Magnitudes of Relevance – Relevance Judgements, Magnitude Estimation, and Crowdsourcing

Anonymous Anonymous

Anonymous Anonymous

satisfaction

ABSTRACT

Magnitude Estimation is a psychophysical scaling technique where the intensity of a stimulus is rated by the assignment of a number. We report on a preliminary investigation on using magnitude estimation for gathering document relevance judgements, as commonly used in test collection-based evaluation of information retrieval systems. Unlike classical binary or ordinal relevance scales, magnitude estimation leads to a ratio scale of measurement, more suitable for statistical analysis and potentially allowing a more precise measurement of relevance. By performing a crowdsourcing experiment, we show that magnitude estimation relevance judgements are consistent with ordinal relevance ones; we study the difference of using a bounded or an unbounded scale; we show that magnitude estimation can be a useful tool to understand the perceived relevance when using an ordinal scale; and we investigate document presentation order effects.

1. INTRODUCTION

Gathering relevance judgements is a common and important activity in Information Retrieval (IR) evaluation. People recruited to perform the judgements can range from experts specifically hired for the task, as for example in TREC, to anonymous workers recruited online by means of crowdsourcing. In campaigns such as TREC and NTCIR, these judges are typically asked to assign relevance to a document using an ordinal scale. Historically, this has been a binary scale: relevant or not relevant. More recently, multi-level judgements from three or more categories have been used. In this paper we explore the use of Magnitude Estimation (ME) for collecting relevance judgements, rather than using an ordinal scale.

ME is a psychometric scaling technique where the intensity of a stimulus is rated by the assignment of a number. Subsequent stimuli are then assigned higher or lower values, depending on the perceived difference in their intensity from the previous stimuli. Magnitude estimation was initially developed for the measurement of physical stimuli such as the brightness of a light, or the frequency of a sound. However, it has also been successfully applied to the measurement of stimuli which do not have a physically measurable underlying scale, such as the perceived severity of crimes [14], and the usability of information technology systems [8].

In this paper, we report on a preliminary investigation of the ME technique for the measurement of document relevance judgements as commonly used in test collection-based evaluation of IR systems. Unlike ordinal scale approaches, the application of ME results in a ratio scale of measurement, making it more suitable for statistical analysis [8]. Also, intuitively ME relevance measurement on a ratio scale could be more precise than a multi-level, or even binary, categorical judgement as the granularity of the scale is chosen by the judge, and not constrained by categories. We also focus on gathering the ME measurements by means of crowdsourcing, a technique that has become popular in recent years [1, 6, 19].

Previous work in the IR field has applied ME for scoring the relevance of carefully curated document descriptions from a database of scientific papers [4], and for searching of a library database while carrying out personal research projects [13]. To the best of our knowledge we are the first to apply ME directly to judging the relevance of documents for test collection-based IR evaluation, and the first to investigate whether reliable ME relevance judgements can be obtained through crowdsourcing.

The key research questions addressed in this paper are as follows.

1. Are crowdsourced, document-level relevance judgements obtained using the magnitude estimation technique consistent with expert judgements obtained on a categorical scale?
2. Magnitude estimation can be applied using a bounded or an unbounded scale; does this choice impact on the quality of the relevance judgements obtained?
3. Can magnitude estimation enable better understanding of ordinal relevance scales, and in particular the differences between scale items?
4. Are document-level magnitude estimation judgements subject to presentation order bias?

EVIA'14, December 9, 2014, Tokyo, Japan.

This paper is organised as follows. In Section 2 we survey related work. Section 3 presents the experimental activity we have carried out. Sections 4 and 5 present and discuss the results of the experiment. Section 6 summarises the paper and sketches future work.

2. BACKGROUND

2.1 Relevance Assessment

Relevance is a fundamental concept in information retrieval, and human-generated judgements of the relevance of potential answer documents in response to an information need are a key component of *test collection* analysis, the most widely-used approach for the evaluation of IR systems.

Historically, relevance assessments were mostly made on a binary scale, where a document is rated as being relevant or not relevant. Many of the most widely-used IR evaluation metrics, such as MAP and P@10, use this notion of binary relevance [18]. However, humans are able to detect more fine-grained notions of relevance (for example, one document may be classified as being highly relevant to an information need, while another may be only marginally relevant). More recent evaluation metrics such as nDCG incorporate such multi-level relevance [7]. And indeed metrics that can take into account relevance judgements on a ratio scale have been proposed [3, 9], although they are rarely used in practice.

Typically, multi-level relevance assessments are made on an ordinal (ranked) scale. However, there has been much debate about the appropriate size of the scale – popular choices include three [2], four [12], or seven [16]. Moreover, no matter the size of the scale, the statistical operations that can be carried out to analyse data collected using an ordinal scale are limited, as it is well known in the measurement theory field [15, 17]. For example, since it cannot be assumed that the items on an ordinal scale have equal distance, the arithmetic mean is not a meaningful measure.

