INTELLIGENT RETRIEVAL
Open Journal

Review

Selecting Intelligent Topics for Cost-Effective Information Retrieval: A Fresh View on Deep vs. Shallow Judging

Mucahid Kutlu[1]; Tamer Elsayed[1]; Matthew Lease[2]

[1]*Department of Computer Science and Engineering, Qatar University, Qatar*
[2]*School of Information, University of Texas at Austin, USA*

[*]***Corresponding author***
Mucahid Kutlu
Department of Computer Science and Engineering, Qatar University, Qatar

Article information
Received: March 25th, 2024; **Revised:** April 2nd, 2024; **Accepted:** May 1st, 2024; **Published:** May 31st, 2024

Cite this article
Kutlu M, Elsayed T, Lease M. Selecting intelligent topics for cost-effective information retrieval: A fresh view on deep vs. shallow judging. 2024; 2(1).
doi: https://doi.org/10.70705/ppp.ir.2024.v02.i01.pp39-45

ABSTRACT

The foundation of Cranfield-based evaluation of information retrieval (IR) systems is test collections, but with today's massive document collections (e.g., ClueWeb12's 700M+ Webpages), it is practically impossible to use traditional pooling techniques to build test collections. This has spurred a surge of research suggesting more efficient and trustworthy ways to assess IR. In order to decrease the amount of search topics—and the associated costs of human relevance judgments—needed for credible IR assessment, we provide a novel intelligent topic selection approach in this study. We bring together two areas of study that were previously separate to conduct a thorough evaluation of our method: intelligent topic selection and deep vs. shallow judging. This refers to the question of whether it is more cost-effective to gather a large number of relevant judgments for a small number of subjects or a smaller number of judgments for a large number of topics. Previous research on the subject of intelligent topic selection has not been tested against baselines for shallow judgment; nonetheless, studies comparing deep and shallow judging have mostly supported shallow judging, supposing that topics are randomly selected. We contend that the ultimate question to address when assessing a subject selection approach is whether it is beneficial to choose topics or to conduct superficial judgments across several topics. We undertake an exhaustive research across a range of pertinent criteria never before examined together in an effort to arrive at a rigorous solution to this general question: 1) how to choose topics; 2) how familiarity with a subject affects the speed with which humans judge; and 3) the effects on budget usage and the quality of judgments made by various topic generating techniques (requiring varied amounts of human work). Experiments conducted on the NIST TREC Robust 2003 and Robust 2004 test collections demonstrate that IR systems can be accurately evaluated using fewer topics. The results also show that 1) intelligent topic selection makes deep judging more cost-effective than shallow judging in terms of evaluation reliability, and 2) the evaluation cost vs. reliability trade-off is heavily influenced by topic familiarity and topic gener-ation costs. By demonstrating that, with intelligent subject selection, deep judgment is often better than superficial judging, our results contradict common belief.

INTRODUCTION

Under the Cranfield paradigm, test collections are the backbone of system-based IR algorithm evaluations (Cleverdon, 1959). First, a collection of documents that will be searched; second, a list of topics that users have already decided to search for relevant information on; third, a brief description of each topic that can be used as a search query in an IR system; and finally, a set of documents that will be searched.

2) an algorithm that uses human-made relevance assessments to determine which collection documents are most relevant to a given search query. The continuous improvement of search algorithm performance is made possible by use of this test collection, which allows community benchmarking and empirical A/B testing of new search algorithms. It is common practice to combine the best-ranked documents from several systems and evaluate just those. This is due to the fact that it would be very expensive to evaluate every document in a realistic document collection. Incomplete judgment by pooling has a well-established track record of dependability, provided the pool depth is big enough (Sanderson, 2010). However, assessment results might be skewed if not enough documents are evaluated; for example, it could be assumed incorrectly that many unjudged papers are irrelevant while in fact they are (Buckley, Dimmick, Soboroff, & Voorhees, 2006). One major issue is that modern document collections are getting bigger and bigger all the time. Another is that in order to evaluate search algorithms realistically, they need to be tested on large enough document collections to be searched in practice. This ensures that results from lab evaluations translate to real-world applications. Inevitably, bigger collections include more relevant

(and apparently relevant) documents, which means that human relevance assessors are required to evaluate an increasing number of documents for each search subject. Results show that conventional pooling methods' assessment costs have skyrocketed (Sanderson, 2010). Therefore, developing novel assessment methods to lower evaluation costs without sacrificing dependability is a significant open problem in IR. Put simply, how can we make the most efficient use of the funds for IR evaluation?

