not forced to judge certain number of documents [16]. The kappa statistic is computed using an online kappa calculator [17]. If a kappa value is above 0.6, it shows an acceptable agreement, while a value above 0.8 represents perfect agreement [18]. In this experiment, the Free-marginal kappa for individual agreement is 0.28 which can be seen as a fair but not high agreement below the acceptance value. As the individual agreement between TREC assessors and crowdsourced workers is not satisfactory, the agreement between TREC assessors and groups of workers is examined to see whether the group agreement is higher than the individual agreement. In addition, the group agreement is tested by 3, 5, 7 and 9 workers for judging the same topic and document.

TABLE I
INDIVIDUAL AGREEMENT BETWEEN CROWDSOURCED WORKERS AND TREC

ASSESSORS

		TREC assessors (qrels)	
		R	NR
Workers	R	37.5%	20.31%
	NR	11.25%	28.18%

B. Group Agreement

The group agreement is to have multiple judgments for each topic and document from different workers. In this part, the group agreement between workers and TREC assessors is calculated to see whether the agreement is improved compared with individual agreement. We examine the agreement by using 3, 5, 7 and 9 workers. For group agreement, we consider 3 random groups of workers and then average the result. The Majority Voting (MV) is used to aggregate the judgments. Table 2 shows the group agreement between TREC assessors and crowdsourced workers. Each topic and document is judged by 3 workers. Table 3 shows the same information but 5 workers used instead to judge the same pair of topic and document. The result of agreement for 7 workers was the same as 5 workers. Finally, the table 4 displays the agreement while using 9 workers to judge the same topic and document.

TABLE II
GROUP AGREEMENT BETWEEN 3 WORKERS AND TREC ASSESSORS

		TREC assessors (qrels)	
		R	NR
Workers	R	43.75%	14.37%
	NR	6.25%	35.62%

TABLE III
GROUP AGREEMENT BETWEEN WORKERS (5 OR 7 WORKERS) AND TREC ASSESSORS

		TREC assessors (qrels)	
		R	NR
Workers	R	44.37%	15%
	NR	5.62%	35%

TABLE IV
GROUP AGREEMENT BETWEEN 9 WORKERS AND TREC ASSESSORS

		TREC assessors (qrels)	
		R	NR
Workers	R	45 %	11.25 %
	NR	5%	38.75 %

The kappa statistics for the group agreement for group of 9 workers is above 0.6 that shows an acceptable agreement comparing with individual agreement which is not acceptable (0.28). In general, relevance judgments generated through groups is more reliable than the one generated individually for evaluating systems as we aim to replace the TREC assessors with workers. The kappa statistics for the group agreement for group of 3, 5 and 7 workers is 0.58 which is a moderate agreement and is close to group agreement of 9 workers.

TABLE V
AGREEMENT BETWEEN TREC ASSESSORS AND WORKERS

Number of workers	1	3	5	7	9
Percentage agreement	65.68	79.37	79.37	79.37	83.75
Free marginal kappa	0.28	0.58	0.58	0.58	0.67

IV. RANK CORRELATION

One of the common statistics in IR evaluation is Kendall's τ [19] which is a non-parametric statistic that utilized to examine the correlation between two ranked lists. In this study, to compare reliability of system ranking using different relevance judgment set, TREC assessors and crowdsourced workers, this test was applied. The high correlation means the ranking in both lists are the same. In IR evaluation, a Kendall's τ above 0.8 is considered as strong correlation.

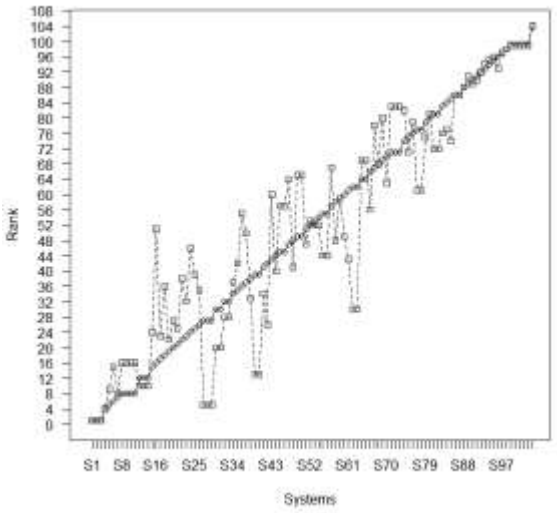
Evaluation of most researches in information retrieval is usually done by calculating precision. Precision is an information retrieval performance measure that quantifies the fraction of the retrieved documents which are relevant (see Equation (1)).