2.2 Magnitude Estimation

Psychophysics is the study of the human perceptual system, and aims to measure the subjective experience of sensory stimuli such as brightness and loudness. ME is a psychophysical technique for ratio scaling, proposed by S.S. Stevens. A series of stimuli at different levels of intensity are presented to an observer, who assigns numbers to the sensations in proportion to their magnitudes [5]. The key advantage of using ME is that it leads to a ratio (continuous) scale, compared to gathering ratings using traditional ordinal (ranked category) scales, meaning that a greater range of mathematical operations and statistics can be applied for analysis [8]. While ME was initially developed to measure the intensity of sensory stimuli that typically have an underlying measurable quantity, the technique is also widely applied to stimuli that are not physically quantifiable, ranging from levels of pain and emotion [5], to the intensity of opinions and attitudes on issues including the seriousness of criminal offences, political views, and the importance of Swedish monarchs [14], to HCI usability analysis [8].

Within the IR field, the application of ME has been limited. Eisenberg [4] investigated the use of ME to judge the relevance of document abstracts from a library cataloguing system, finding that the ratio scale obtained through ME was less influenced by ordering effects that a 7-point ordinal scale. Spink and Greisdorf [13] investigated the use of a bounded ME approach for measuring the relevance of items that participants found when conducting research for a personal project using a library database. Relevance was indicated on a complex worksheet that included both a 4-point ordinal scale, a bounded ME scale represented as a 77 mm line, and additional questions about the nature of the relevance of the items.

In contrast to previous work, we apply ME to the task of judging the relevance of whole documents, in a scenario that is typical of how relevance judgements are made in the context of test collection-based evaluation of IR systems. This paper is also the first to investigate whether reliable ME relevance judgements can be obtained through crowdsourcing.

3. EXPERIMENTAL DESIGN

To investigate our research questions concerning the use of ME to obtain document-level relevance judgements, we carried out a user study.

3.1 Participants

The study was carried out using CrowdFlower, a crowdsourcing platform (`www.crowdflower.com`). Each work task consisted of making four magnitude estimation assignments. Participants were paid $0.20 per task. CrowdFlower allows requestors to select the "performance level" of the workers that are allowed to participate in the experiment. Three worker levels are possible: "Good" for top performance workers who account for 60% of monthly judgements; "Great" who account for 36% of monthly judgements; and "Best", who account for 7% of monthly judgements. We selected the intermediate performance level, "Great".

3.2 Scales

To obtain a ratio scale, the instructions for ME typically require that judges assign any positive number. There is no upper limit, and fractional numbers are allowed, meaning that there is also no lower limit to the magnitude that can be assigned. We refer to this setup as an *unbounded* scale. As this kind of scaling may be unfamiliar to participants, so we also investigated the use of a *bounded* scale, with judges magnitude assignments being limited to a maximum level of 100, and still greater than zero.

3.3 Topics

why? how do you select these topics?

In order to be able to compare the ME scores with ordinal relevance ratings, three topics were chosen from the TREC-7 and 8 ad hoc tracks: 351, 355, and 408. These topics have existing ratings on a 4-point scale as assessed by carefully trained judges [12]: not relevant (`N`), marginally relevant (`M`), relevant (`R`), and highly relevant (`H`); we refer to this set of assessments as *expert judgements*.

3.4 Documents and orderings

ME requires participants to assign scores to multiple stimuli. As such, we constructed a number of pre-defined templates of document orderings, based on the relevance levels of the underlying expert judgements: increasing (`NMRH`), decreasing (`HRMN`), non-relevant (`NNNN`), and medium (`MRMR`). For each relevance level "slot", a document with a known judgement was randomly selected from the existing expert relevance judgements. To minimise possible variability from documents, the same documents were used where possible

(for example, for a particular topic, the H level document in the NMRH ordering was the same as the H document in the HRMN ordering). Overall therefore, 1 H, 3 R, 3 M and 4 N documents were used for each topic.[1]

3.5 Process

Participants were first shown instructions about the task. These included a short explanation of the ME process, and detailed the precise steps that the participant should carry out. Next, they were shown the *description* and *narrative* fields of one of the chosen TREC topics. For quality control purposes, participants were then shown a simple question to test their understanding of the topic. The question was in multiple-choice format, with the participant making a selection from four possible answers. After this, the main experiment began, with four documents being displayed in turn. For each document, the participant was instructed to "assign a number to every document in such a way that your impression of how large the number is matches your judgement of how relevant the document is". The instructions were adapted from Stevens (as cited in Gescheider [5] and Eisenberg [4]); the full text shown to participants is included in Appendix A. The workers typed the magnitudes into a text box available under each document. Participants were also required to enter text comments to explain why they entered the number that they did. Figure 1 shows a snapshot of the interface used.

not common

Each participant was assigned a single topic from the pool of three topics, and the documents that they were shown followed one of the four orderings described previously. They were instructed to use either a bounded or unbounded scale when carrying out the ME task. The input by the participant was parsed to check that it conformed to the required instructions (i.e., that the relevance magnitude was a real number, strictly larger than zero and, in the bounded case, strictly smaller than 100, and that the comment field was not empty).