Many research have looked at the question of whether, given an assessment budget, it is preferable to gather a small number of relevance judgments for a large number of subjects (WaS) or a large number of relevance judgments for a small number of topics (NaD) when it comes to relevance. As an illustration, in the TREC Million Query Track study (Carterette, Pavlu, Kanoulas, Aslam, & Allan, 2009), IR systems were applied to approximately 10,000 queries taken from two massive query logs. A subset of topics, for which a human assessor could infer some purpose from the query, underwent shallow judging. This allowed for the back-fitting of a topic description and the making of relevance determinations. On the surface, because

It is reasonable to test systems on a broad range of search subjects and questions since individuals use different queries to convey interest in different things. It is recommended to employ a wide variety of themes for system evaluations to ensure consistent results, since empirical evidence shows that search accuracy might vary significantly across various topics for the same system (Banks, Over, & Zhang, 1999). Anderson and Zobel (2005), Carterette and Smucker (2007), and Bodoff and Li (2007) are only a few of the previous research that found that, when comparing WaS with NaD judgment, the former generally provides a more reliable assessment for the same amount of human effort. Although there are certain circumstances when this result does not apply, the number of exceptions has been quite small. In their study, Carterette, Pavlu, Kanoulas, Aslam, and Allan (2008) found that a total of 5000 judgments over 250 themes was just as reliable as 600 judgments across 10 topics. In this study, we highlight an important point: all previous research comparing NaD vs. WaS judgment has assumed that search subjects are chosen at random.

A further line of inquiry has focused on intelligent topic selection, the process of selectively include search subjects in a test collection with the goal of reducing the total number of search topics required for a reliable assessment. Reducing the number of subjects used immediately lowers judging costs, since human relevance assessments are required for each topic. Based on a simple, effective, and expensive topic creation process that involves collecting initial judgments for each candidate topic and manually selecting the final topics to retain, NIST TREC test collections have traditionally used 50 search topics (manually selected from a larger initial set of candidates) (Voorhees, 2001). Stable assessment requires at least 25 subjects (preferably 50), according to Buckley and Voorhees (2000), while Zobel (1998) demonstrated that one set of 25 questions reasonably predicted relative system performance on another set of 25 topics. A systematic research by Guiver, Mizzaro, and Robertson (2009) shown that comparing IR system evaluations spanning all themes and evaluations using the "right" selection of topics produces almost identical findings. Having said that, they failed to provide a strategy for actually identifying such a useful subset of topics. Along with Hosseini, Milic-

The iterative approach suggested by Frayling, Shokouhi, and Yilmaz (2012) for the purpose of discovering effective topic subsets has shown promising outcomes. We note that current favored technique for IR assessment cost reduction is using shallow judgment baselines, and that previous work on intelligent subject selection has not tested against these. Our main point is that we think people should just do WaS judgment across a large number of subjects instead of trying to decide which ones to choose.

Our Job. In order to decrease the amount of search topics—and the associated costs of human relevance judgments—needed for credible IR assessment, we provide a novel intelligent topic selection approach in this work. We combine two lines of research—one on intelligent topic selection and the other on NaD vs. WaS judging—in order to test our central hypothesis that topic selection is more beneficial than WaS judgment techniques. In particular, we look at a whole suite of important elements that have never been studied together before: First, the technique for selecting topics; second, how familiarity with a subject affects the speed with which humans judge; and third, the effects on budget utilization and judgment quality of various topic creation procedures (requiring varied amounts of human work). We point out that previous research on NaD vs. WaS judgment has ignored the potential financial effects of the relationship between judging depth and judging speed. In a similar vein, previous research on NaD vs. WaS judging has ignored the time required to build subjects; taking this into consideration might make WaS assessing many topics a lot less appealing (Voorhees, 2016). Therefore, our results contribute to the ongoing discussion over NaD vs. WaS judgment under the assumption of random subject selection. Research question one (RQ-1) is where we'll start: Given that different IR systems rate documents differently for each subject, how can we choose search terms that will optimize the validity of evaluations? Here, we provide a fresh take on subject selection using learning-to-rank (L2R). Section 4.3 details the iterative process of subject selection using a greedy strategy that maximizes correct system ranking. We utilize MART (Friedman, 2001) as our L2R model, although we're open to using alternative models if needed. For this subject selection assignment, we identify and extract 63 characteristics that reflect the relationship between topics and system ranking (Section 4.3.1). We provide a way to automatically create valuable training data from pre-existing test collections (Sec-tion 4.3.2) so that our model may be trained. Our method is more applicable and effective in a wide range of contexts since it relies only on existing test collections for model training, allowing us to build a new collection of tests without having to make any previous relevance decisions. Our method is compared to recent past work (Hosseini et al., 2012) and random topic selection (Section 5), and it is evaluated on the NIST TREC Robust 2003 (Voorhees, 2003) and Robust 2004 (Voorhees, 2004) test collections. Final Product

