Experiential Design and Research Plan

Ziying Yang
ziyingy@student.unimelb.edu.au
692959

Part A

The experimental design of paper reviewed in paper is *The benefits of magnitude estimation relevance assessments for information retrieval evaluation* [1].

1. A psychological scaling technique, *magnitude estimation* is involved for gathering document-level relevance judgments (used in Information Retrieval to assess the performance of IR systems) from crowdsourced users for TREC-8 [10]. There are three hypotheses being tested:
   (a) magnitude estimation is an appropriate technique for collecting relevance judgments;
   (b) using magnitude scores, the ranking of top systems that is evaluated by binary (TREC) [10] or Sormunen judgments [7] will not change;
   (c) compared to TREC binary and Sormunen judgments, magnitude scores provide better understandings of document relevances and the gain function used by gain-based evaluation metrics such as nDCG can be obtained (or calibrated) for crowdsourced users.
2. The hypotheses above can be regarded as successful by criteria described as follow:
   (a) relevance judgments gathered by the magnitude estimation (called magnitude scores) respect (highly agree with) the ordinal relevance judgments that we already have and use currently for document-level effectiveness scoring;
   (b) The variations of the rankings for top-performing systems evaluated using magnitude-based, TREC and Sormunen judgments are low;
   (c) the settings of gain profile which represents user perceptions of relevance should be found according to the ratios of magnitude scores assigned to the Sormunen categories. Although they are different for distinct type of users, they should be found as closer to one of the *linear*, *exponential* and other settings.
3. The agreements of relevance judgments using binary scale (named as TREC in this paper), Sormunen categories and magnitude estimation are measured in
   - document level: the proportion of document pairs that agree in terms of relevance order in the three judgments are computed and plotted (and so compared with each other);

[1] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. *In Proc.* SIGIR, pages 565 - 574, Santiago, Chile, 2015

- system level: (a) correlations of system orderings assessed according to three sets of judgments are tested using Kendall's τ; (b) the variations in the top of ranking lists using three judgments are measured by the overlap between the top part of three lists, computed by a paired Wilcoxon signed-rank test with $p < 0.05$.

4. (a) No special software was implemented for this paper. But JavaScript codes were created to build the task on the CrowdFlower crowdsourcing platform.

   (b) The magnitude scores for documents assigned by users are collected. In addition, users are allowed to go back to view the previous documents by clicking *back* buttons. The back button usage is recorded as well. The further data that could have been collected is the time spent and the range of scores used on each unit, by each user. The type of users (for instance, *generous* or *parsimonious* users) could then probably be found via this data.

   (c) The judgments of three relevance scales: magnitude estimation, binary and Sormunen category are compared in levels of (a) document: for each document, its magnitude score is compared with its binary score and Sormunen category respectively (shown in Figure 1 of the paper); (b) topic: for each topic, document scores of three judgments are compared (shown in Figure 2); (c) document-pair: the agreement of relevance ordering for each document pair between judges is computed and plotted in Figure 3; (d) system score: system scores evaluated based on three judgments, assessed by metric nDCG@10 and ERR@10 respectively, are calculated and correlations (Kendall's τ) between system scores are compared in Figure 4 and 5; (e) system ranking: the correlations of system rankings using three judgments and two metrics are measured by Kendall's τ and compared in Figure 6.

   (d) The user's perception of relevance can be observed via this experiment. According to the distribution of magnitude scores for each Sormunen category, the gain function (mapping the relevance categories/scores to the *weights* according to the intensity of relevance perceived by humans for each level) used in gain-based IR evaluation metrics can be obtained.

   As mentioned in 4(b), the type of users might have been found via assessing time or score range. The gain functions can also be calibrated according to distinct distributions of magnitude scores given by different types of users.

   (e) The baseline used in this experiment is the TREC and Sormunen judgments that are commonly used for IR evaluation. The newly proposed magnitude judgments are compared with these two sets of judgments in several ways, as described in 4(c). The judgments using other relevance scales such as multi-level and preference could also have been collected and used as baseline in this experiment.