4. DATA

With three topics, four document relevance orderings, and two possible scales (bounded and unbounded) in total we had 3 x 4 x 2 unique combinations, each of which was repeated by 10 workers, for a total of 240 workers and 960 document judgements. In this section we explain the data cleaning process, and provide a description of the raw ME scores.

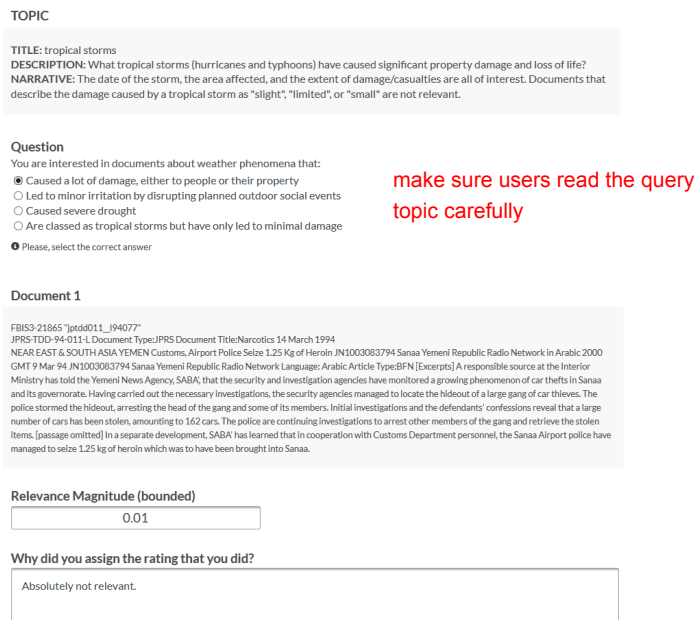
4.1 Data Cleaning

25 of the participants failed to correctly answer the question designed to test their comprehension of the topic, and were excluded from the analysis. As a further quality control, three of the authors judged the textual comments that participants were required to enter explaining why they had given the document a particular rating. The judging was blind, without knowledge of the particular topic being judged, and the assessment was simply to determine whether the comments were feasible responses for someone who was serious in their attempt to complete the task. For 26 of the remaining responses, two or more of the checks agreed that the comments were not reasonable, these participants were removed from the analysis, leaving a total of 189 participants whose responses are analysed in the remainder of this paper.

rather subjective

[1]To promote reproducibility, details of the documents used are available at `http://anonymised-for-review`

expression



make sure users read the query topic carefully

Figure 1: Interface used by participants to judge documents.

4.2 Normalisation of Scores

As the ME process allows people to assign values in an unrestricted way, it is usual to normalise the scores using geometric averaging, to obtain responses on a comparable scale [5]. We adopt the approach recommended by McGee [8] as follows. If ℓ_{jd} is the logarithm of the score assigned by judge j to document d, and J is the total number of unique judges, then we compute the normalised score

$$s^*_{jd} = 10^{\ell_{jd}+\mu-\mu_j}$$

where

$$\mu_j = \frac{1}{4}\sum_{d=1}^{4} \ell_{jd} \quad \text{is the mean score used by judge } j, \text{ and}$$

$$\mu = \frac{1}{4J}\sum_{j=1}^{J}\sum_{d=1}^{4} \ell_{jd} \quad \text{is the over all mean score.}$$

All logarithms are base ten in this paper. Note that for clarity of presentation we use "Log Normalised" scores in the figures and tables, which is simply $\ell_{jd} + \mu - \mu_j$. This does not alter most of the analysis as it is done on ranks (non-parametric analysis). Where appropriate, we use s^*_{jd} for comparisons.

4.3 Descriptive statistics

A summary of the raw (unnormalised), cleaned data is presented in Table 1, including the minimum, median and maximum ME scores that participants assigned when using the bounded and unbounded scales. Although the unbounded maximum is much higher than the bounded maxi-