demonstrate steady progress compared to baselines, with relative progress increasing with decreasing subject use.

In addition to demonstrating that our approach of subject selection is an advance over previous work, we think it is crucial to evaluate intelligent topic selection in relation to the true central question: how can we accomplish cost-effective IR evaluation? Should we just conduct WaS judgment across a large number of subjects, or is intelligent topic selection truly useful? We utilize our intelligent topic selection process to do a thorough examination including a collection of targeted research topics that have not been addressed in the previous work. Our goal is to examine this:

2 Related Work

It takes a lot of time and energy to build test collections by hand. Researchers have so come up with a number of ideas to lessen the financial burden of making test collections. New assessment tools and statistical approaches for incomplete judgments are among the suggested approaches (Aslam, Pavlu, & Yilmaz, 2006; Buckley &

Voorhees, 2004; Sakai, 2007; Yilmaz & Aslam, 2006, 2008), finding the best sample of documents to be judged for each topic (Cormack, Palmer, & Clarke, 1998; Carterette, Allan, & Sitaraman, 2006; Jones & van Rijsbergen, 1975; Moffat, Webber, & Zobel, 2007; Pavlu & Aslam, 2007), inferring relevance judgments (Aslam & Yilmaz, 2007), topic selection (Hosseini et al., 2012; Hosseini, Cox, Milic-Frayling, Vinay, & Sweeting, 2011; Mizzaro & Robertson, 2007; Guiver et al., 2009), evaluation with no human judgments (Nuray & Can, 2006; Soboroff, Nicholas, & Cahan, 2001), crowdsourcing (Alonso & Mizzaro, 2009; Grady & Lease, 2010), and others. For a more comprehensive analysis of previous research on techniques for cost-effective IR assessment, the reader is directed to (Moghadasi, Ravana, & Raman, 2013) and (Sanderson, 2010).

2.1 Topic Selection

Mizzaro and Robertson (2007) conducted the first research that we are aware of that attempted to determine which themes would be the most suitable for review. Before using the HITS algorithm, they constructed a system-topic graph to depict the connection between IR systems and subjects. They postulated that those with more hubris would be better able to differentiate between systems. But experimental evidence provided by Robertson (2011) disproved their notion.

The experimental work of Guiver et al. (2009) shown that a ranking of systems may be achieved with a selection of topics that is extremely close to the ranking when all topics are used. Nevertheless, they failed to provide guidance on how to choose the most appropriate group of subjects. Other scholars have been inspired to delve more into this issue by this work. A well-chosen subset of subjects used to assess one set of systems may be sufficient to assess another set of systems, as shown by Berto, Mizzaro, and Robertson (2013), who placed an emphasis on generalizability. According to Hauff, Hiemstra, Azzopardi, and De Jong (2010), the Jensen-Shannon Divergence technique did not effectively minimize the number of subjects by utilizing the simplest topics. For the purpose of increasing the collection's re-usability, Hosseini et al. (2011) zeroed down on the subset of subjects to expand upon. Examining the evolution of subjects' predictive power in TREC test collections over time was the focus of Culpepper, Mizzaro, Sanderson, and Scholer (2014). If you're looking to save money on preference-based IR assessment, Kazai and Sung (2014) suggest adopting dissimilarity-based query selection.