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved} \quad (1)$$

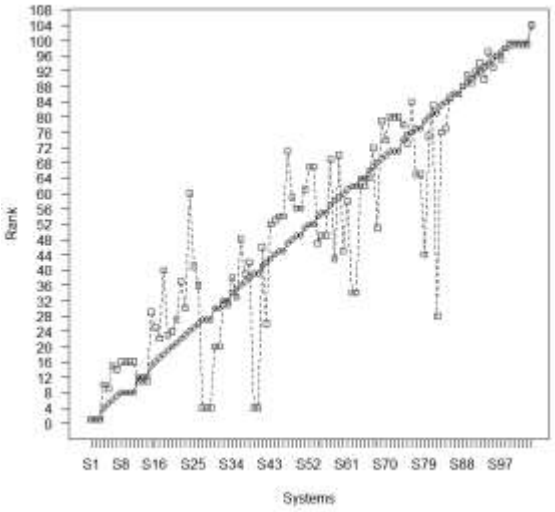
In this experiment, the systems are ranked based on the Mean Average Precision (MAP) which is calculated for each system by averaging the precision. In the first step, system ranking is found based on the relevance judgment set which created by TREC assessors. Then, the system ranking is created based on relevance judgment set which generated by crowdsourced workers while testing with 3, 5, 7 and 9 workers for judging the same topic and document. In the second step, the resulting ranked lists of TREC assessors are compared to workers through Kendall τ correlation. Table 6 displays the Kendall τ correlation. The ranked list of systems is shown in Figure 2. Based on relevance judgments created by 3, 5, 7 and 9 workers for each topic and document. The blue line shows ranked list of systems when using relevance judgment set which created by TREC assessors and the red line shows the ranked systems when crowdsourcing used as relevance judgment set.

TABLE VI
TAU VALUE USING MAP

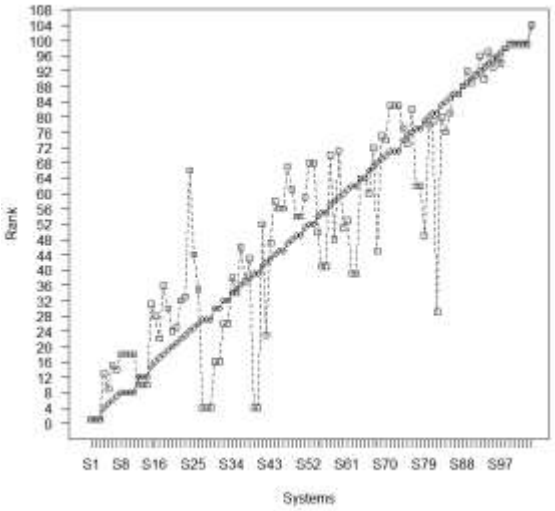
Number of workers	3	5	7	9
Kendall's τ	0.595	0.598	0.598	0.680



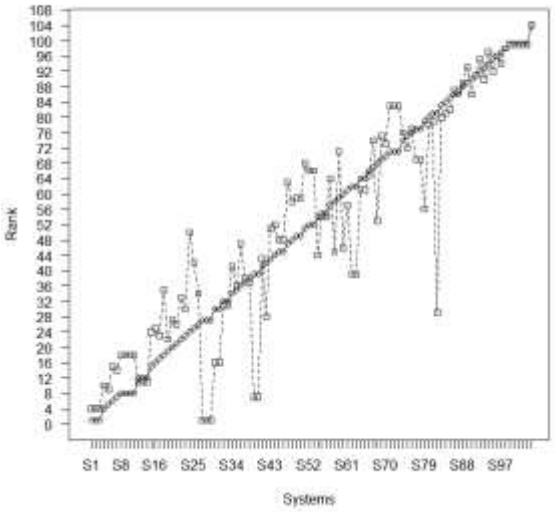
3 workers



5 workers



7 workers



9 workers

Fig. 2: System ranking based on qrels and workers (each topic and document was judged by 3, 5, 7 and 9 workers)

The Kendall's τ shows moderate positive correlation in ranking between crowdsourced workers and TREC assessors for all cases. The tau value for groups of 3, 5 and 7 workers are the same while 9 workers shows better correlation but not significantly different from groups of 3, 5 and 7 workers.

As Fig 2 shows, if the figures divide by two fractions, the two system ranking (TREC assessors and crowdsourced workers) are more comparable in the second fraction. So the tau value is calculated for two fractions of ranked list (first fraction and second fractions contains 52 systems).

TABLE VII
KENDALL'S T VALUE USING MAP FOR TWO FRACTIONS OF RANKED LIST

Workers	First fraction (System rank 1 – 52)	Second fraction (System rank 53 – 104)
3	0.58	0.76
5	0.60	0.74
7	0.56	0.72
9	0.60	0.75

As shown in Table 7, the second fraction shows the higher correlation in all cases. To recap, we can conclude that for low performance systems, the relevance judgments created by crowdsourcing produces a more reliable systems ranking.