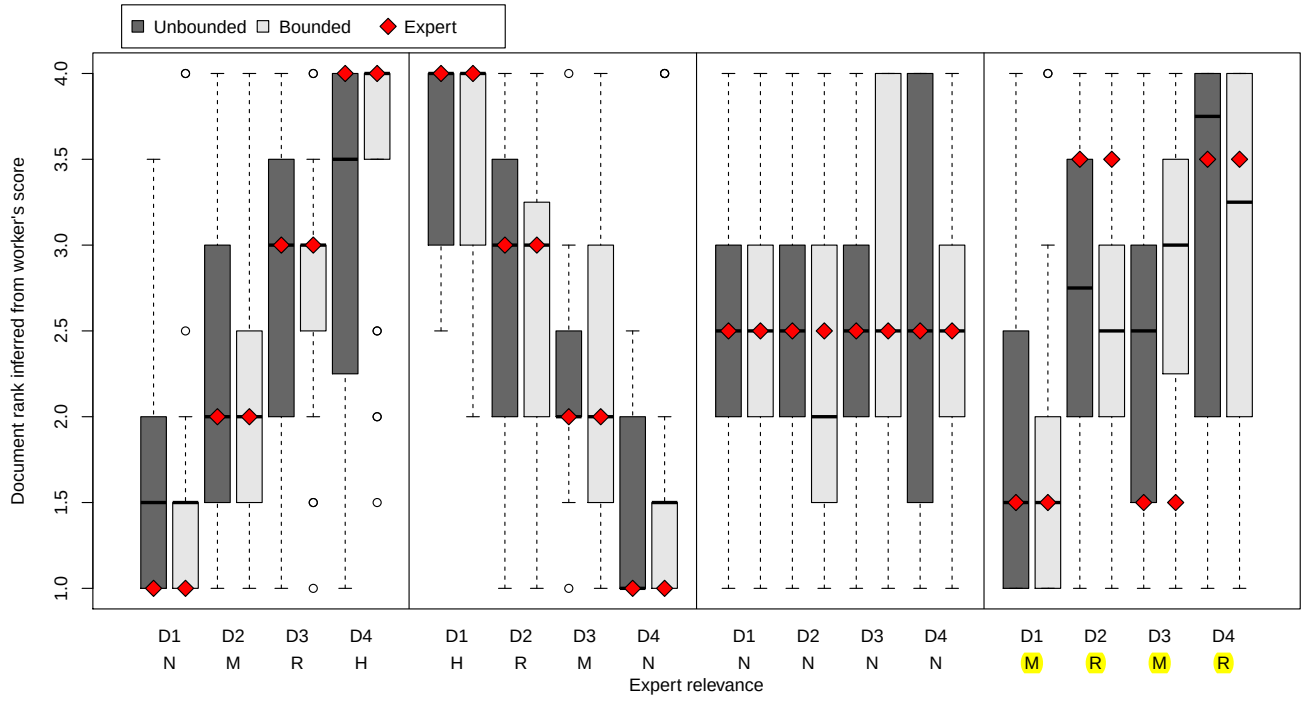


Figure 2: The rank of documents in each experimental condition (NMRH, HRMN, NNNN, and MRMR) inferred from worker's ME scores and the expert judgements on a 4-point scale. D1 to D4 indicates the presentation order of documents.

Number of judges	189
Number of judgements	756
Unbounded scores minimum	0.01
Unbounded scores median	4
Unbounded scores maximum	11,000,000
Bounded scores minimum	0.01
Bounded scores median	20
Bounded scores maximum	99
Number in NMRH Condition	50
Number in HRMN Condition	42
Number in MRMR Condition	42
Number in NNNN Condition	55

Table 1: Descriptive statistics of data after cleaning.

mum, as might be expected, the unbounded median is, perhaps surprisingly, lower than the bounded median.

5. RESULTS

In this section we present the results of our experiments and discuss how they answer the research questions.

5.1 Consistency of ME & Ordinal Judgements

To answer the first research question, whether the magnitude estimation technique can be applied to obtain consistent document-level relevance judgements, we compare the relative relevance orderings assigned using ME and the 4-level ordinal expert judgements. The results are shown in Figure 2, where the expert judgements are shown on the x-axis, and the y-axis shows the rank of documents inferred by sorting the ME scores of each judge for the four documents they were shown. Each of the four panels corresponds to one of the document orderings. The darker boxes show the data from the unbounded ME scale, while the lighter boxes show results for the bounded scale. The red diamonds show the nominal rank inferred from the expert judgements in each condition.

There is a clear and consistent agreement between the median of the ranks inferred from the ME scores and those inferred from the expert judgements for the NMRH, HRMN and NNNN document orderings. In all cases but the non relevant document (N) in the NMRH case the median rank inferred from the ME scores (black lines) are equal to the rank inferred from the expert judgements (red diamonds). For the MRMR ordering, the unbounded scores follow the expert relevance pattern (up-down-up), while the bounded scores are increasing from the first to the fourth document. However, this inconsistency may be due to an ordering effect in the MRMR setting, since only a single sequence of documents was used; in future work we plan to study this issue by testing further document orderings. Overall, we therefore conclude that crowdsourced ME judgements can indeed be used to obtain document-level relevance scores, and that the scores rank documents consistently with ratings made by expert judges on an ordinal scale.

Another possible analysis of consistency between the ranks

should involve more median relevance documents and check whether the ME results cannot be separated from each other

	NMRH		HRMN	
	Unbounded	Bounded	Unbounded	Bounded
N < M	50	54	72	50
M < R	62	73	61	67
R < H	58	69	72	71
N < R	75	81	72	83
M < H	62	85	83	79
N < H	75	88	94	88

Table 2: Percent of workers whose scores agreed in order with the experts in each for document pairs in the NMRH and HRMN condition.

of documents inferred from ME scores and those inferred from the expert judgements is to look at all pairs of documents that are judged by a worker, and count how many are consistent with the expert judgements. For example, for the first and last documents in the NMRH condition, if ME is consistent with the expert judgements we would expect that the score for the first document would be less than the score for the fourth for all workers. Table 2 shows the percentage of workers who assigned scores that agreed with the expert orderings for each document pair in the NMRH and HRMN conditions. If crowdsourcing ME judgements perfectly agreed with the expert judgements, we would expect this table to contain all 100% entries. This is clearly not the case. When the categorical score assigned by experts differs by only one (N<M, M<R, R<H) the percentage of workers that assigned scores in agreement with these orderings can be as low as 50%. When there is a clear gap between relevance categories (N<H), the agreement is closer to 90%.