   (f) In the crowdsourcing platform, CrowdFlower, users (participants) are paid to complete human intelligence tasks (HITs). First, users need to pass two quality control tests (see section 5(a)). Then in each HIT, *units*, each of which is composed by 1 topic and 8 documents (randomly selected from the document pool for this topic, but ensuring that each document will be assessed), are randomly assigned to participants. Participants can assign any positive score to the first document in the given unit and judge the next document with regard to the previous one during at least 20 seconds. After participant

finishes score assigning in crowdsourcing platform, score for each document is then be normalized by the formula given by the paper's authors.

5. Variables that may have influenced the outcomes:

   (a) **Assessors/Users/Participants:** as the experiment is done through the crowdsourcing platform, the types of users are unknown. Some users may have trouble with completing the task (for example, they do not know about ME, they do not have sufficient knowledge or background to assess documents fairly, or they misunderstand the given topic, or they lose consistency, or they do not really care and so on), which leads noise to the relevance data that collected. The qualities of assessments are controlled by four tests in the task:

   - participants are required to complete a practice task to show that they have understood ME;
   - participants need to correctly answer a multiple choice question which focuses on the key concepts of the given topic;
   - a known highly relevant document H_k and a known irrelevant document N_k are included in each unit and the judge for a unit is deemed to be qualified if the assigned score of H_k is higher than the score of N_k;
   - users have to spend at least 20 seconds on assigning score for each document.

   (b) **Topics and documents:** contents of topics and documents could influence the assessing of users. Topics and documents used in this experiment are selected from a well-known text collection, TREC. The document and topic contents in TREC-8 are somewhat dated, thus the assessing for young workers may be more difficult. The documents are collected from USA so users who are not American may not be familiar with the contents. The select method is not described in the paper but the authors aim to control this variable by filtering out inappropriate testing data for the task. In addition, each *unit* is composed of only 8 documents. Units are randomly assigned to users and documents in a unit are presented to users in a random order as well. But these methods do not control all the influences described above. The paper's authors probably should have picked documents and topics from several different text collections and tracks.

6. - The TREC-8 documents and topics have been selected for this experiment because some of them do not have Sormunen judgments and so the comparing of magnitude and Sormunen results cannot be done for them. The selection leads that the number of documents (and so unit number) for each topic is different. This selection may or may not affect the results by magnitude estimation, which is not considered, tested or discussed by authors. Other text collections and other relevance judgments (despite TREC and Sormunen) that contain the full data could have been used in this experiment.
   - Some documents are judged as highly relevant by users using magnitude estimation but they are classified into the *not relevant* category by Sormunen or are not in the pool (and so were not judged) for TREC (these documents are obvious found in Figure 1). They indicate that either magnitude judgments or Sormunen and TREC judgments may contain noise. Alternatively, the distinction may be due to different types of assessor. It may affect the effectiveness of the magnitude estimation but hard to be tested and verified. These document-topic pairs whose judgments are disparate using distinct relevance scales could have been selected out and checked again. The paper's authors should have been more careful when treating these pairs and comparing the judgments.

Part B

What are we trying to do With an IR text collection, TREC-8, we firstly break documents for a topic into pairs. If there are n_t documents for the topic t, then $n_t(n_t - 1)/2$ pairs will be generated.

Then document pairs will be randomly distributed to assessors (users) in a crowdsourcing platform, CrowdFlower. We can also control the number of assessors for each pair (depends on the budget) using tools provided by CrowdFlower. The user will be expected to read the specification and the topic carefully and have to pass tests for quality control before they can start a task. During the task, assessors need to make a preference choice of which document is more relevant to the topic than the other by clicking the button for the preferred document. The preference data is recorded by CrowdFlower and can be downloaded after all pairs have been assessed.

Ranking learning algorithms will be employed for generating an ideal list for each topic according to preference judgments collected from CrowdFlower. The ranking lists given by systems will be compared with the ideal list and evaluated by IR metrics such as *ppref* and *rpref* proposed by Carterette and Bennett [3], which have been extended for preference judgments.