V. DISCUSSION AND CONCLUSION

Summarizing, this paper examined the reliability of crowdsourcing for creating relevance judgment set while investigating whether increasing number of workers for judging the same pair of topic and document improves the relevance judgment set's quality. The agreement between TREC assessors and crowdsourced workers are measured examining individual and group agreement. The results show that when we use individual agreement, the percentage agreement between the TREC assessor and each worker is 65% and the kappa statistics show an agreement of 0.28 which is considered as a low agreement, but when using group assessment, the percentage agreement between them is 79.37% and the kappa statistics is 0.58 for 3, 5 and 7 workers while it is even higher for 9 workers, 83.75% and kappa of 0.67 which is an acceptable agreement. This leads to the conclusion that the crowdsourcing based on the individual judgment is not reliable and to have more reliable results, each topic and documents should be judged by multiple workers. However, increasing number of workers does not improve results significantly.

A further experiment investigated how the rank ordering of systems change when replacing human assessors' judgment set with crowdsourcing. In addition, we examined whether increasing number of workers to judge the same topic and document improve the system ranking. The coefficient correlation value shows moderate correlation between the two ranked lists. The tau value for groups of 3, 5 and 7 workers are the same while 9 workers shows better correlation but not significantly different from groups of 3, 5 and 7 workers.

However, when comparing the high and low performance systems, we can see higher correlation between TREC assessors and workers for low performance systems. A deeper analysis about using crowdsourcing' judgment set for system ranking will be part of future work. Investigating other factors that effect on the judgments quality is our future plan.

ACKNOWLEDGMENT

This research was supported by High Impact Research Grant UM.C/625/1/HIR/MOHE/FCSIT/14 and Exploratory Research Grant Scheme ER027-2013A.

REFERENCES

[1] C. Cleverdon: The Cranfield tests on index language devices. In: Aslib proceedings. pp. 173-194 MCB UP Ltd, 1967. http://dx.doi.org/10.1108/eb050097

[2] J. Howe: The rise of crowdsourcing. Wired magazine. 14(6), pp. 1-4, 2006.

[3] R. Baeza-Yates and B. Ribeiro-Neto: Modern Information Retrieval: The Concepts and Technology Behind Search. Addison Wesley Professional, 2011.

[4] H. Yang, et al.: Collecting high quality overlapping labels at low cost. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 459-466 ACM, 2010. http://dx.doi.org/10.1145/1835449.1835526

[5] V.S. Sheng, F. Provost, and P.G. Ipeirotis: Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 614-622 ACM, 2008. http://dx.doi.org/10.1145/1401890.1401965

[6] P. Welinder and P. Perona: Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. pp. 25-32 IEEE, 2010. http://dx.doi.org/10.1109/cvprw.2010.5543189

[7] P.G. Ipeirotis, et al.: Repeated labeling using multiple noisy labelers. Data Mining and Knowledge Discovery. 28(2), pp. 402-441, 2014. http://dx.doi.org/10.1007/s10618-013-0306-1

[8] M. Hosseini, et al.: On aggregating labels from multiple crowd workers to infer relevance of documents, in Advances in Information Retrieval. Springer. pp. 182-194, 2012. http://dx.doi.org/10.1007/978-3-642-28997-2_16

[9] R. Snow, et al.: Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on empirical methods in natural language processing. pp. 254-263 Association for Computational Linguistics, 2008. http://dx.doi.org/10.3115/1613715.1613751

[10] M. Hirth, T. Hoßfeld, and P. Tran-Gia: Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. Mathematical and Computer Modelling.(0), 2012.

[11] W. Tang and M. Lease: Semi-supervised consensus labeling for crowdsourcing. In: SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR). 2011.

[12] O. Alonso and S. Mizzaro: Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In: Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation. pp. 15-16, 2009.

[13] Crowdflower; Available from: http://crowdflower.com/,

[14] D. Hawking: Overview of the TREC-9 Web Track. In: TREC. 2000.

[15] J. Cohen: A coefficient of agreement for nominal scales. Educational and psychological measurement. 20(1), pp. 37-46, 1960. http://dx.doi.org/10.1177/001316446002000104

[16] S. Nowak and S. Rüger: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval. pp. 557-566 ACM, 2010. http://dx.doi.org/10.1145/1743384.1743478

[17] J.J. Randolph: Online Kappa Calculator. Available from: http://justus.randolph.name/kappa, 2008

[18] J.R. Landis and G.G. Koch: The measurement of observer agreement for categorical data. biometrics. pp. 159-174, 1977.

[19] M.G. Kendall: A new measure of rank correlation. Biometrika. 30(1/2), pp. 81-93, 1938. http://dx.doi.org/10.2307/2332226.