This analysis highlights one of the potential traps of using crowdsourced ME data (or indeed, crowdsourced data in general): presuming that many workers do not actively engage in the task, one should only use aggregated data, and not individual data to form judgements (as is also discussed in other studies [1, 6]). Hence the median over workers shown in Figure 2 agrees with the expert judgements, but the individual data in Table 2 does not.

5.2 Normalised Magnitude Estimation Scores

The ME scores, normalised using geometric averaging as explained in Section 4.2, are shown in Figure 3 for the unbounded scale (left) and bounded scale (right). The scores are pooled across all ordering conditions and workers.

Since the ME scores are real numbers, we can test whether the perceived differences between the document relevance levels differ statistically significantly for the different expert judgement levels. In other words, we can check whether the previously assumed differences between the levels of the 4-point ordinal scale are in fact reflected in the perceptions of the participants.

The normalised ME scores using the unbounded variant (no maximum number in the assigned scores) are on a ratio scale, so it would be possible to carry out parametric statistical tests of the differences in perceived relevance levels. However, it is not clear that this assumption holds for the bounded scale, so we simply used the Wilcoxon signed rank test, a non-parametric test of whether two samples represent populations with different median values. In the analysis

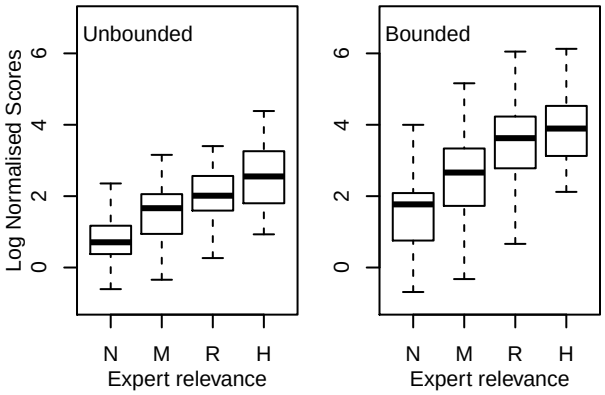


Figure 3: Log normalised scores for each category of expert judgement pooled over all conditions and all workers.

Scale	Comparison (A vs B)	Mean A	Mean B	Wilcoxon
Unbounded	N vs M	0.715	1.443	$p < 0.0001$
	M vs R	1.443	2.129	$p = 0.0001$
	R vs H	2.129	2.649	$p = 0.0073$
Bounded	N vs M	1.579	2.395	$p < 0.0001$
	M vs R	2.395	3.489	$p < 0.0001$
	R vs H	3.489	3.902	$p = 0.0469$

Table 3: Log normalised scores for each category of expert judgement pooled over all conditions and all workers.

that follows, the standard threshold of $p < 0.05$ is used to decide whether a difference is "statistically significant".

The results of the analysis of whether the ordinal expert judgements (on a 4-point scale) correspond to consistent perceived differences in relevance magnitudes and shown in Table 3. For both the unbounded and the bounded scales, the perceived differences between the ordinal relevance levels as measured using ME are statistically significant. This provides evidence that the assumptions of the ordinal relevance scale are borne out by the perceptions of the crowdsourced assessors.

Since ME judgements are made on a continuous (ratio) scale, they also enable a deeper analysis of the relationship between the categories of the ordinal scale (third research question). For example, it is now possible to answer the question: is the perceived difference between non-relevant and marginally relevant documents the same as the difference between relevant and highly relevant documents? To investigate this, the median ratios between the s^*_{jd} scores of adjacent ordinal category groups were calculated. From this, an inferred relevance scale can be obtained, by anchoring the ME score of the highest ordinal category (H) at 1. Note that the units of the scale are arbitrary: it is the ratios that are meaningful. The inferred relevance scores are plotted in Figure 4, for judgements obtained using the two ME scales (bounded and unbounded), and for the increasing (NMRH) and decreasing (HRMN) document orderings.