(Hosseini et al., 2012) uses an adaptive algorithm to choose topics, which is the most similar research to ours. It chooses the first subject at random. After settling on a subject, the relevance judgments are gathered and used to guide the selection of related subjects. This means that in subsequent cycles, we will choose the subject that we think will have the highest current Pearson correlation. They do this by training a Support Vector Machine (SVM) model using the decisions from the subjects that have been chosen so far to estimate the likelihood of relevance of qrels for the ones that are still not included. At each cycle, the training data is enhanced by incorporating the relevance evaluations from each picked topic, allowing for improved topic selection.

Topic selection has been the subject of additional research for a variety of purposes, including but not limited to: improving supervised data fusion algorithms through data selection; developing low-cost datasets for training learning-to-rank algorithms; and estimating system ranks (Hauff, Hiemstra, De Jong, & Azzopardi, 2009; Lin & Cheng, 2011). Topic selection for low-cost IR system assessment is not taken into account in these research.

4.3.1 Features

In this section, we describe the features we extract in our L2R approach for each candidate topic. Hosseini et al. (2012) mathematically show that, in the greedy approach, the topic selected at each iteration should be different from the already-selected ones (i.e., topics in P) while being representative of the non-selected ones (i.e., topics in $\bar{P}$). Therefore, the extracted set of features should cover the candidate topic as well as the two sets P and $\bar{P}$. Features should therefore capture the interaction between the topics and the IR systems in addition to the diversity between the IR systems in terms of their retrieval results.

We define two types of feature sets. Topic-based features are extracted from an individual topic while set-based features are extracted from a set of topics by aggregating the topic-based features extracted from each of those topics.

The topic-based features include 7 features that are extracted for a given candidate topic tc and are listed in Table 2. For a given set of topics (e.g., currently-selected topics P), we extract the set-based features by computing both average and standard deviation of each of the 7 topic-based features extracted from all topics in the set. This gives us 14 set-based features that can be extracted for a set of topics. We compute these 14 features for each of the following sets of topics:

- not-yet-selected topics ($\bar{P}$)
- selected topics with the candidate topic ($P \cup \{tc\}$)
- not-selected topics excluding the candidate topic ($\bar{P} - \{tc\}$)

In total, we have 63 features for each data record representing a candidate topic: $14 \times 4 = 56$ features for the above groups + 7 topic-based features. We now describe the seven topic-based features

that are at the core of the feature set.

• Average sampling weight of documents ($\overline{fw}$): In the statAP sampling method (Pavlu & Aslam, 2007), a weight is computed for each document based on where it appears in the ranked lists of all IR systems. Simply, the documents at higher ranks get higher weights. The weights are then used in a non-uniform sampling strategy to sample more documents relevant to the corresponding topic. We compute the average sampling weight of all documents that appear in the pool of the candidate topic tc as follows:

$$f\sigma w\,(tc) = \sigma\{w(d, S) \mid \square d \square\ Dtc\ \} \quad (3)$$

4.4 Selecting A Fixed Number of Topics

In our first set of experiments, we evaluate our proposed L2R topic selection approach vs. baselines in terms of Kendall's τ rank correlation achieved as a function of number of topics (RQ-1). We assume the full pool of judg- ments are collected for each selected topic and evaluate with MAP.

Figure 1 shows results on Robust2003 and Robust2004149 collections. Given the computational complexity of Hosseini et al. (2012)'s method, which re-trains the classifier at each iteration, we could only select 63 top- ics for Robust2003 and 77 topics for Robust2004149 after 2 days of execution4, so its plots terminate early. The higher limit Overly fond of In Robust2003, Oracle managed to attain a τ score of 0.90, which is considered a satisfactory correlation (Voorhees, 2000). In Robust2004, the score increased to 20 topics. Unless 70% or 80% of topics are chosen, our suggested L2R technique beats Robust2003 baselines and Robust2004149 baselines. Reducing the number of subjects is accompanied with an increase in relative improvement over baselines. It seems that our L2R technique performs better with either a smaller set of subjects or a larger pool of topics from which to pick while working with a fixed set of topics.

Our next experiment takes a more frugal approach by considering a judgment condition where statAP is used to pick 64 or 128 papers to be assessed for each subject, rather than assuming the complete pool gets judged. Figure 2 displays the mean τ scores of all methods. Displayed in the vertical bars are the standard deviations across all trials. In general, our method is superior to the baselines in almost every trial and becomes better with fewer topics chosen, just as in the first set. When 70 or 80 topics are chosen for Robust 2003, our L2R technique performs rather poorly, same as the prior experiment with complete pooling. Our prior experiment yielded results that were somewhat poorer than random for 70 or 80 of these themes, while our L2R method here is on par with random selection (and slightly better).