For each topic, judgments of pairwise preference, TREC and Sormunen will be compared in levels of

1. document pair: for pairs that contain two documents which were classified into distinct relevance categories in TREC and Sormunen, check their relevance order in preference judgments. High agreement between pairwise preference and absolute relevance judgments will testify that preference judgments are suitable for gathering document relevance for IR.

2. system grouping: systems will be divided into groups by their performance scores assessed using these three judgments. High agreements of system grouping show that the collected preference judgments respect the knowledge of relevance judgment that we currently have and use.

3. system ordering: the agreement (overlap) of system rankings sorted by system scores evaluated using these three judgments can be calculated and compared; the low agreements of top system orderings will indicate that preference judgments are useful and make differences.

Review of current works Arguello et al. [1] proposed and used preference judgments for aggregate search results of text, photos, video and so on. They divided the document list into several blocks and create other blocks for different types of media results retrieved for the given topic. Then collect the ordering preference of blocks from participants in a crowdsourcing platform, Amazons Mechanical Turk (AMT). Then they used Schulze voting method [6] to generate the ideal ranking list, *reference*, and evaluated system performance by measuring distance between their ranking lists and the reference using Kendall's τ, a well-known rank-based correlation metric.

The preference judgment was not compared with other competitive methods to show that it was suitable to collect judgments for IR from real users. Instead, for each topic, they collected pairwise preferences of ranking lists given by systems from AMT participants again and compare the results with Kendall's τ scores (the distances between the reference and ranking lists) of systems. This method is expensive and cannot be used for big datasets. It is impossible to let users compare

whole ranking lists at document level. Thus, further improvements are needed if we want to collect preference judgments for document relevance.

What's new and why it will succeed Relevance judgments using absolute relevance scales assessed by *experts* have been argued as unreliable and not objective [2, 4, 5, 8] for IR evaluation. The problems arise because:

- The judgments were not collected from real users, but from experts in this area.
- No matter which absolute relevance scale is chosen, the distinction between relevance levels expected by users may be disparate. User perceptions of relevance are hard to incorporate into a single relevance scale [9].
- Assessment consistency may be lost as more documents are read and evaluated.
- It is impossible to tell which document is preferred to another if they receive the same relevance scores (or are classified into same categories).

Arguello et al. [1] proposed preference judgments that only record which element is preferred in each pair. Preference judgments have not been used to collect relevance preference from real users in document level in IR area. If we employ this technique as a relevance scale to collect document relevances from participants on crowdsourcing platform, all problems listed above may be solved. Assessors do not need to interpret differences between relevance levels but only need to make a preference choice for each pair of documents instead of assigning relevance scores or classifying them into categories. Compared to absolute judgments that are commonly used, using preference judgments will reduce the assessing complexity and improve the assessing consistency because only two documents are compared each time and users do not need to remember previous assessed documents. More than that, with the ideal ranking list which is obtained by preference judgments of all document pairs, the preference between any two documents can be discovered and so there is no tied document relevance in the judgment.

Who cares If preference judgments prove to be useful for collecting document relevance from real users,

- users of IR system may receive better search results for given queries because systems will have been adjusted and improved using preference of real users;
- organizers of text collections for IR such as TREC may try to collect preference judgments for their documents and topic;
- IR system developers may test and improve their systems with preference judgments to meet information needs of real users;
- IR evaluation metrics may need to be proposed, extended or adjusted for preference judgments;
- program developers who implement software to evaluate IR systems may need update the program for considering preference judgments.

What difference will it make If preference judgments are proven to be useful for collecting relevance judgments, two benefits will be obtained: (1) we can have a better understanding of user's perceptions of relevance; (2) the relevance assessing accuracy can be improved due to the reduction of noise associated with the distinction of user's perceptions of relevance and the assessing complexity; (3) the organizer do not need to choose a specific relevance scale for assessing documents, which may lead unpredictable relevance judgment variations.