A consistent trend that can be observed in all results is

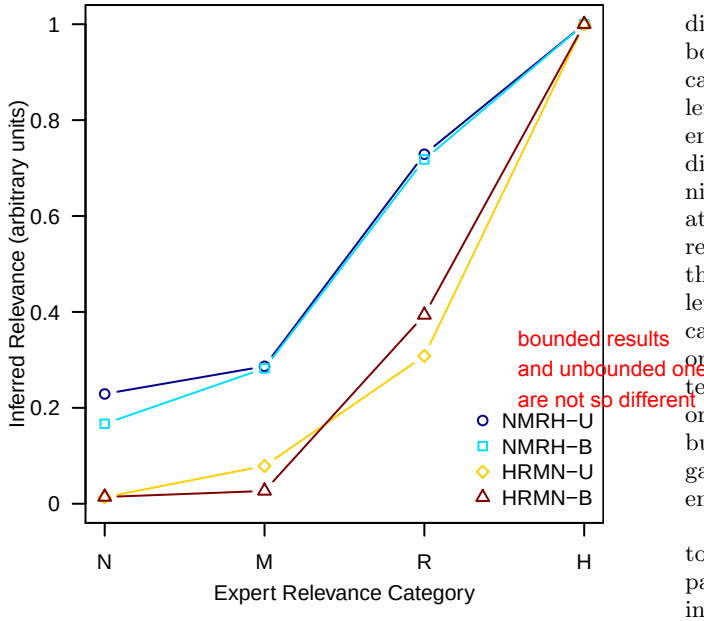


Figure 4: Inferred relevance, derived from the median ratio between ME scores for each category of the ordinal 4-point scale. An arbitrary level of 1 is assumed for H. The data is discrete, lines are added for presentation purposes only.

that the difference between the lowest two ordinal groups (N and M) is much smaller than the difference between the middle (M and R) or highest ordinal groups (R and H). This indicates that participants perceived a much smaller difference between the relevance of non relevant and marginally relevant documents, than between documents at the other levels. This is consistent with observations made in previous work [11], which analysed the relevance thresholds of users when searching with systems instantiated at different levels of P@1, and concluded that when multiple-level relevance scales need to be folded into a binary scale, marginal and non relevant documents should be bundled together.

5.3 Ordering Effects

To investigate the fourth research question on ordering effects, participants were asked to judge documents shown in one of four orderings, based on the underlying expert relevance judgements. The normalised ME scores are shown in Figure 5, for the unbounded scale (left) and bounded scale (right). Within each scale grouping, the four panels indicate one of the document orders, as shown on the x-axis.

The trends in the figure suggest that ordering effects are present. For example, for the unbounded scale, the ME scores for the NMRH are much closer together than for the HRMN ordering. Some differences, particularly on the unbounded scale, also appear to be asymmetrical between the NMRH and the HRMN conditions.

Repeating the analysis of differences in ME scores between the ordinal relevance levels, but taking ordering into account, shows that for the NMRH document order and the unbounded ME scale, there is no significant difference between ordinal levels N and M, and no difference between levels R and H, while there is a significant difference between levels M and R. In contrast, for the HRMN ordering, the ME scores differ significantly between all four ordinal levels. For the bounded ME scale, the NMRH document order shows significant differences between N, M, and R levels, but not between levels R and H. However, for the HRMN ordering, no difference is found between N and M, while there are significant differences between all other ordered pairs. Given the significant differences between all four ordinal relevance groups at a global level, as shown in the previous subsection, these results are unexpected. However, it should be borne in mind that far fewer observations are available for analysis at the level where ordering is taken into account, so it may be the case that a lack of statistical power is the cause of finding only a few differences. For example, a power analysis for a test of the difference between the ME scores at the N and M ordinal groups shows power of 0.99 at the unordered level, but only 0.16 at the ordered level. In future work we plan to gather more data and re-visit the issue of significant differences between ME scores when order is taken into account.

Another way to consider ordering effects is with reference to the inferred relevance scale, as shown in Figure 4. In particular, the decreasing (HRMN) ordering (the top two lines in the figure show this ordering with the bounded and unbounded ME scores) leads to a much smaller perceived difference between the top two ordinal relevance groups, H and R, being only around half of the perceived difference between these two groups for the increasing (NMRH) ordering. This suggests that there is indeed a relationship between presentation order and the perceived difference in relevance as measured by ME. One possible explanation is that this may be related to a priming effect, as has also been found to occur when users make judgements on an ordinal scale [10].

Eisenberg, studying relevance ratings of curated document descriptions, concluded that that ratings obtained using ME may be less subject to ordering bias than ratings obtained on an ordinal scale [4]. However, as noted in his analysis, the results possibly lacked statistical power.

With regard to the fourth research question, it would appear that the ME judgements on both the ordered and unordered scales are potentially subject to ordering bias.

6. CONCLUSIONS AND FUTURE WORK

The technique of ME has been used to study the perception of stimuli at different levels of intensity. In this paper, we have investigated the application of the technique to the gathering of relevance judgements at the document level, a key component of test collection-based IR evaluation.

Our preliminary analysis using a small data set suggest that the ME technique is promising: relevance judgements made using ME are consistent with ordinal judgements made by expert assessors. This held for different document presentation orders, and for using either a bounded or an unbounded relevance scale. The one exception was that for the MRMR ordering, the bounded scale showed a slight inconsistency, increasing with presentation order, rather than showing a "zig zag" trend as required by the underlying expert judgement. The unbounded scale did not show this inconsistency.