In our studies, the topic selection approach proposed by Hosseini et al. (2012) actually performed worse than random, which contradicts their published findings. As a result, we looked into this thoroughly. We observed that, even though we meticulously followed their stated process for creating the baseline, our own random baseline performed, on average, $\tau \approx 0.12$ better than theirs across the 20 outcomes they published, using 10, 20, 30,..., 200 topics. We also compared our findings to those given by Guiver et al. (2009) and ran our random baseline on TREC-8 to further analyze this variation in baseline performance. Our findings were very congruent with those of Guiver et al. During our conversation with Hosseini et al. (2012), we learned that they took "special care when considering runs from the same participant"—which could mean that our two studies used different methods for preprocessing participant runs—which could explain why our results differ empirically.

Across both test sets, our method beats the baselines in almost every instance. We are able to train our approach using the abundant test sets made available by TREC and other shared task campaigns, even if the baseline methods don't need them. In addition, our tests show that we can leverage existing test collections in building models that are useful for constructing other test col- lections. This suggests that there are common characteristics across different test collections that can be leveraged even in other scenarios that are out of the scope of this work, such as the prediction of system rankings in a test collection using other test collections.

4.5 Feature Ablation Analysis

In this experiment, we conduct a feature ablation analysis to study the impact of each core feature and also each group of features on the performance of our approach.

We divide our feature set into mutually-exclusive subsets in two ways: core-feature-based subsets, and topic- group-based subsets. Each of the core-feature-based subsets consists of all features related to one of our 7 core features (defined in Table 2). That yields 9 features in each of these subsets; we denote each of them by $\{f\ \}$, where f represents a core feature. In the other way, we define 5 groups of the topics: the candidate topic tc (which has 7 core features) and four other groups of topics defined in Section 4.3.1 (each has a subset of features using average and standard deviation of the 7 core features, yielding a total of 14 features). We denote each of these feature subsets by F (g), where g represents a group of topics.

In our ablation analysis, we apply leave-one-subset-out method in which we exclude one subset of the features at a time and follow the same experimental procedure with the previous experiments using the remaining features. We evaluate the effectiveness of systems using MAP. For each subset of features, we report the average Kendall's τ correlation over all possible topic set sizes (1 to 100 for Robust2003 and 1 to 149 for Robust2004149) to see its effect on the performance. The results are shown in Table 4.

Table 4: Feature ablation analysis. The percentages in parenthesizes show how much the performance is decreased by removing the corresponding subset of features.

The table shows four interesting observations. First, $\{f\sigma\$\ \}$ and $\{f\sigma w\ \}$ are the most effective among the core- feature-based subsets, while F (P □ {tc}) and F (tc) are the most effective among the topic-group-based subsets, when testing on Robust2003 and Robust2004149 respectively. Second, $\{f\sigma\tau\ \}$ has the least impact in both test collections. Third, the feature subset of the candidate topic F (P □ {tc}) is the best on overage over all subsets,

which is expected as it solely focuses on the topic we are considering to add to the currently-selected topics. Finally, testing on both test collections, we achieve the best performance when we use all features.

4.6 Robustness and Parameter Sensitivity

The next set of experiments we report assess our L2R method's effectiveness across different training datasets and parameterizations. We evaluate the effectiveness of systems using MAP. In addition to presenting results for all topics, we also compute the average τ score over 3 equal-sized partitions of the topics. For example, in Robust2004, we calculate the average τ scores for each of the following partitions: 1-50 (denoted by τ1−33%), 51-100 (denoted by τ34−66%) and 101-149 (denoted by τ67−100%). These results are presented in a table within each figure.

Effect of Label Range in Training Set: As explained in Section 4.3.2, we can assign labels to data records in various ranges. In this experiment, we vary the label range parameter (K in Line 12 of Algorithm 3) and com- pare the performance of our approach with the corresponding training data on Robust2003 and Robust2004149 test collections. The results are shown in Figure 3. It is hard to draw a clear conclusion since each labeling range has various instances with different levels of performance. For example, with only five labels (i.e., Labeling 0-4), it performs well with a small set of themes. Its performance approaches that of a random technique as the number of subjects rises. In a combined analysis of Robust2003 and Robust2004149, it was shown that labeling 0-49, or 50 labels, produced the most consistent and superior outcomes. In most cases, employing 25 labels is preferable than using 10 or 5. Consequently, we find that our L2R method works better with fine-grained labeling.

