Review of Relevance Judgment Variation

Ziying Yang
ziyingy@student.unimelb.edu.au

August 19, 2016

Abstract

Relevance judgments, recording the pertinence of documents and topics, are widely used criterion for assessing the performance of *Information Retrieval* (IR) systems by IR evaluation metrics. They should be objectively and reliably assessed by groups of real users. The viewpoints of relevance are different from user to user, which leads the variation of relevance judgments, and so systems ordering may vary. A probabilistic model for computing the chance of the systems ordering change using different judges is reviewed and analyzed in this article. In addition, certain factors such as the relevance scale used for judgments potentially affect the relevance assessing because user's individual sense of distinction between relevance levels is various. We review an newly proposed approach, an psychophysical scaling technique, magnitude estimation, cooperating with a crowd sourcing platform, that tries to collect objective relevance judgments from real users.

The performance of a new IR system design should be examined by retrieval experiments in order to decide if the new design has meaningful advantages. In this evaluation process, the system returns a ranked list containing the *documents* (in text collections) topically related to the given topic. A variety of established evaluation *metrics* can be used to examine the performance of IR systems by comparing the retrieved results with the relevance judgments.

In text collections built for IR evaluation, relevance judgments are usually in the form of *qrels* composed by a list of tuples, each containing a topic, a document and the *relevance* score about how much overlap (pertinence) there is between the document and the topic. Conventionally, relevance judgments are assessed by human *experts*. Relevance judgments are also heavily relied on by evaluation metrics to compute scores for IR systems. So it is crucial to keep relevance judgments relatively objective and representative of the opinions of real users with respect to the original information need.

Relevance Judgment Variation In early work, Katter [5] stated that the reliability of relevance judgments was low. Relevance was usually subjectively judged by individuals (experts) independently. Even when relevance was assessed by groups, as was argued by Fairthorne [4], there were still great disparities between assessors' opinions [1]. In the relevance assessment of TREC4, up to 200 relevant documents plus 200 randomly selected irrelevant documents judged by the primary assessor (who created the topic) became a *pool* which was then passed over to another two secondary assessors [4]. Although these assessors had similar background and were trained for TREC task, the studies showed that the *overlaps* of their assessment agreements were under 50% [4]. Turpin

and Scholer [10] carry out similar tests (using another data set, TREC GOV2) and concluded that assessors did not agree well with each other in both task and topic levels.

There are several factors that may affect relevance assessments, including these noted by Burgin [2], Turpin and Scholer [10] and Kelly [6]:

- *human*: gender, age, background, preference, experience;
- *document*: representation, style, presented order, other documents in the set;
- *task*: time limitation, expectation, relevance scale used; and
- *topic*: rarity, saliency, ambiguity.

Users (or judges) may observe and measure *relevance* according to their own definitions [6], which may lead to conflicting results for the evaluation of IR systems.

In contrast, other researchers such as Salton and Lesk [7], have demonstrated that if the assessor consistency is low, their disagreements tend to be mostly found in deep ranks, and that the top relevant documents tended to be evaluated as relevant by most assessors, meaning that the variation of judges would not affect the system ranking in most cases.

Modeling Relevance Variation Turpin and Scholer [10] describe a probabilistic model of measuring the disagreement between relevance judges and how it affects the assessment of IR system performance made by batch-based evaluation. It estimates the agreement proportion between two judges (users): α_0 is defined as the probability that the new judge agrees with the old judge that document is irrelevant to the query; and given the document which is judged as relevant to the query, the new judge has a probability of α_1 to agree. That is, if the new judge always completely agrees with the old judge, then $\alpha_0 = \alpha_1 = 1$. If two judges always totally disagree with each other, then $\alpha_0 = \alpha_1 = 0$.

Given two systems A and B, denote the difference in their metric scores (measured using the old judge) cut off at depth d as $\Delta(d)$. If $\Delta(d) > 0$, system A is said to be better than B. Otherwise, the metric has voted for system B instead of A. Let $\Delta'(d)$ be their difference using the new judge, whose expectation is calculated as [10]:

$$E[\Delta'(d)] = (\alpha_0 + \alpha_1 - 1)\Delta(d). \tag{1}$$

The variance of $\Delta'(d)$, $\mathrm{Var}[\Delta'(d)]$, can be computed by $\mathrm{E}[\Delta'(d)^2] - \mathrm{E}[\Delta'(d)]^2$ and further simplified with assumptions made by metrics. If the $\mathrm{E}[\Delta'(d)]$ and $\mathrm{Var}[\Delta'(d)]$ are both known, the probability of $\Delta'(d) \leq 0$ (that is, system A remains superior) can be found.

This model is found to be useful for selecting metrics when appropriate choice of depth d and assumptions have been made. It tells the probability that the systems ranking varies according to the agreements α_0, α_1 of the new judge and the $\Delta(d)$ calculated by the old judge. So that the confidence of systems ordering using the old judge can be confirmed. For example, Turpin and Scholer [10] tested various combinations of α_0 and α_1 values for *Precision* at depth 1, assuming $\Delta(d) > 0$. It shows that if α_0 and α_1 are both greater than 0.8, the confidence of $\Delta'(d)$ exceeding 0 will be more than 95%.