Ratings obtained using ME are on a ratio scale, and can therefore be used to investigate the extent to which perceptions of relevance actually differ between the ordinal categories assigned by experts. The ME scores assigned by participants at the four ordinal levels, across all users and orderings, showed statistically significant differences. This

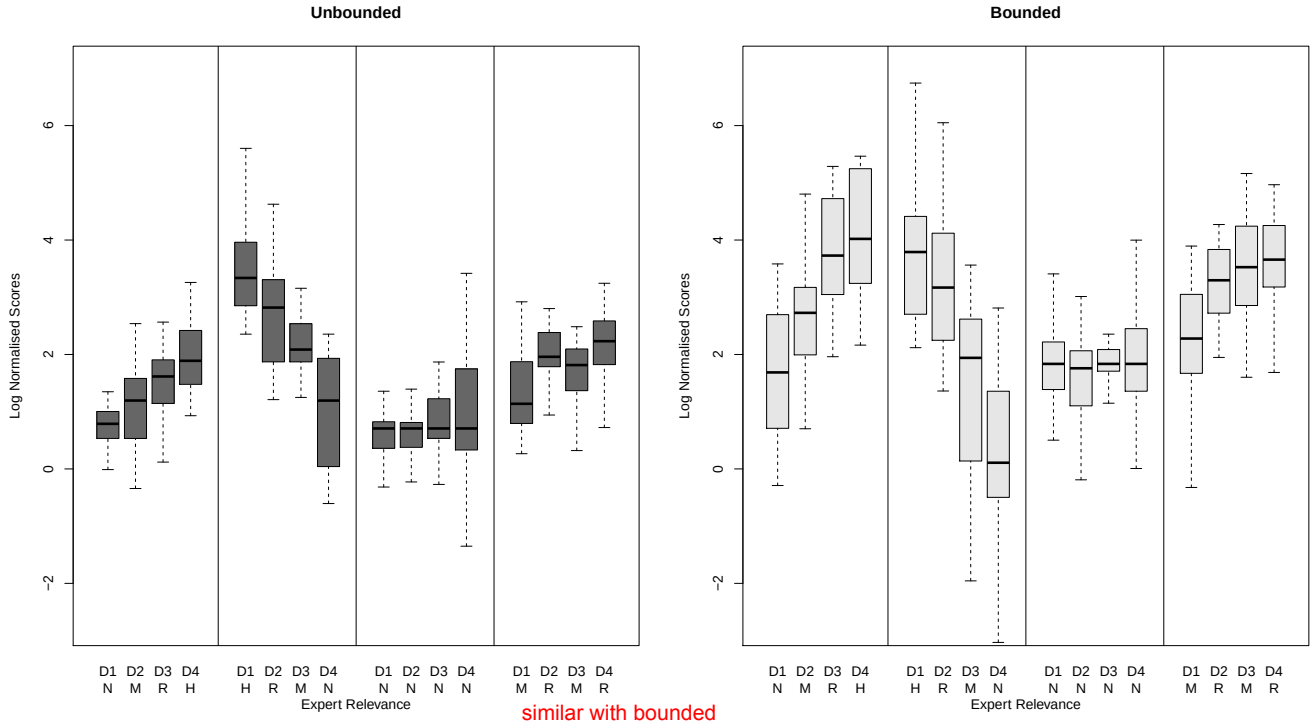


Figure 5: Log ME scores after geometric averaging. Scores for the bounded scale are shown on the left, and for the bounded scale on the right. Within the scale groups, the four panels show results for the different document orderings based on the ordinal expert judgements. D1 to D4 indicates the presentation order of documents.

validates the ordinal scale, providing evidence that when judges assign documents to "not relevant", "marginally relevant", "relevant" or "highly relevant" categories, these correspond to different perceptions of the relevance content.

Moreover, ME enables the difference between items on a nominal relevance scale to be made. This analysis showed that the difference in the perceptions of relevance between the lowest two ordinal relevance categories were much smaller than between the other categories. From a relevance perspective, this implies the difference between non relevant and marginally relevant documents is much smaller than the difference between marginally relevant and relevant, or relevant and highly relevant.

Analysis of ordering effects indicated that ME relevance judgements may be influenced by presentation order. A marked example of this occurs when using the unbounded ME scale for the `NMRH` and `HRMN` orderings. Moreover, the perception of the difference between the two highest ordinal relevance categories, relevant and highly relevant, was substantially affected by presentation order. This may be due to a number of factors: inherent ordering differences; priming effects; the choice of unbounded or bounded ME scale; or a lack of statistical power, and requires further investigation.

Another research question of this paper was whether the choice of bounded or unbounded ME scale has an impact on the quality of the obtained relevance judgements. As the summary of the results above suggests, there is not yet enough evidence to draw a firm conclusion about this.