The Influence of Tuning Dataset Size: Here, we test how well our method holds up when faced with a smaller pool of potential tuning topics. To do this experiment, we take Robust2003 and randomly choose 50 or 75 topics, respectively, and then filter out the ones that weren't picked. The R3(50) and R3(75) reduced tuning sets are what we call them. For the purpose of evaluating Robust2004149, we use these reduced tuning settings. We use a similar strategy while testing on Robust2003. In other words, we use the same method to choose fifty, seventy-five, and one hundred subjects at random from Robust2004149).

In order to get the average τ score, we carry out this procedure five times.

In Figure 4, you can see the outcomes. Each of the five trials' standard deviations is shown by the vertical bars. Tuning using all 100 subjects (i.e., real) yields the greatest performance over Robust2004149, as anticipated.

Robust2003); employing 75 topics is slightly better than employing 50 topics. Over Robust2003, when the number of selected topics is ≤33% of the whole topic pool size, tuning with 149 topics gives the best results. For the rest of the cases, tuning with 75 topics gives slightly better results than others. As expected, tuning with only 50 topics yields the worst results in general. Intuitively, using test collections with more tuning topics is seen to yield better results.

Effect of Test Collections Used in Training: In this experiment, we fix the training data set size, but vary the test collections used for generating the training data. For the experiments so far, we had generated 100K data records for each topic set size from 0-49 with TREC-9 and TREC-2001 and subsequently combined both (yielding 200K records in total). In this experiment, in addition to this training data, we generate 200K data records for each topic set size from 0-49 with TREC-9 and TREC-2001, and use them separately. That is, we have 3 different datasets (namely, T9&T1, T9 and T1) and each dataset has roughly the same number of data records. The results are shown in Figure 5. As expected, using more test collections leads to better and more consistent results. Therefore, instead of simply generating more data records from the same test collection, diversifying the test collections in present in the training data appears to increase our L2R method's effectiveness.

(a) Robust 2003 - 20 Hours (b) Robust 2004149 - 20 Hours

(c) Robust 2003 - 30 Hours (d) Robust 2004149 - 30 Hours

(e) Robust 2003 - 40 Hours (f) Robust 2004149 - 40 Hours

Illustration 11: Reusability Performance. We do 20 runs of statAP and average the tau scores to arrive at our technique. For the random technique, we choose 5000 questions and compute statAP once for each set of queries. Our assumptions include a topic generation cost of 76 seconds and an increase in judgment speed with the number of documents they evaluate. The vertical bars show the dispersion of the data. We used all of the themes, and the dashed horizontal line shows how well it worked.

assess IR systems using a reduced number of subjects; nevertheless, it should be noted that: 1) well chosen topics often result in more cost-effective assessment reliability using deep judging rather than shallow judging; and 2) the evaluation cost vs. reliability trade-off is significantly affected by topic familiarity and topic production costs. By demonstrating that, with intelligent subject selection, deep judgment is often better than superficial judging, our results contradict common belief.

To be more precise, our study's primary conclusions are as follows. Firstly, compared to the baselines, our suggested method almost always chooses superior themes that provide more accurate evaluations. Secondly, if subjects are chosen at random, corroborating results from previous studies, superficial judgment is better than deep judging. When themes are chosen well, nevertheless, deep judgment may typically get more reliable evaluations done with the same money as shallow grading. Third, there should be another parameter to consider in the deep vs. shallow judgment trade-off; this is because, if judging speed grows with the number of documents assessed for the same subject, increasing judging speed significantly affects assessment reliability. Fourth, thorough judgment is better than superficial judging when the cost of topic creation rises. Finally, it is preferable to collect fewer judgments of higher quality than more judgments of lower quality, given that short topic generation periods affect topic quality and, by extension, consistency of relevance judgments. The higher topic generation cost is another reason why deep judgment is often better than superficial judging.