However using preference judgments, additional processes for (1) breaking documents into pairs; (2) waiting participants from crowdsourcing platform to complete assessing tasks; (3) generating ideal lists are necessary. More than that, traditional IR metrics such as Precision and Recall need to be extended for preference judgments.

Risk and payoffs As we cannot control that how many users will participant and complete tasks we published, so the time spent on collecting judgments from CrowdFlower cannot be guaranteed. But if tasks are successfully completed, we will receive relevance judgments from users.

As this is the first time that using preference judgments collected from crowdsourcing platform for IR evaluation, results of the experiment may reject our hypothesis. But we can still get better understandings of the relevance preference of human and improve the experiment.

Cost If we use the same dataset as Turpin et al. [9] used for magnitude estimation (see Table 1 in their paper), there are 18 topic and for each topic t there are $n_t(n_t - 1)/2$ pairs, so totally we have 560,327 pairs of documents need to be assessed. If we set 60 pairs as a unit and pay \$0.2 for each, then totally the cost of assessing will be $\$0.2 \times 9338$ units = \$1867.

How long The time for this experiment is composed by time for (1) developing code for breaking documents into pairs and generating data file for CrowdFlower; (2) participants on CrowdFlower to complete all the tasks; (3) developing code for learning algorithm to obtain ideal ranking list according to the collected preference judgments; (4) developing code for measuring systems using preference judgments; (5) comparing results using preference judgments and judgments collected by other relevance scales.

Exams After collecting judgments for all pairs from CrowdFlower, a midterm exam can be done by checking the agreement between preference and TREC binary judgments. Select the pairs that contain one *relevant* and one *irrelevant* document assessed using TREC binary relevance scale. Check the relevance ordering of documents in those pairs in preference judgments and calculate the proportion that preference judgments agree with the TREC. And similarly, for pairs whose documents are from different Sormunen categories, test how many pairs whose documents have the same relevance ordering in preference judgments. The high agreements indicate the success of employing preference judgments for collecting document relevance in IR.

The previous exams test the eligibility of preference judgments. The final exam, which will be hold after obtaining the ideal ranking list, tests if using preference judgments will make differences to the IR evaluation as well as allowing better understanding of users' perceptions of relevance. The systems will be sorted by scores evaluated using preference, TREC and Sormunen judgments. The overlap and correlation of the top part of these three lists will be measured by Kendall's τ and Rank Biased Overlap. We expect that the top systems are almost same in these three lists, but their orderings may be different. The system ordering using preference judgments represents more

about the opinions of participants on crowdsourcing platform, instead of experts, with respect to the original information need.

References

[1] Jaime Arguello, Fernando Diaz, Jamie Callan, and Ben Carterette. A methodology for evaluating aggregated search results. In *Proc. European Conference on Information Retrieval*, pages 141–152, Dublin, Ireland, 2011.

[2] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: Are judges exchangeable and does it matter. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 667–674, Singapore, 2008.

[3] Ben Carterette and Paul N. Bennett. Evaluation measures for preference judgments. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–686, Singapore, 2008.

[4] Robert. A. Fairthorne. Implications of test procedures. *J. Information Retrieval in Action*, pages 109–113, 1963.

[5] Robert. V. Katter. The influence of scale form on relevance judgments. *J. Information Storage & Retrieval*, 4(1):1–11, 1968.

[6] Markus Schulze. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *J. Social Choice and Welfare*, 36(2):267–303, 2011.

[7] Eero Sormunen. Liberal relevance criteria of TREC: counting on negligible documents? In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–330, Tampere, Finland, 2002.

[8] Andrew Turpin and Falk Scholer. Modelling disagreement between judges for information retrieval system evaluation. In *Proc. Australian Document Computing Symposium*, page 51, Sydney, Australia, 2009.

[9] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 565–574, Santiago, Chile, 2015.

[10] Ellen M. Voorhees and Donna Harman. Overview of the eighth text retrieval conference (TREC-8). In *Proc. Text REtrieval Conference*, Maryland, USA.