This preliminary investigation suggests a lot of interesting future work. Of course, a larger analysis, involving more topics, more documents, and more workers is mandatory. We plan a more controlled investigation of ordering effects, e.g., by including the reverse ordering of `MRMR`. Once the reliability of ME is assessed, it can be a powerful tool to understand whether the categorical/ordinal relevance scales that have been used in past evaluation exercises are indeed grounded on a hidden relevance ratio scale. This would mean that it makes sense to speak of "amount of relevance", and this would potentially lead to reconsideration of many assumptions made in IR, ranging from the choices made in IR effectiveness metrics (e.g., gain levels in nDCG) to the "probability of relevance" concept at the basis of all probabilistic models. In this respect, a specific target would be to apply ME to investigate whether the TREC Web Track 2009–2013 relevance construct, which seems to conflate navigational relevance and informational relevance into 5 relevance levels, is actually an operational ordinal scale.

More generally, we also aim to use ME to re-visit the debated issue of whether there is an "optimal" number of ordinal relevance levels, or whether judgements should be made using ME in general. Applying ME to measure non-topical aspects of relevance might also be interesting. Finally, in the future ME relevance judgements could also be used in (specific tracks of) test collection evaluation initiatives.

7. ACKNOWLEDGEMENTS

Anonymised for review.

Appendix A. Full Instructions Shown To Participants

In this task, you will be shown a statement that expresses a need for information. This will be displayed at the top of the screen.

Please read the statement carefully. Next, you will be asked a question about the statement, to test your understanding.

You will then be shown 4 documents that have been returned by a search system, in response to the information need.

Your task is to indicate how RELEVANT these documents appear to you, in relation to the information need.

As a preliminary exercise, can you imagine a document which would be highly relevant? Can you imagine a document that you would judge to be low in relevance? Can you imagine a document that you would judge to be medium in relevance?

Now do the same for numbers. Imagine of a large number. A small number. A medium number.

As indicated above, you will be shown 4 documents, one at a time. Your task will be to assign a number to every document in such a way that your impression of how large the number is matches your judgment of how relevant the document is.

Write the number for each document in the box under the document description.

- [Item shown for the Unbounded scale only] You may use any numbers that seem appropriate to you – whole numbers, fractions, or decimals. However, you may not use negative numbers, or zero.
- [Item shown for the Unbounded scale only] Don't worry about running out of numbers – there will always be a larger number than the largest you use, and a smaller number than the smallest you use.
- [Item shown for the Bounded scale only:] You may use any numbers strictly greater than zero (i.e., zero excluded) and up to 100 (100 excluded) – whole numbers, fractions or decimals.
- Treat each document individually and don't worry about the numbers assigned to preceding documents.
- You are able to indicate your best judgment of relevance of a document description at the time it is presented to you. Respond as spontaneously as you can. Do not go back to documents you have already rated.

Is it required by ME or because this will hurt the effect of ordering?

8. REFERENCES

[1] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for TREC relevance assessment. *Information Processing and Management*, 48(6):1053–1066, November 2012.

[2] Charles Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2004 terabyte track. In *The Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, 2005. National Institute of Standards and Technology.

[3] V. Della Mea and S. Mizzaro. Measuring retrieval effectiveness: A new proposal and a first experimental validation. *J of the American Society for Information Science and Technology*, 55(6):530–543, 2004.

[4] Michael Elsenberg. Measuring relevance judgements. *Information Processing and Management*, 24:373–389, 1988.

[5] George Gescheider. *Psychophysics: The Fundamentals*. Lawrence Erlbaum Associates, 3rd edition, 1997.

[6] Mehdi Hosseini, Ingemar J. Cox, Nataša Milić-Frayling, Gabriella Kazai, and Vishwa Vinay. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Proceedings of the 34th European Conference on Advances in Information Retrieval*, ECIR'12, pages 182–194, Berlin, Heidelberg, 2012. Springer-Verlag.

[7] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[8] Mick McGee. Usability magnitude estimation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 47(4):691–695, 2003.

[9] S. M. Pollack. Measures for the Comparison of Information Retrieval Systems. *American Documentation*, 19(4):387–397, 1968.

[10] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S. Lee, and William Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 623–632, New York, NY, USA, 2013. ACM.

[11] Falk Scholer and Andrew Turpin. Metric and relevance mismatch in retrieval evaluation. In *The Fifth Asia Information Retrieval Symposium (AIRS 2009)*, pages 50–62, Sapporo, Japan, 2009.

[12] Eero Sormunen. Liberal relevance criteria of TREC: Counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 324–330, New York, NY, USA, 2002. ACM.

[13] Amanda Spink and Howard Greisdorf. Regions and levels: Measuring and mapping users' relevance judgments. *J of the American Society for Information science and Technology*, 52(2):161–173, 2001.

[14] Stanley S Stevens. A metric for the social consensus. *Science (New York, NY)*, 151(3710):530–541, 1966.

[15] Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103 (2684):677–80, 1946.

[16] Rong Tang, William M Shaw, and Jack L Vevea. Towards the identification of the optimal number of relevance categories. *J of the American Society for Information Science*, 50(3):254–264, 1999.

[17] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[18] Ellen M. Voorhees and Donna K. Harman. *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[19] Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M Jose, and Leif Azzopardi. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval*, 16(2):267–305, 2013.