In the future, we want to use qualitative analysis to determine what factors are at play when determining which topics are better at predicting the relative average performance of IR systems, and we will also look into how well our topic selection method works with other

evaluation metrics. This section draws inspiration from earlier qualitative research that aimed to identify the factors that made some subjects more challenging than others (Harman & Buckley, 2009). This newfound knowledge has the potential to pave the way for more efficient and trustworthy IR assessment in the future, as well as to inspire revolutionary ideas for the construction of subject sets.

REFERENCES

It was written by Allan, J., Aslam, J. A., Carterette, B., Pavlu, V., and Kanoulas, E. in 2008. Summary of the Million-Query-Track in 2008. This was presented at the 17th annual conference on text retrieval.
The authors of the report are Allan et al. (2007) and Carterette and Aslam and Kanoulas. Technical report on the Million-Query Track in 2007. Form DTIC.

In 2009, Alonso and Mizzaro wrote the article. Would trec assessors be able to be eliminated? using mechanical turk to evaluate significance. Volume 15, page 16, of the 2009 SIRG workshop proceedings deals with the topic of the future of IR assessment.

In 2006, Aslam, Pavlu, and Yilmaz published a work. A statistical approach to evaluating systems with imperfect judgments. Pages 541–548 of the 29th International ACM SIRGIR Conference on Research and Development in Information Retrieval Proceedings.

E. Yilmaz and J. A. Aslam (2007). Using missing data to infer the importance of a document. Volume 16, Issue 4, Pages 636–642, in The Proceedings of the Sixteenth Annual Conference on Information and Knowledge Management.

In 1999, Banks, Over, and Zhang published a paper. Elephants and blind men: six ways to look at trail data. Articles 7–34 in Information Retrieval, volume 1, issue 2, 2019.

Robertson, S., Berto, A., & Mizzaro, S. (2013). Concerning the matter of evaluating information retrieval with fewer subjects. In Volume 9, page 9, of the 2013 theory of information retrieval conference proceedings.

P. Li and D. Bodoff (2007). Theoretical framework for evaluating their test sets. Page numbers 367–374, collected from the 30th annual international acm SIGRID conference on information retrieval research and development.

In 2006, Buckley, Dimmick, Soboroff, and Voorhees published a work. Prejudice and the boundaries of amalgamation. Page numbers 619–620 from the 29th annual international acm SIRGID conference on research and development in information retrieval.
In 2000, Buckley and Voorhees published a book. Stability of assessment measures evaluated. Pages 33–40 of the 23rd International ACM SIRGIR Conference on Information Retrieval Research and Development Proceedings.

Publication year: 2004 by Buckley and Voorhees. Evaluate retrieval efforts when data is inadequate. Volume 27, Issue 1, Pages 25–33, of the International Conference on Information Retrieval Research and Development (SIGRAS) (27th Annual International Conference on).

An article published in 2006 by Carterette, Allan, and Sitaraman. Little datasets for retrieval assessment. Page numbers 268–275 from the 29th annual international acm SIRGID conference on information retrieval research and development.

In 2008, Carterette, Pavlu, Kanoulas, Aslam, and Allan published a paper. Thousands of requests were evaluated. Pages 651-658 are included in the proceedings of the 31st annual international acm SIGRID conference on information retrieval research and development.

In 2009, Carterette, Pavlu, Kanoulas, Aslam, and Allan published a paper. I would have a million questions. Pages 288–300 of the European conference on information retrieval.

In 2007, Carterette and Smucker published their work. Using partial relevance assessments for hypothesis testing. Pages 643-652 in the Proceedings of the sixteenth Annual Conference on Information and Knowledge Management (pp. acm).

Written by Cleverdon in 1959. Analysis of data retrieval systems. Article published in Volume 1 of the proceedings of the international conference on scientific information, pages 687-698.

In 1998, Cormack, Palmer, and Clarke published a work. Effective building of extensive test sets. Presented during the 21st annual international acm SIRGID conference on research and development in information retrieval, this book covers page numbers 282-289.

With contributions from Mizzaro, Sanderson, Culpepper, and Scholer (2014). Issue with subject engineering (Trec). Page numbers 1147–1150 from the 37th International ACM SIGRID Conference on Research and Development in Information Retrieval.

O'Riordan, C., Jose, J., and Cummins, R. (2011). Enhanced standard deviation-based query performance prediction optimization. This is part of the 34th annual SIRGID conference on information retrieval research and development, which is published in the proceedings of the conference (989-1090).