They also tried to discover the most likely value of α_0 and α_1 in the practical web-based search task via user experiments. They concluded that for different topics or tasks, the agreement of users

varied above 50%. That is, the values of α_0 and α_1 are not stable across topics or tasks, so they need to be re-examined for each user and topic, which is too expensive when the dataset is big.

More than that, only the metric $P@d$ in the extreme case was tried in their experiments and it is difficult to extend this model to other metrics without proper assumptions that can simplify the calculation. Thus this probabilistic model is not practical in real cases without further improvement.

Relevance Scales On the other hand, researchers tried to reduce the judgment variation caused by objective factors, such as the relevance scale employed for the assessing.

Relevance scales allow users to measure the pertinence of document and topic in multiple degrees. The pertinence of documents and topics can be interpreted into several levels or categories. The relevance scale used for judgment assessing is usually preset by experiment creators instead of users. But the distinction between relevance levels expected by users may be disparate. Thus relevance scales used for relevance assessing can also cause judgment variation.

In the text collection TREC, relevance is divided into multiple levels. For example in TREC2005 [3], the relevance level for irrelevant documents is 0, for relevant documents is 1, and for highly relevant documents is 2. However the criteria that users have to determine the distinctions of relevance levels is diverse. *Generous* users (defined as assessing more than 50% of level 0 documents as relevant [9]) are more likely to judge documents as relevant, but *parsimonious* users (defined as assigning less than 50% of level 1 documents as relevant [9]) usually only judge the level 2 documents as relevant [8].

Scholer and Turpin [8] investigated the average time that users spend to assess documents of different levels as relevant. Their results show that the time for documents in level 0 being judged as relevant by users is the longest. The highly relevant documents (level 2) required the least time to be judged relevant, as expected.

Magnitude Estimation Since user perception of relevance is various and hard to incorporate into a single relevance scale, Turpin et al. [11] involved a psychophysical scaling technique, *magnitude estimation* (ME), working with a crowd sourcing platform, CrowdFlower, to collect relevance judgments from users.

Participants were paid to complete task *units*, in which participants need to assess the relevance of 8 documents (contain a known highly relevant document and a not relevant one) to a topic. If participants pass two tests for quality control (to ensure that participants have been familiar with ME and understood the topic), documents in the unit will be presented to them in random order. They can assign any positive score to the first assessed document. And then judge the next document with regard to the previous one during at least 20 seconds. For instance, if the document is twice as relevant as the previous one, its score should be double the assigned score of the previous document. The scores of two documents whose relevance level are known will be checked (that is, the score of highly relevant document should be greater) to ensure the judgment quality.

The score of document i assigned by a participant is denoted as S_i. Suppose the arithmetic mean of log scores of 8 documents in a unit u is μ_u, and of all the documents judged for a topic is μ, the final score of document i is normalized by

$$S'_i = \exp(\log S_i - \mu_u + \mu). \tag{2}$$

Magnitude estimation is considered as a suitable relevance scale for IR evaluation by Turpin et al. [11], since its judgments are consistent with ordinal judgments. They agree more with each other than binary judgments in document orderings.

With different relevance scales and gain profiles (mapping functions for metrics such as nDCG) used, the system ordering is substantially disparate. But the distribution of magnitude demonstrates that user's perceptions of relevance are neither readily fitted by a single profile, nor easily captured by any approach. Therefore, breaking this limitation is a key to evaluate IR systems accurately.

References

[1] 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 *Proc. SIGIR*, pages 667–674, 2008.

[2] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Inf. Proc. & Man.*, 28(5):619–628, 1992.

[3] C. L. A. Clarke, F. Scholer, and I. Soboroff. The TREC 2005 terabyte track. In *Proc. TREC*, 2005.

[4] R. A. Fairthorne. Implications of test procedures, in center for documentation and communication research. *J. Inf. Retr. in. Act.*, pages 109–113, 1963.

[5] R. V. Katter. The influence of scale form on relevance judgments. *Inf. Str. & Retri.*, 4(1): 1–11, 1968.

[6] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *J. FTIR*, 3(12):1–224, 2009.

[7] G. Salton and M. Lesk. Computer evaluation of indexing and text processing. *J. ACM*, 15(1): 8–36, 1968.

[8] F. Scholer and A. Turpin. Metric and relevance mismatch in retrieval evaluation. In *Proc. Asia Info. Retri. Soc. Conf.*, pages 50–62, 2009.

[9] F. Scholer, A. Turpin, and M. Wu. Measuring user relevance criteria. In *Proc. EVIA*, pages 50–62, 2008.

[10] A. Turpin and F. Scholer. Modelling disagreement between judges for information retrieval system evaluation. *Proc. Aust. Doc. Comp. Symp.*, page 51, 2009.

[11] A. Turpin, F. Scholer, S. Mizzaro, and E. Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proc. SIGIR*, pages 565–574, 2015.