In 2001, J. H. Friedman wrote. A gradient-boosting machine for greedy function approximation. Statistics journal, 1189–1232.

Published by Grady and Lease in 2010. Mechanical Turk for crowdsourcing document relevance evaluation. Presented at the 2010 NAACL Human Language Technology (HLT) workshop,

this paper details the use of Amazon Mechanical Turk to generate language and voice data (pp. 172-179).

Mizzaro, S., Robertson, S., & Guiver, J. (2009). A few of solid ideas: Experiments in topic set reduction for retrieval assessment. "The ACM Transactions on Information Systems" (TOIS), volume 27, issue 4, page 21, March 2019.

The authors of this work are Hall et al. (2009) and Frank et al. Updating the Weka data mining program. Volume 11, Issue 1, pages 10–18, ACM SIGKDD Explorations Newsletter.

Authors: Harman and Buckley (2009). Overview of the dependable information access workshop. Article number: 615 in the journal Information Retrieval.

Azzopardi, L., Hauff, C., Hiemstra, D., & De Jong, F. (2010). Implications of automated system assessment. Pages 153–165 of the European Conference on Information Retrieval.

In 2009, Hauff, Hiemstra, De Jong, and Azzopardi published a working paper. The estimate of system rankings using topic subsets. Volume 18, Issue 6, Pages 1859–1862, 18th American Conference on Information and Knowledge Management.

Daniel Hawking (2000). Overview of the trec-9 web track. Regarding Trec. Published in 2002 by Hawking and Craswell. Web track overview for trec-2001. pages 61–67 of the NIST special issue.

The authors of this work are Hosseini, M., Cox, I. J., Milic-Frayling, N., Shokouhi, M., and Yilmaz, E. (2012). To assess their systems, they developed an uncertainty-aware query selection model. This is part of the 35th annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 901-910).

Milic-Frayling, N., Vinay, V., Hosseini, M., & Sweeting, T. (2011). To get further relevance judgments, choose a subset of searches. Pages 113–124 of the Conference on the Theory of Information Retrieval.

In 1975, Jones and van Rijsbergen published a paper. The need of and access to a "ideal" information retrieval test set (report no. 5266 from the British Library's research and development division), 43.

With Sung, H. (2014), Kazai, G. Efficient preference-based ir assessment using dissimilarity-based query selection. In "European conference on information retrieval" (pp. 172-183).

G. Kendall (1938). A novel measure of rank correlation. Article ranging from 81 to 93 pages in Biometrika, volume 30, issue 1/2.

In 2015, Lewandowski published a dissertation. Using a statistically valid sample of queries, we test how well various online search engines retrieve specific information. Article 66(9), pages 1763–1775, published in the Journal of the Association for Information Science and Technology.

In 2011, Lin and Cheng published. Query sampling for learning data fusion. Included in the 20th Annual ACSM Conference on Knowledge and Information Management (pp. 141-146).

A study conducted by McDonnell, Lease, Kutlu, and Elsayed in 2016 was published. Just why does it matter? gathering the reasoning behind relevance assessments made by annotators. Held in conjunction with the 4th Annual AAAI Conference on Human-Centric Computing and Crowdsourcing (HCOMP) (pp. 139-148).

In 2015, Mehrotra and Yilmaz published a paper. Learn to rank using submodular functions with representative and informative query selection. The paper is part of the 38th annual SIRGID conference on information retrieval research and development, which was published in the proceedings (pp. 545–554).

S. Robertson and S. Mizzaro (2007). Hits hits trec: using network analysis to investigate the outcomes of the IR assessment. With the help of the pages 479–486 included in the proceedings of the 30th annual international acm SIGRID conference on research and development in information retrieval.

In 2007, Moffat, A., Webber, W., and Zobel, J. Comparing strategic systems via targeted relevance evaluations. Included in the proceedings of the 30th annual international acm SIRGID conference on information retrieval research and development (pp.375–382).

The authors of this work are Moghadasi, Ravana, and Raman (2013). Assessment methods for information retrieval systems that are inexpensive: a literature study. Public Health Informatics, 7(2), 301-312.

Authors: Nuray and Can (2006). Classification of data fusion-based information retrieval systems automatically. Journal of Information Processing and Management, 42(3).